MA-SBI: COMPUTABLE CEILINGS DECIDE WHICH MISSPECIFICATION CORRECTOR TO FIT

Anonymous authors
Paper under double-blind review

ABSTRACT

Simulation-based inference degrades when the simulator misspecifies the process that produced the real observations, and the existing repairs ask for data the practitioner rarely has, such as ground-truth parameters paired with real observations, a functional form for the noise, or a credible alternative prior. What practitioners often do have is a written record of the conditions under which each observation was collected, such as the instruction text given to participants or the policy bulletin in force that week. We show that the share of the simulator's error this side information can remove has a ceiling that is computable in advance from held-out residuals, without evaluating any posterior, and that a second ceiling says whether the correction also needs to see the observation. We introduce misspecification-aware simulation-based inference (MA-SBI), a family of observation-space correctors that the two ceilings select between, fitted in front of a frozen amortised posterior. In the Gaussian-linear model the conditional mean provably attains both ceilings, and the fitted correction is identically zero when the simulator is well specified, however the prior behaves. A second result separates the estimators by sample complexity, since the structured one pays for the rank of the embedded design where an unstructured one pays for the width of the encoder. Across four benchmarks and three physical datasets the ceilings are tight, and given adequate calibration every arm whose closed-form ceiling exceeds five per cent lands between 0.988 and 1.0 of its own, so the number tells a practitioner what an estimator will achieve before one is fitted. On a wind-tunnel benchmark a corrector that uses no parameter labels matches a calibration-based method that uses them, and we report the negative results too, including an epidemiological dataset whose apparent coverage gain does not survive our pre-registered test.

1 INTRODUCTION

Simulation-based inference (Cranmer et al., 2020) fits parameters of a mechanistic simulator when its likelihood is intractable; neural posterior estimation trains a conditional density estimator $q(\theta \mid y)$ on simulated pairs $(\theta, y_{\mathrm{sim}})$ and returns an approximate posterior in one forward pass. Its persistent weakness is misspecification: real observations y_{obs} come from a process the simulator does not reproduce, and the resulting posteriors are biased, overconfident, or both (Wehenkel et al., 2025).

Existing repairs ask for data the practitioner rarely has: ground-truth parameter values paired with real observations, an explicit functional form for the noise, or a credible alternative prior. What practitioners do have is a record of the conditions under which each observation was collected, written down in words: the instruction text given to participants, the policy bulletins in force that week. We write this side information as z and ask how much of the simulator's error is predictable from z, and whether a corrector needs to look at the observation as well.

Two ceilings. Write $\delta = y_{\mathrm{obs}} - y_{\mathrm{sim}}$ for the misspecification. A corrector that reads only z competes with $\mathbb{E}[\delta \mid z]$, so the fraction of the error it can remove is bounded by $C_z = 1 - \mathbb{E}\|\delta - \mathbb{E}[\delta \mid z]\|^2 / \mathbb{E}\|\delta\|^2$; one that also reads the observation competes with $\mathbb{E}[\delta \mid y_{\mathrm{obs}}, z]$ and has ceiling $C_{yz} \geq C_z$. Both are δ-space quantities estimated from held-out calibration data without training a corrector or running a posterior. Their difference $C_{yz} - C_z$ is the measurable value of letting the correction depend on y, and decides which estimator to fit.

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.

Contributions.

1. Two computable ceilings and a decision rule that picks between a z-only corrector, a (y, z)-conditional corrector, and no corrector, with achievability proved in the Gaussian-linear model (Theorem 1).
2. A sample-efficiency theorem (Theorem 2) whose calibration multiplier $(d_y + d_z + 1)/K$ counts d_z as the rank of the embedded calibration design, not the encoder width.
3. Pairing-free correction from posterior-predictive residuals, with an exact account of what the residual recovers (Theorem 3) and a pre-flight criterion validated as a classifier.
4. Open-vocabulary side channels, with a one-hot-with-label control that isolates the mechanism, and an evaluation against calibration-based, noise-model, estimator-repair and domain-transfer baselines on three physical datasets and two negative controls.

What the method needs. The point corrector is trained on observation-aligned triples $(z_i, y_{\mathrm{sim},i}, y_{\mathrm{obs},i})$, with the simulator run at the inputs that produced the real observation. That is a different requirement from parameter-labelled calibration, not a lighter one, and it holds only where a simulator can be re-run at a recorded operating point. Section 4 therefore gives two pairing-free variants first, each with its assumption written out, and the wind-tunnel result below is obtained with one of them.

2 Related work

We group prior work by what it asks the practitioner for, in the taxonomy of Elsemüller et al. (2025): methods target (1) the estimator, (2) the observation or likelihood, or (3) the prior. Ours is target (2).

Calibration-based methods. RoPE (Wehenkel et al., 2025) treats the simulator-reality gap as optimal transport between learned summary representations fitted on calibration pairs (θ^*, y^*); FRISBI (Senouf et al., 2025) uses mini-batch transport and FMCPE (Ruhlmann et al., 2025) flow matching. All three need parameter labels on real observations.

Noise models and robust summary statistics. NNPE and RNPE (Ward et al., 2022) add spike-and-slab noise to simulator outputs and denoise observed summaries under that model, so the practitioner supplies a functional form for the contamination. Huang et al. (2023), Bharti et al. (2026) and Dellaporta et al. (2022) work from repeated observations, learning summary statistics or noise models whose simulated and observed distributions match. Khoo et al. (2026) correct the observed summary at test time on a frozen estimator by minimising a distance to the simulated summary distribution, the closest mechanism to ours without a side channel.

Model discrepancy. Modelling $y_{\mathrm{obs}} - y_{\mathrm{sim}}$ as a function of recorded conditions is Bayesian calibration with a discrepancy term (Kennedy & O'Hagan, 2001; Higdon et al., 2004), whose non-identifiability against the parameter is well known (Brynjarsdóttir & O'Hagan, 2014). Two things differ here. We never estimate θ and δ jointly: the backbone is frozen and the corrector is fitted in δ-space alone, which makes the slice-wise condition of Theorem 5 the right one. And we test whether the discrepancy is predictable from the recorded conditions, the quantity C_z measures; that literature assumes it. Diagnostics and robust variants for misspecification in likelihood-free inference (Frazier et al., 2020; Frazier & Drovandi, 2021; Cannon et al., 2022) do not condition on a recorded side channel.

Estimator repair. Self-consistency methods reuse the simulator likelihood on unlabelled real observations to penalise inconsistency between prior, likelihood and posterior (Schmitt et al., 2023; Mishra et al., 2025), and unsupervised domain adaptation aligns simulated and real embeddings (Elsemüller et al., 2025). Both leave the observation alone, so they compose with observation-space correction instead of competing with it; Section 6.7 runs the composition and reports where it loses. Kelly et al. (2023), Kelly et al. (2026), Gao et al. (2023) and Tomaselli et al. (2025) sit here too, adding per-observation adjustment parameters, a preconditioning step, or a generalised Bayesian target.

Domain transfer and prior adaptation. SPIN (Anonymous, 2026) learns a map from unpaired unlabelled real observations to the simulated distribution, the data regime closest to ours minus the side channel. PriorGuide (Yang et al., 2026) steers a diffusion posterior toward a different prior at test time; that is target (3) work on a different failure, and Theorem 4 says why the two families cannot substitute for one another.

Side information. Brehmer et al. (2020) extract side information from the simulator itself, the joint score and joint likelihood ratio recorded during a run, which requires instrumenting it; ours is external, recorded by whoever collected the data, with the simulator a black box. Text has also been used as a direct predictor of outcomes (Wang et al., 2024; Li et al., 2024), the opposite role: here z indexes how the simulator is wrong and never predicts θ.

3 Setting and target

Model. A simulator $y_{\mathrm{sim}} = f(\theta, \varepsilon) \in \mathbb{R}^d$ is driven by a parameter $\theta \sim \pi$ and internal randomness ε. Real observations satisfy $y_{\mathrm{obs}} = y_{\mathrm{sim}} + \delta$, where δ is the misspecification, and side information $z \in \mathcal{Z}$ is recorded alongside each. We assume throughout **(A1′)**: z is independent of (θ, ε), and (δ, z) is independent of θ given y_{sim}. Both mechanisms used in the experiments satisfy (A1′), the additive per-context shift because δ is drawn independently of the simulator, the multiplicative one because δ is a function of y_{sim}, z and external noise. We also assume **(A2)**: $\mathbb{E}\|\delta\|^2 < \infty$ and $\mathbb{E}\|\theta\|^2 < \infty$, and all statements range over measurable correctors h with $\mathbb{E}\|h\|^2 < \infty$.

Which posterior we target. We target

$$p_{\mathrm{sim}}\big(\theta \mid y_{\mathrm{obs}} - \delta^*(z)\big), \qquad \delta^*(z) = \mathbb{E}[\delta \mid z], \tag{1}$$

the simulator's own posterior evaluated at a corrected observation. The ceiling we can reach is the simulator-oracle posterior $p_{\mathrm{sim}}(\theta \mid y_{\mathrm{sim}})$, not the posterior of a correctly specified model. Section 6.6 measures what that costs.

Three data regimes. With *observation-aligned triples* $(z_i, y_{\mathrm{sim},i}, y_{\mathrm{obs},i})$, the simulator run at the inputs that produced $y_{\mathrm{obs},i}$, δ_i is observed and the point corrector of Section 4.3 applies. With *context pools*, real and simulated observations available per context but not paired within it, distribution matching applies (Section 4.2). With *observations and text only*, the posterior-predictive bootstrap applies at the price quantified by Theorem 3. We use *context* for the K indices of the misspecification mechanism and *condition* for the three evaluation settings, well-specified, misspec-informative and misspec-uninformative; Table 15 fixes the symbols.

4 Method

Misspecification-aware simulation-based inference (MA-SBI) corrects in observation space. Its general member is a residual corrector, MA-SBI-R, which replaces the observation by $y_{\mathrm{obs}} - h(y_{\mathrm{obs}}, z)$ before a frozen amortised posterior evaluates it (Figure 1). The additive z-only corrector $h(z)$ is its sample-efficient special case, the right choice whenever C_{yz} is close to C_z.

4.1 The two ceilings and the decision rule

Define

$$C_z = 1 - \frac{\mathbb{E}\|\delta - \mathbb{E}[\delta \mid z]\|^2}{\mathbb{E}\|\delta\|^2}, \qquad C_{yz} = 1 - \frac{\mathbb{E}\|\delta - \mathbb{E}[\delta \mid y_{\mathrm{obs}}, z]\|^2}{\mathbb{E}\|\delta\|^2}. \tag{2}$$

We estimate both by regressing held-out δ on z and on (y_{obs}, z), using per-context means where z is categorical and a fitted network otherwise. Neither estimate touches the posterior, so the decision costs one δ-space regression. Both ceilings need observation-aligned triples; in the two pairing-free regimes only the bootstrap surrogate of Section 4.2 is available, which estimates C_z up to the known deflation of Theorem 3 and does not estimate C_{yz} at all. The rule reads off three numbers. If $C_z \approx 0$, no corrector of this family helps and the practitioner should stop. If $C_{yz} - C_z$ is small relative to the estimation cost of Theorem 2, fit the additive z-only corrector; if it is large, fit the residual corrector.

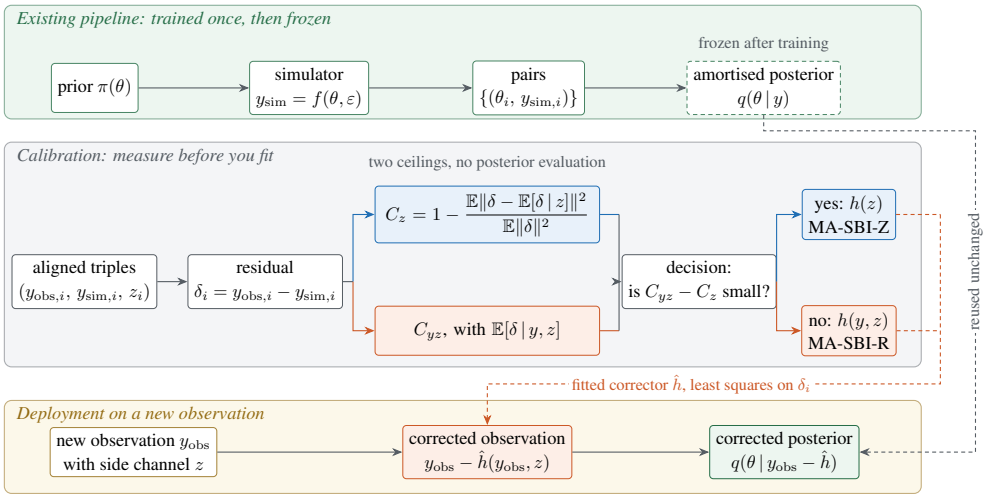


Figure 1: The pipeline. The amortised posterior is trained once on simulator output and frozen. Observation-aligned calibration triples give residuals δ_i, from which the two ceilings are computed with no posterior evaluation. Their gap picks the corrector. That corrector is fitted by least squares on the residuals and applied to the observation before the frozen posterior reads it.

A large gap says that conditioning on y pays, not that the mechanism is multiplicative: a purely additive δ independent of z already gives $C_{yz} > C_z$, because y_{obs} carries information about δ that z does not.

4.2 Pairing-free variants

Distribution matching. With context pools rather than pairs, we fit h so that the law of $y_{\text{obs}} - h$ within a context matches the law of y_{sim} within that context, by minimising an energy distance between the two sample sets. *Assumption:* within a context, the parameter marginal that generated the real observations matches the prior used for the simulations. If the real observations in a context were collected at systematically different θ, the matched shift absorbs that difference. The assumption is testable only against a θ sample, which is what this regime lacks, so we state it and leave it unverified.

Posterior-predictive bootstrap. With observations and text only, we use the residual of the frozen posterior itself,

$$\delta^{\text{boot}}(y_{\text{obs}}) = y_{\text{obs}} - m(y_{\text{obs}}), \qquad m(y) = \mathbb{E}_{\theta \sim q(\cdot \mid y)} \mathbb{E}_{\varepsilon} f(\theta, \varepsilon), \tag{3}$$

and fit the corrector on δ^{boot} pooled within a context. *Assumption:* the posterior-predictive mean map m is contractive, with Lipschitz constant $L_m < 1$, so the part of δ it absorbs into θ is a bounded fraction of δ. Being far from the identity is not enough: $m(y) = y - c$ has $L_m = 1$ and recovers nothing. The label-free route estimates only C_z and must be pooled within a context; Appendix A gives the deflation factor and the reasons.

Pre-flight criterion. Before either variant is run we compute the per-cell residual norm divided by the per-context median residual norm. This statistic separates the regime where the bootstrap route is worth running from the regime where it is not, and Section 6.1 reports it as a classifier with an area under the receiver operating characteristic curve.

4.3 The point corrector

With observation-aligned triples, $\delta_i = y_{\text{obs},i} - y_{\text{sim},i}$ is observed and we fit by least squares,

$$\hat{h} = \arg\min_{h} \frac{1}{n} \sum_{i=1}^{n} \left\| \delta_i - h(y_{\text{obs},i}, z_i) \right\|_2^2, \tag{4}$$

over a three-layer multilayer perceptron with Gaussian error linear unit (GELU) activations, 128 to 256 hidden units, trained for 300 epochs at batch size 128 with early stopping on a held-out split of δ. Where a context may be unlike anything in calibration, we use a kernel corrector instead, a similarity-weighted average of calibration residuals whose weights flatten off support, so it shrinks to the calibration mean instead of extrapolating; Section 6.8 shows this is the difference between a usable and an unusable corrector. Dropping the y input gives the additive form $h(z)$, the corrected posterior is $q(\theta \mid y_{\text{obs}} - \hat{h})$, and the backbone is never retrained.

5 THEORY

Proofs are in Appendix B. Two lemmas carry the section, the conditional-mean projection identity (Lemma 7) and a transfer lemma (Lemma 8) that converts δ-space error into excess parameter risk whenever the posterior-mean map is linear and correct. Write $\Delta\text{MSE}(h) = \mathbb{E}\|\mu(y_{\text{obs}} - h) - \theta\|^2 - \mathbb{E}\|\mu(y_{\text{sim}}) - \theta\|^2$, $\text{MMSE}(\delta \mid W) = \mathbb{E}\|\delta - \mathbb{E}[\delta \mid W]\|^2$, $\text{GC}(h) = 1 - \mathbb{E}\|\delta - h\|^2/\mathbb{E}\|\delta\|^2$ for the gap closure of h in δ space, and $\bar{\delta} = \mathbb{E}\delta$. (GL) denotes the Gaussian-linear model $\theta \sim \mathcal{N}(0, \sigma_p^2 I_d)$ with $d_\theta = d$, $y_{\text{sim}} = \theta + \varepsilon$ and $\varepsilon \sim \mathcal{N}(0, \sigma_s^2 I_d)$, under which the well-specified posterior mean is $\mu(y) = \alpha y$ with $\alpha = \sigma_p^2/(\sigma_p^2 + \sigma_s^2)$.

Theorem 1 (TH1, achievability of the two ceilings)**.** *Assume (A2). Among all measurable $h : \mathcal{Z} \to \mathbb{R}^d$ the ℓ_2-optimal corrector is $h^\star(z) = \mathbb{E}[\delta \mid z]$ and the best attainable value of $1 - \mathbb{E}\|\delta - h(z)\|^2/\mathbb{E}\|\delta\|^2$ equals C_z. Among all measurable $h : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}^d$ the optimum is $h^{\star\star}(y_{\text{obs}}, z) = \mathbb{E}[\delta \mid y_{\text{obs}}, z]$, the value is C_{yz}, and $C_{yz} \geq C_z$. Under (A1′) and (GL) the gap closure in excess parameter risk equals C_z and C_{yz} exactly. If in addition $\delta \sim \mathcal{N}(0, \sigma_\delta^2 I_d)$ is jointly Gaussian with z, with canonical correlations $\rho_1, \ldots, \rho_d$ under the convention $\rho_k = 0$ for $k > \dim z$, then $1 - C_z = d^{-1} \sum_k (1 - \rho_k^2) \geq \exp(-2I(\delta; z)/d)$, with equality exactly when all ρ_k are equal.*

The identity is exact only where the posterior-mean map is linear, so on a flow backbone C_z is a diagnostic and the general statement is one-sided (Remark 9); the Gaussian clause recovers the mutual-information bound of earlier work (Appendix B.9).

Theorem 2 (TH2, sample efficiency)**.** *Fit both families by least squares under isotropic residual noise σ^2 and correct specification of each target, taking the minimum-norm (pseudoinverse) solution when the design is rank deficient, and write $p = \text{rank}(X)$ for the design X of the family in question. The in-sample excess risk is $\sigma^2 d\, p/n$. Under a random design with independent sub-Gaussian rows, positive definite second moment and $n \gtrsim p \log p$, the out-of-sample excess risk is $\sigma^2 d\, p/n\,(1 + o(1))$. For the additive z-only estimator over K contexts $p = K$, and under proportional allocation $n_k = np_k$ its out-of-sample excess risk is $\sigma^2 dK/n$ whatever the context frequencies. For the residual linear estimator on the design $[y, z, 1]$, $p \leq d_y + d_z + 1$, and with K distinct embedded contexts $p \leq d_y + \min(d_z, K) + 1$ whatever the width of the encoder. To reach the same excess risk the residual estimator needs p/K times as many calibration examples.*

Encoder width does not enter the count. What does is the rank of the embedded calibration design, and the comparison holds at matched approximation error, so the additive form stays competitive only while $C_{yz} = C_z$. Appendix B.9 works the rank count through and compares it with the fitted multipliers of Section 6.3.

Theorem 3 (TH3, the posterior-predictive residual is a projection)**.** *Let m be L_m-Lipschitz. Then $\|\delta^{\text{boot}}(y_{\text{obs}}) - \delta\| \leq \|y_{\text{sim}} - m(y_{\text{sim}})\| + L_m\|\delta\|$, the two terms being the well-specified predictive bias and the part of δ the posterior absorbs into θ. Under (A1′) and (GL), $\mathbb{E}[\delta^{\text{boot}} \mid z] = (1 - \alpha)\mathbb{E}[\delta \mid z]$, so pooling within a context recovers the direction of the target exactly and a known fraction $1 - \alpha$ of its magnitude, and a corrector fitted on the pooled residual attains gap closure $(1 - \alpha^2)\, C_z$. The iteration $\delta_{t+1} = y_{\text{obs}} - m(y_{\text{obs}} - \delta_t)$ contracts at rate α to the degenerate fixed point $\delta^\star = y_{\text{obs}}$, at which the corrected input is zero, so the limit returns the prior whatever the data and the iteration is not a refinement of the one-step estimate.*

Circularity is the obvious worry here, and the theorem answers it. The bootstrap estimates a known multiple of the target, and α, or its non-linear analogue L, is the pre-flight criterion (Appendix B.9).

Theorem 4 (TH4, scope)**.** *If the simulator is well specified, so $\delta = 0$ almost surely, then the population-optimal corrector is $h^\star \equiv 0$ for every prior and for every joint law of z and the prior perturbation. The prior never enters the corrector's objective, so no amount of information about*

prior misspecification carried by z can be used. When both simulator and prior are wrong, the corrector acts only on the conditioning argument of a fixed q, so its target remains $q(\theta \mid y_{\mathrm{sim}})$ whatever the prior error.

Theorem 5 (TH5, identifiability)**.** *Let $Z(y,z) = \int_\Theta p(y \mid \theta', z)\pi(\theta')\, d\theta'$ and $\tilde{q}(\theta \mid y, z) = p(y \mid \theta, z)\pi(\theta)/Z(y,z)$. Under the regularity condition (R1) of Appendix B, $\tilde{q}$ is well defined. Say that θ is identifiable from the joint law of (y,z) if distinct $\theta_1 \neq \theta_2$ induce distinct joint laws of (y,z), the law of z being fixed. Then θ is identifiable if and only if the augmented map $(\theta, z) \mapsto (p(\cdot \mid \theta, z), z)$ is injective, equivalently if and only if $\theta \mapsto p(\cdot \mid \theta, z)$ is injective for P_z-almost every z. In the necessity direction the set of contexts on which slice injectivity fails is assumed to have positive P_z measure.*

The slice-wise form matters: since z is observed, a collision between (θ_1, z_1) and (θ_2, z_2) with $z_1 \neq z_2$ costs nothing. Appendix B.9 maps the three preconditions onto the instrumental variable triple (Angrist et al., 1996).

Theorem 6 (TH6, sub-Gaussian extension)**.** *Let δ satisfy the multivariate sub-Gaussian condition $\mathbb{E}[e^{u^\top(\delta - \mathbb{E}\delta)}] \leq e^{\|u\|^2\tau^2/2}$ for all $u \in \mathbb{R}^d$. Then $\mathbb{E}\|\mathbb{E}[\delta \mid z] - \mathbb{E}[\delta]\|^2 \leq 2\tau^2 I(\delta; z)$, and if μ is L_μ-Lipschitz the induced posterior-mean displacement is at most $2L_\mu^2\tau^2 I(\delta; z)$.*

Note the quantifier. It is multivariate, and charging a per-coordinate proxy is exactly the error that makes the bound appear to fail (Appendix B.9).

6 Experiments

6.1 Setup

Benchmarks come from the simulation-based inference benchmark suite (Lueckmann et al., 2021), which supplies reference posteriors, plus repository simulators for the tasks whose reference implementations are unavailable. Misspecification is a context-indexed shift on the simulator output, with $K = 4$ contexts and an interpolation parameter α_{mix} that moves the mechanism from purely additive to purely multiplicative. The side channel is a sentence per context, encoded by a sentence transformer by default, with a character n-gram hash as an ablation, and the backbone is a normalising flow with standardised inputs and outputs, trained to convergence. Metrics are the classifier two-sample test (C2ST) against the reference posterior, gap closure in δ-space and in posterior-mean squared error, and two one-sided tests (TOST) against the oracle. Tables report the median over seeds, with interquartile range or across-seed standard deviation where the spread carries the claim. Tables built from matched seeds or cells can differ from Table 1 by one to three points for the same method; the registry records each table's cell set. Gap closure in δ-space is $1 - \mathbb{E}\|\delta - h\|^2/\mathbb{E}\|\delta\|^2$; in posterior-mean squared error and in C2ST it is the improvement as a fraction of the uncorrected-to-oracle gap, so 100 matches the simulator oracle and a few points above 100 is seed noise. The neural oracle reaches C2ST 0.72 on SLCP instead of 0.5 because the backbone does not match the reference posterior exactly at this budget. Gap closure is stated against that oracle, and because C2ST saturates near it, the δ-space and posterior-mean figures carry the quantitative claims. The bootstrap pre-flight statistic separates the easy from the hard regime, labelling a cell hard when its shift is at least twice the posterior-predictive spread, with an area under the receiver operating characteristic curve of 0.975 over five seeds.

6.2 Main comparison

One anchor run fixes the protocol that every benchmark table uses. On the simple-likelihood-complex-posterior task over five seeds, uncorrected estimation reaches C2ST 0.839 and the oracle 0.72; the corrector with sentence encoding reaches 0.716. That closes 104.1 per cent of the C2ST gap and 100.0 per cent of the δ-space gap, at a TOST p value of 0.0099 against the oracle at an equivalence margin of 0.02. Across four tasks at ten seeds (Table 1) both corrector forms reach the oracle on every misspecified-informative cell, the transport baseline recovers a small fraction of the gap, and direct regression from z to θ or to y is worse than doing nothing.

Table 7 adds the other two conditions and measures graceful degradation (Corollary 12): with an uninformative side channel the z-only corrector degrades to the best constant shift, not to zero, and the residual arm to what the observation alone supports, which is higher (Appendix A.5).

Table 1: Main results. Median C2ST gap closure against the simulator oracle, four benchmarks, misspecified z-informative condition, pooled over $n_{\text{cal}} \in \{100, 500, 2000\}$ and ten seeds. † marks methods given parameter labels. The component rows are the two halves of the RoPE transport map; further baselines are in Tables 16 and 9.

Method	SLCP	Two Moons	Gaussian Linear	Bernoulli GLM
NPE (baseline)	0	0	0	0
RoPE-EMD†	10.6	10.4	12.5	31.5
$g(z) \to \theta$†	−139.6	−4.5	−540.7	−21.2
$g(z) \to y$†	−134.0	−4.5	−482.5	−21.3
Marginal component only	5.1	11.1	17.8	7.8
Conditional component only	84.6	21.9	81.2	59.4
MA-SBI-Z (additive, z only)	100.4	100.1	102.7	99.7
MA-SBI-R (residual)	99.1	100.6	102.2	99.7
Oracle (neural)	100	100	100	100

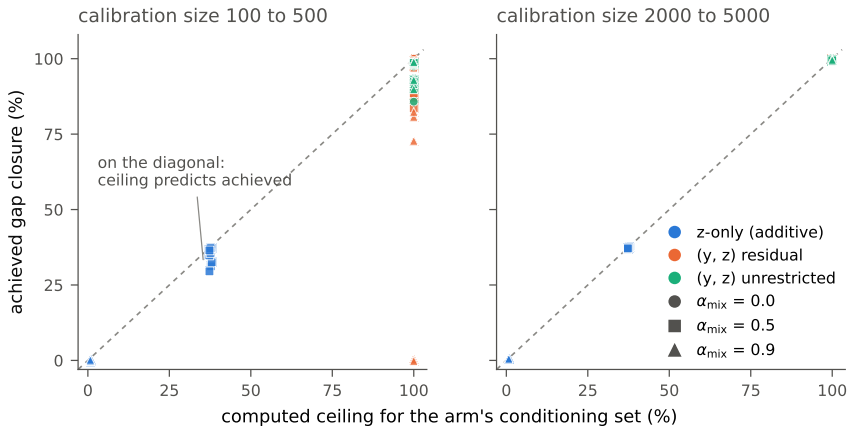


Figure 2: Achieved gap closure against the ceiling of each arm's own conditioning set, five seeds, three mixtures; the diagonal is equality. At large n_{cal} every arm sits on it; at small n_{cal} the z-only arm gets there first. Arms with closed-form ceilings only; marker shape is α_{mix}.

6.3 Estimator family and sample efficiency

Table 13 and Figure 2 decide the method. At $n_{\text{cal}} = 2000$ on SLCP, across five seeds and three misspecification mixtures, every arm with a closed-form ceiling above 5 per cent reaches between 0.988 and 1.0 of that ceiling, with median 0.998 and seed standard deviations under half a point. The δ-space ceiling is attained on this grid, so it functions as a prediction as well as a bound, and in C2ST the same arms reach the oracle too (Table 1). Two cases are excluded: the y-only arm's ceiling is a nearest-neighbour lower bound, so the arm can exceed it, and the additive arm at $\alpha_{\text{mix}} = 0.9$ has a ceiling of 0.8 per cent, so what fraction of it the arm attains says nothing. That additive ceiling collapses from 100 to 0.8 as the mechanism turns multiplicative, the collapse $C_{yz} - C_z$ is designed to detect, while the residual arm holds at its ceiling throughout.

The full grid of 360 cells shows the same separation throughout, and where calibration is short the residual arm pays for its extra inputs (Appendix A.6).

Table 18 measures the multiplier directly: the fitted ratio falls below the rank-based bound in every row, so the ordering of Theorem 2 survives while its magnitude does not (Appendix A.1).

6.4 Beyond fixed shifts

A fixed per-context shift matches the additive corrector's own assumptions, so we break it three ways. First, the shift is made stochastic, $\delta = \text{shift}(\text{context}) + \tau\mathcal{N}(0, I)$, and Table 19 tracks the closed-form prediction with a mean absolute error of 0.47 percentage points and no overshoot at $\tau = 0$;

the C2ST column saturates while posterior-mean gap closure is still large, which says something about the metric and nothing about the method. Second, z is made continuous and three bands are deleted from calibration at the peak, the zero crossing and the trough of a sine, with evaluation only inside the deleted bands. The sentence encoder reaches 98.0 and 90.6 per cent there where the hash reaches 52.2 and -53.4 (Table 20): a lookup table fails at a sign-flipping trough it never saw, and the hash fails exactly there. Third, Section 6.3 already interpolates the mechanism toward multiplicative (Table 13). The stochastic and paraphrase curves are in Figure 3.

Every backbone passes a simulation-based calibration gate before it carries a number, and on the cells that pass, the corrector arms move far less between a flow and a diffusion backbone than the baselines do (Appendix A.2).

6.5 Open-vocabulary side channels

The obvious objection is that z is a categorical variable in disguise, and the test is the slope, not the level. We go from one to 10 training paraphrases per context and evaluate on a disjoint held-out pool (Table 14). The sentence encoder rises from 16.7 to 34.4 per cent gap closure. The hash falls from 11.9 to -92.7, and one-hot over exact strings stays flat at 10.7 to 12.0. One-hot given the true context label reaches 100 per cent and marks the ceiling of the lookup interpretation. A categorical variable in disguise cannot improve as paraphrases are added, and the sentence encoder does.

Larger encoders raise held-out transfer without reaching the lookup ceiling and without helping transfer to unseen continuous z (Appendix A.3).

6.6 Scope and prior misspecification

Hold the simulator well specified and perturb the test prior by a mean shift, a scale change, bimodality or Student-t tails. The fitted correction norm is at most 0.001 under every perturbation and gap closure stays within 0.6 per cent of zero (Table 21), which is what Theorem 4 predicts. When both simulator and prior are wrong, the corrected posterior lands within 0.001 of the simulator oracle and the fitted correction norm recovers the true shift norm of 0.300. So a prior shift is representable as an input correction here, and the method still does not find it, because nothing in the corrector's objective points at the prior. That is a sharper negative result than an inability to represent the shift (Appendix A.4).

6.7 Estimator repair and composition

Composing estimator repair with observation-space correction loses to correction alone. Self-consistency fine-tuning on unlabelled real observations closes 45.9 per cent of the context gap over five seeds, observation-space correction closes 105.2, and composing the two closes 69.6; the composition loses because self-consistency fine-tuning costs accuracy on clean inputs, moving the oracle-input C2ST from 0.72 to 0.771.

Table 16 adds the denoising and domain-adaptation arms, which close 5.1 and -1.9 per cent on SLCP against 100.0 for MA-SBI-R, and preconditioned robust NPE (Kelly et al., 2026) reaches -17.1 per cent on SLCP and 0.1 on Two Moons (Table 9). None of these baselines is broken: a spike-and-slab error model, a marginal alignment penalty and data-dependent preconditioning have no way to express a shift that depends on the operating context, and a whole-observation shift is still a plausible simulator draw, so there is nothing for them to flag.

6.8 Real data

The positive real-data experiment is the causal-chamber wind tunnel, used by calibration-based and domain-transfer methods: the parameter is a hatch position, the observation a downwind pressure trace, the side channel the experiment protocol. Misspecification norm is 13.39 Pa against 1.85 Pa of simulator noise, between 6.6 and 7.7 times in every experiment, and the text leakage test returns $R^2 = -0.004$. The z-only ceiling is 0.651 in sample and 0.575 under leave-one-experiment-out, the figure quoted below. A single global mean reaches 0.607; conditioning on z removes a further 0.112 of the residual variance under this simulator, and 0.338 once the simulator is given the true exhaust load (Table 12).

Two findings, against the recorded hatch angle. The multilayer perceptron corrector is fragile when the held-out context sits off the support of the five training contexts. It extrapolates in a 384-dimensional embedding and collapses the posterior: root mean squared error moves from 7.27 to 17.97 degrees, and the 90 per cent interval width from 1.25 to 0.01. A kernel corrector averaging calibration residuals by similarity in embedding space, with weights that flatten off support, fixes this. Its bootstrap version, with no parameter labels, lowers posterior-mean error from 7.27 to 6.23 degrees and beats uncorrected estimation on 15 of 15 fold-seed comparisons (five experiments, three seeds), against 6.45 degrees for a transport baseline with labels.

The ceilings also work as a decision, with one caveat: from the four training experiments alone they pick the (y, z)-conditioned corrector on every fold, so agreement tests calibration, not the ability to switch. That arm wins 12 of 15, the three losses at the extreme exhaust load where the held-out context is off support (Table 11, Figure 7). Bootstrap residuals, with no labels, rank the arms the same way on every fold (Table 3), with the Section 4.2 caveat that the bootstrap (y, z) ceiling is not informative alone. Figure 6 gives per-method error and win share on all three datasets.

On a wind-tunnel dataset with 28 random-walk protocols (Table 4) the ceilings say in advance that the side channel adds 1.7 points over a constant shift, and the z-only corrector then wins 12 of 28 folds, which is chance. The (y, z) ceiling is high there, 72.4, and the (y, z) arms do win more often, 16 and 18 of 28. On absolute error, though, the transport baseline with labels beats every arm we fit: 23.05 against 25.35 and 25.75 degrees uncorrected. We report this as a loss. A high (y, z) ceiling says the information is present, not that a kernel corrector on 27 protocols finds it. The light tunnel splits the two routes. Misspecification there is fifty times the simulator noise and the text leaks nothing, but the leave-one-experiment-out ceiling is 0.261 against the wind tunnel's 0.575. The corrector without labels ties uncorrected estimation (9 of 15, $p = 0.847$) while the transport baseline with labels wins (11 of 15, $p = 0.035$; Table 10). The labelled multilayer perceptron corrector, unusable on the wind tunnel, is best here, 0.371 against 0.483 uncorrected and 0.424 for transport, winning 12 of 15 ($p = 0.004$), because the held-out contexts interpolate the training ones.

Three datasets sit outside the working range (Table 22): one defeat, two caught in advance. The COVID policy dataset is the one we expected to work, and it does not. With the case level made identifiable by treating the initial infected count as a parameter, no pooled variant improves predictive negative log likelihood. Median change runs between -1.2 and -0.0 per cent, win rates between 0.33 and 0.5, and none of the six leave-one-country evaluations passes the pre-registered criterion of a median improvement of at least 2 per cent at a win rate above 0.6. The coverage gain that a first look suggests is interval width. On influenza surveillance, with the public surveillance narrative as z, the text passes the leakage screen ($R^2 = 0.051$) but the held-out-season ceiling proxy is -0.011, indistinguishable from zero, so the criterion rules it out in advance. On a well-specified cognitive dataset the fitted correction is indistinguishable from zero.

7 Limitations and conclusion

Prior misspecification is out of reach: Theorem 4 says the objective never sees the prior, and Section 6.6 measures a fitted correction of zero under four perturbations, so a wrong prior needs a target-(3) method. Where the simulator can be rebuilt to condition on the context, rebuild it, and where a discrepancy can be made identifiable, nothing is left to correct (Section 6.8). The data requirement is real too: observation-aligned triples need a simulator re-runnable at a recorded operating point, each pairing-free variant carries the assumption of Section 4.2, and the exact ceiling identity needs a linear posterior-mean map, so on a flow backbone C_z is a diagnostic and the method is for a frozen backbone.

Ethics statement

This work uses public simulation benchmarks, a public physical-system dataset, public epidemiological case counts with public policy text, and a public behavioural dataset. No personal data, no human subjects recruited by us, and no identifiable individuals are involved. The behavioural dataset was collected under the original authors' approvals and is redistributed by them in anonymised form. The method corrects posteriors for known simulator error and can, if misapplied, transfer a correction fitted in one context to a context where it does not hold. The ceiling estimates and the pre-flight

criterion in Section 4 exist to make that failure visible before deployment rather than after. We report negative results in the main text for the same reason. The compute used is modest, on the order of two hundred processor-hours and thirty accelerator-hours in total.

REPRODUCIBILITY STATEMENT

Every table and figure in Section 6 is generated from result files by a single script, and every number in a table or quoted from a result in the prose is a generated macro; the hand-typed exceptions are constants of the design such as seed counts, grid sizes and the closed-form multipliers. A registry maps each table and figure to the script, arguments, seeds and commit that produced it, and the build fails if a referenced result is missing or older than its registry entry. Theorem statements in Section 5 have full proofs in Appendix B, including the assumptions each one needs and the cases where the identity degrades to a one-sided bound. Pre-registered pass and fail criteria for each experiment were written before the corresponding run and are included in the supplementary material, along with the code, the seeds, the data-preparation scripts and the licences of every dataset used.

AI USE STATEMENT

Language models appear in this work as the side-channel encoder under study: all-MiniLM-L6-v2 by default, with bge-large, e5-large, gte-large and gte-Qwen2-1.5B in the encoder sweep of Appendix A.3 and a hashed character 3-gram code as the ablation. Each is frozen and used only to embed text recorded independently of this work. The synthetic side-channel sentences of Section 6.5 were written by the authors, and the side channels in the real-data experiments (policy bulletins, surveillance narratives and chamber protocols) are recorded rather than generated. Beyond that, LLM assistance was limited to grammatical checks on text. No claim, result, figure or citation in this paper was generated by an LLM

REFERENCES

Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

Anonymous. SPIN: Simulator-to-population inference under distribution shift. *arXiv preprint arXiv:2605.05652*, 2026.

Ayush Bharti, Louis Filstroff, and Samuel Kaski. Neural score matching for robust simulation-based inference. *arXiv preprint arXiv:2602.11325*, 2026.

Johann Brehmer, Gilles Louppe, Juan Pavez, and Kyle Cranmer. Mining gold from implicit models to improve likelihood-free inference. *Proceedings of the National Academy of Sciences*, 117(10): 5242–5249, 2020.

Jenný Brynjarsdóttir and Anthony O'Hagan. Learning about physical parameters: the importance of model discrepancy. *Inverse Problems*, 30(11):114007, 2014.

Patrick Cannon, Daniel Ward, and Sebastian M. Schmon. Investigating the impact of model misspecification in neural simulation-based inference. *arXiv preprint arXiv:2209.01845*, 2022.

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

Charita Dellaporta, Jeremias Knoblauch, Theodoros Damoulas, and François-Xavier Briol. Robust bayesian inference for simulator-based models via the MMD posterior bootstrap. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Lasse Elsemüller, Valentin Pratz, Mischa von Krause, Andreas Voss, Paul-Christian Bürkner, and Stefan T. Radev. Does unsupervised domain adaptation improve the robustness of amortized bayesian inference? a systematic evaluation. *arXiv preprint arXiv:2502.04949*, 2025.

David T. Frazier and Christopher Drovandi. Robust approximate Bayesian inference with synthetic likelihood. *Journal of Computational and Graphical Statistics*, 30(4):958–976, 2021.

David T. Frazier, Christian P. Robert, and Judith Rousseau. Model misspecification in approximate Bayesian computation: consequences and diagnostics. *Journal of the Royal Statistical Society: Series B*, 82(2):421–444, 2020.

Richard Gao, Michael Deistler, and Jakob H. Macke. Generalized bayesian inference for scientific simulators via amortized cost estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Dave Higdon, Marc Kennedy, James C. Cavendish, John A. Cafeo, and Robert D. Ryne. Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing*, 26(2):448–466, 2004.

Daolang Huang, Ayush Bharti, Amauri Souza, Luigi Acerbi, and Samuel Kaski. Learning robust statistics for simulation-based inference under model misspecification. In *NeurIPS*, 2023.

Ryan P. Kelly, David J. Nott, David T. Frazier, David J. Warne, and Christopher Drovandi. Misspecification-robust sequential neural likelihood for simulation-based inference. *arXiv preprint arXiv:2301.13368*, 2023.

Ryan P. Kelly, David J. Nott, David T. Frazier, David J. Warne, and Christopher Drovandi. Preconditioned neural posterior estimation under model misspecification. *arXiv preprint arXiv:2602.18004*, 2026.

Marc C. Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B*, 63(3):425–464, 2001.

Sherman J. J. Khoo, Ayush Bharti, Chris J. Oates, and François-Xavier Briol. Minimum distance summaries for robust simulation-based inference. *arXiv preprint arXiv:2602.09161*, 2026.

Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. EconAgent: Large language model-empowered agents for simulating macroeconomic activities. In *ACL*, 2024.

Jan-Matthis Lueckmann, Jan Boelts, David Greenberg, Pedro Goncalves, and Jakob Macke. Benchmarking simulation-based inference. In *AISTATS*, pp. 343–351, 2021.

Aayush Mishra, Valentin Pratz, Marvin Schmitt, and Stefan T. Radev. Self-consistency losses for amortized bayesian inference on unlabeled real data. *arXiv preprint arXiv:2501.13483*, 2025.

Pierre-Louis Ruhlmann, Pedro L. C. Rodrigues, Michael Arbel, and Florence Forbes. Flow matching for robust simulation-based inference under model misspecification. *arXiv preprint arXiv:2509.23385*, 2025.

Marvin Schmitt, Valentin Pratz, Ullrich Köthe, Paul-Christian Bürkner, and Stefan T. Radev. Leveraging self-consistency for data-efficient amortized bayesian inference. *arXiv preprint arXiv:2310.04395*, 2023.

Ortal Senouf, Antoine Wehenkel, Cédric Vincent-Cuaz, Emmanuel Abbé, and Pascal Frossard. Inductive domain transfer in misspecified simulation-based inference. In *NeurIPS*, 2025.

Francesco Tomaselli, Charita Dellaporta, and Theodoros Damoulas. On the frequentist properties of generalised bayesian inference under model misspecification. *arXiv preprint arXiv:2508.02404*, 2025.

Xinlei Wang, Maike Feng, Jing Qiu, Jinjin Gu, and Junhua Zhao. From news to forecast: Integrating event analysis in LLM-based time series forecasting with reflection. In *NeurIPS*, 2024.

Daniel Ward, Patrick Cannon, Mark Beaumont, Matteo Fasiolo, and Sebastian M. Schmon. Robust neural posterior estimation and statistical model criticism. In *NeurIPS*, 2022.

Antoine Wehenkel, Juan L. Gamella, Ozan Sener, Jens Behrmann, Guillermo Sapiro, Jörn-Henrik Jacobsen, and Marco Cuturi. Addressing misspecification in simulation-based inference through data-driven calibration. In *ICML*, 2025.

Table 2: Calibration gate. Maximum mean-rank deviation in simulation-based calibration for the flow backbone on well-specified data, under misspecification, and after the residual corrector, three seeds. A backbone that fails the well-specified column carries no calibration claim.

task	gate	well spec.	misspecified	corrected
SLCP	pass	0.023	0.042	0.022
Two Moons	pass	0.014	0.053	0.017
Gaussian Linear	pass	0.030	0.062	0.029
Bernoulli GLM	pass	0.034	0.154	0.037

Table 3: The ceiling decision on the wind tunnel made from bootstrap residuals with no parameter labels. Levels are inflated because the residual is itself a function of the observation, so the ordering is what carries the decision; it agrees with the labelled version on every fold.

seed	held-out experiment	C_{const}	C_z	$C_{y,z}$	predicted	actual	agree	same as $\delta^\star$
0	L_{out}=0.01, O_{dw}=4	70.2	71.4	99.8	(y, z)	(y, z)	✓	✓
0	L_{out}=0.5, O_{dw}=2	75.7	77.7	99.9	(y, z)	(y, z)	✓	✓
0	L_{out}=0.5, O_{dw}=4	72.0	73.7	99.8	(y, z)	(y, z)	✓	✓
0	L_{out}=0.5, O_{dw}=8	72.7	74.6	99.8	(y, z)	(y, z)	✓	✓
0	L_{out}=1, O_{dw}=4	71.5	73.1	99.8	(y, z)	z-only	×	✓
1	L_{out}=0.01, O_{dw}=4	72.8	73.9	99.9	(y, z)	(y, z)	✓	✓
1	L_{out}=0.5, O_{dw}=2	78.2	80.2	99.9	(y, z)	(y, z)	✓	✓
1	L_{out}=0.5, O_{dw}=4	74.2	75.9	99.9	(y, z)	(y, z)	✓	✓
1	L_{out}=0.5, O_{dw}=8	75.2	76.8	99.9	(y, z)	(y, z)	✓	✓
1	L_{out}=1, O_{dw}=4	74.1	75.6	99.9	(y, z)	z-only	×	✓
2	L_{out}=0.01, O_{dw}=4	72.5	73.2	99.8	(y, z)	(y, z)	✓	✓
2	L_{out}=0.5, O_{dw}=2	78.5	80.3	99.9	(y, z)	(y, z)	✓	✓
2	L_{out}=0.5, O_{dw}=4	74.1	75.8	99.9	(y, z)	(y, z)	✓	✓
2	L_{out}=0.5, O_{dw}=8	74.3	75.9	99.9	(y, z)	(y, z)	✓	✓
2	L_{out}=1, O_{dw}=4	74.7	76.3	99.9	(y, z)	z-only	×	✓

Yang Yang, Severi Rissanen, Paul E. Chang, Nasrulloh Loka, Daolang Huang, Arno Solin, Markus Heinonen, and Luigi Acerbi. Priorguide: Test-time prior adaptation for simulation-based inference. In *ICLR*, 2026.

A Deferred tables

The bootstrap surrogate ceiling. Under (GL) with δ independent of y_{sim}, the surrogate ceiling of Section 4.2 is $C_z^{\text{boot}} = C_z\,\mathbb{E}\|\delta\|^2/(\mathbb{E}\|y_{\text{sim}}\|^2 + \mathbb{E}\|\delta\|^2)$. It says nothing about C_{yz}: $\delta^{\text{boot}}(y_{\text{obs}})$ is a deterministic function of y_{obs}, so the surrogate for C_{yz} is identically one and a non-degenerate estimate would need cross-fitting over disjoint splits of y. Pooling within a context is not optional, since per observation the residual carries a nuisance term as large as the signal.

These tables are referenced from Section 6 and sit here for space. Each comes from the same generator as the tables in the main text.

A.1 Sample efficiency on neural arms

Table 18 measures the multiplier on neural arms the theorem does not cover, so the rank-based value is an upper bound rather than a prediction. At $K = 4$ the fitted ratios are 2.91 and 1.79 against the rank-based 3.0; ten paraphrases per context raise the rank to 39 and the bound to 12.0 while the fitted ratio reaches 1.36. The fitted ratio falls below the bound in every row and the bound loosens by nearly an order of magnitude at the larger rank, so the theorem's ordering survives while its magnitude does not; the nominal-width values of 18.2 and 98.2 are wrong by an order of magnitude in the other direction. The comparison also needs matched ceilings: where the additive ceiling is 37.6 per cent and the residual one 100, the ratio inverts below one, because reaching 90 per cent of a low ceiling is harder than reaching 90 per cent of a high one. Table 18 is therefore restricted to the additive mechanism, where both ceilings are 100.

Table 4: Wind tunnel, 28 random-walk protocols, one seed, 28 leave-one-protocol-out folds. Ceilings per fold and per-arm error. Section 6.8 reads the ceiling, the win rates and the transport baseline's lead off this table.

residual used	C_{const}	C_z	$C_{y,z}$
true $\delta = y - f(\theta^\star)$	0.2 ± 0.0	1.8 ± 0.1	72.4 ± 3.3
bootstrap residual	0.3 ± 0.1	2.1 ± 0.2	63.9 ± 2.1

Method	RMSE (°)	MAE	cov_{90}	wins vs. NPE	wins vs. constant
NPE	25.75 ± 5.81	20.74	0.19	0/28	18/28
RoPE-EMD†	23.05 ± 5.54	18.77	0.24	26/28	26/28
Constant shift	25.78 ± 5.82	20.78	0.19	10/28	0/28
MA-SBI-Z (kernel, z)	25.77 ± 5.82	20.75	0.18	12/28	13/28
MA-SBI-YZ (kernel, y, z)	25.75 ± 6.94	20.59	0.20	16/28	16/28
MA-SBI-YZ, bootstrap residual (no $\theta^\star$)	25.35 ± 6.70	20.20	0.21	18/28	18/28
Oracle (simulated input)	8.31 ± 3.88	6.00	0.84	28/28	28/28

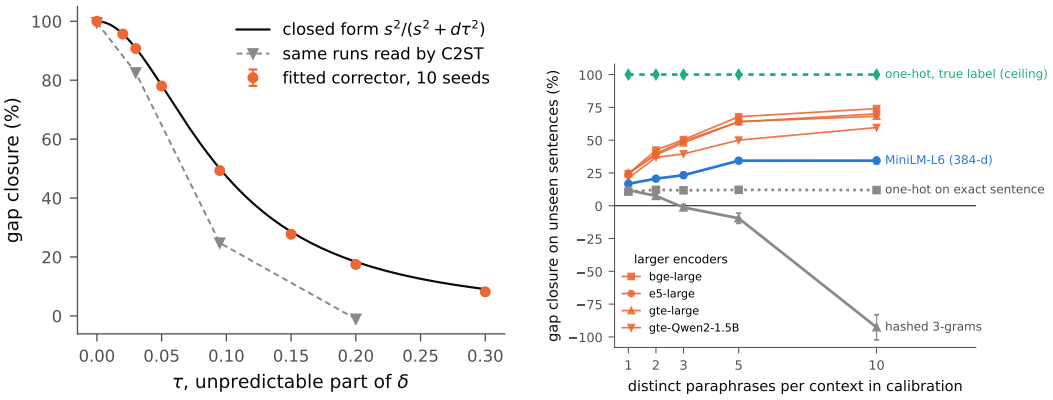


Figure 3: Left: gap closure under stochastic misspecification against the closed form, with the same runs read by C2ST to show where that metric saturates. Right: transfer to unseen sentences as calibration paraphrases increase, per encoder; the one-hot code over sentences is flat and the one-hot code given the true context is the ceiling.

A.2 Calibration gate and backbone agnosticism

The flow backbone passes simulation-based calibration on well-specified data on every task (Table 2), on three of three seeds except Bernoulli GLM, where it passes on two of three, with maximum mean-rank deviation 0.014 on Two Moons and 0.03 on Gaussian Linear. Under misspecification the deviation rises to 0.053 and 0.062 and the residual corrector returns it to 0.017 and 0.029.

The same gate is applied to a diffusion backbone. Unstandardised it fails outright ($p = 4 \times 10^{-70}$); retrained on standardised inputs with early stopping it passes on 18 of 20 task-seed cells, and the two that fail are dropped, not filled. On the cells that pass, the residual arm moves by at most 0.5 points between backbones and the additive arm by at most 8.1, on Gaussian Linear, where its across-seed standard deviation is 8.9 and both arms already exceed the neural oracle. The baselines move more, up to 19 points for the conditional transport component on SLCP, so backbone agnosticism is a claim about the corrector and not about the protocol (Table 8, Figure 5).

A.3 Encoder sweep

Replacing the 384-dimensional encoder with four larger ones raises held-out transfer at ten paraphrases from 34.4 to as much as 74.1 per cent against the lookup ceiling of 100 (Table 6), and the small encoder is flat from five to ten paraphrases while the best is still climbing. The sweep has two limits. It does not reach the ceiling, leaving the best encoder 26 points short. It also does not help transfer to unseen continuous z, where no encoder improves on the smallest one by more than 1.5

Table 5: Calibration size and misspecification mixture, median gap closure in δ space over ten seeds. Rows are n_{cal} and α_{mix}; Section 6.3 reads the separation off this grid.

		SLCP		Two Moons		Gaussian Linear		Bernoulli GLM	
α_{mix}	n_{cal}	MA-SBI-Z	MA-SBI-R	MA-SBI-Z	MA-SBI-R	MA-SBI-Z	MA-SBI-R	MA-SBI-Z	MA-SBI-R
0.0	100	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.9
0.0	500	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
0.0	2000	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
0.5	100	0.7	57.2	19.1	99.9	0.9	45.4	69.6	96.7
0.5	500	3.8	97.8	24.5	100.0	3.9	99.8	70.7	99.8
0.5	2000	4.3	99.6	24.9	100.0	4.4	99.9	70.7	100.0
0.9	100	−0.2	52.6	−0.9	99.9	−2.6	43.3	68.5	96.6
0.9	500	−0.2	91.8	4.8	100.0	−0.9	99.8	69.5	99.8
0.9	2000	−0.1	93.1	5.2	100.0	−0.2	99.9	69.5	100.0

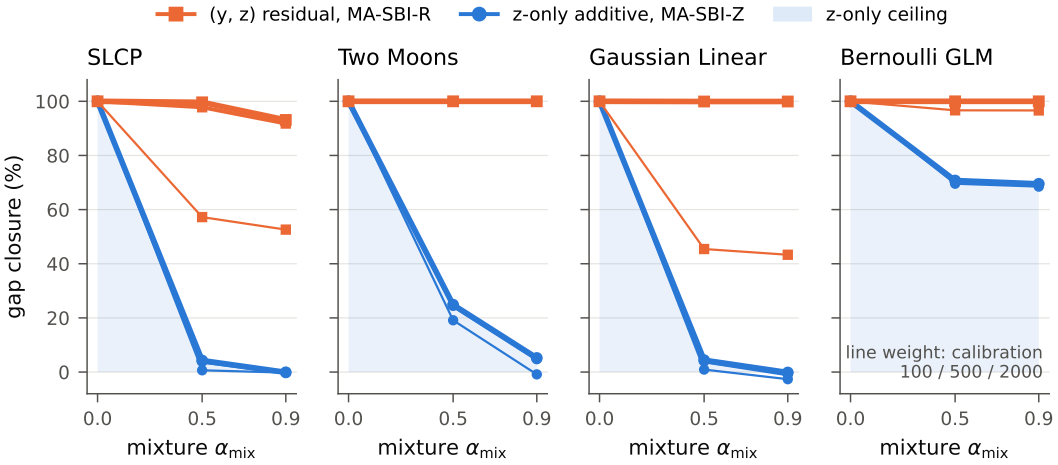


Figure 4: The full grid as curves. Median gap closure against α_{mix} for the two estimator arms, one line per calibration size. The shaded region is the z-only ceiling, which the additive arm tracks down as the shift becomes y-dependent. The residual arm holds near 100 except where calibration is short: at $n_{\text{cal}} = 100$ it falls to 53 on SLCP and 43 on Gaussian Linear at $\alpha_{\text{mix}} = 0.9$, and on SLCP at $\alpha_{\text{mix}} = 0.9$ even $n_{\text{cal}} = 2000$ reaches 93.

points; that limit is the corrector and the numeral representation, not encoder capacity. Scale is not the axis either: the 1536-dimensional instruction-tuned model is the weakest of the four large encoders.

A.4 Prior misspecification detail

The prior-perturbation result of Section 6.6 is consistent with Theorem 4 and a weak test of it, because the calibration targets are identically zero. Reference posteriors there are exact per-coordinate quadrature, not importance sampling, which makes the Student-t row interpretable at its C2ST floor of 0.499.

A.5 Graceful degradation in numbers

With an uninformative side channel the z-only corrector degrades to the best constant shift, whose gap closure $\|\bar{\delta}\|^2/\mathbb{E}\|\delta\|^2$, with $\bar{\delta} = \mathbb{E}\delta$, is 13.2 to 19.3 per cent across the four tasks. In δ-space the fitted additive corrector measures 12.3 to 17.9, never more than 0.6 points above that ceiling, which is within seed noise. Read in C2ST, as Table 7 does, the same arm sits at 4.3 to 18.6, below the ceiling on every task, so the corollary bounds the degradation and the metric decides how much of the bound is reached.

Table 6: Side-channel encoders. Gap closure on sentences and on z values the corrector never saw, for the encoder used throughout the paper and four larger ones. The one-hot row is the lookup ceiling: a code given the true context label, so it never reads the sentence and marks the ceiling rather than a transfer result.

		Held-out paraphrase		Held-out z band	
Encoder	d	k=1	k=10	linear	$\sin(2\pi z)$
all-MiniLM-L6-v2 *(baseline)*	384	16.7	34.4	98.0	90.6
e5-large-v2	1024	24.1	70.0	94.7	88.3
bge-large-en-v1.5	1024	24.5	74.1	98.4	89.9
gte-large	1024	25.0	68.2	98.3	90.7
gte-Qwen2-1.5B-instruct	1536	21.0	59.6	99.5	86.5
One-hot on the true context label *(ceiling)*	—	100.0	100.0	—	—

Table 7: The three evaluation conditions, median over seeds. Where the simulator is already right there is no gap to close, so the corrected C2ST is reported instead of a ratio to zero. Where the side channel is uninformative, each arm degrades to a different floor, which is why the two rows differ (Section 6.2).

condition	quantity	SLCP	Two Moons	Gaussian Linear	Bernoulli GLM
simulator well specified	NPE−Oracle C2ST	−0.001	−0.000	−0.002	0.000
	MA-SBI-R C2ST	0.734	0.505	0.517	0.551
misspecified, z informative	NPE−Oracle C2ST	0.118	0.475	0.074	0.373
	MA-SBI-Z gap closed	100.4	100.1	102.7	99.7
	MA-SBI-R gap closed	99.1	100.6	102.2	99.7
misspecified, z uninformative	NPE−Oracle C2ST	0.110	0.476	0.074	0.373
	MA-SBI-Z gap closed	4.3	8.6	18.6	7.7
	MA-SBI-R gap closed	11.1	5.4	24.8	40.3

A.6 The full grid

The grid of 360 cells over three calibration sizes and three mixtures at ten seeds shows the same separation throughout (Table 5 and Figure 4): at $\alpha_{\mathrm{mix}} = 0.5$ the additive arm's median falls to 13.8 against its own ceiling of 14.7 while the residual arm holds at 99.8. Cutting n_{cal} to 100 costs the residual arm only 97.0 against 100.0 at $\alpha_{\mathrm{mix}} = 0$. Where calibration is short the residual arm pays for its extra inputs: at $n_{\mathrm{cal}} = 100$ and $\alpha_{\mathrm{mix}} = 0.5$ it reaches 0.87 of its ceiling in Table 13 against 0.86 for the additive arm, and at $\alpha_{\mathrm{mix}} = 0.9$ it fails outright on two of five seeds, the direction Theorem 2 predicts for the linear case.

B Proofs

Throughout, (A1′) is the exogeneity condition of Section 3, $z \perp (\theta, \varepsilon)$ together with $(\delta, z) \perp \theta$ given y_{sim}, and (A2) is square integrability, $\mathbb{E}\|\delta\|^2 < \infty$ and $\mathbb{E}\|\theta\|^2 < \infty$. (GL) is the Gaussian-linear model of Section 5, $\theta \sim \mathcal{N}(0, \sigma_p^2 I_d)$ with $d_\theta = d$, $y_{\mathrm{sim}} = \theta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma_s^2 I_d)$, under which the well-specified posterior mean is $\mu(y) = \alpha y$ with $\alpha = \sigma_p^2/(\sigma_p^2 + \sigma_s^2)$. Every statement ranges over measurable h with $\mathbb{E}\|h\|^2 < \infty$. We distinguish the posterior-mean displacement $D(h) = \mathbb{E}\|\mu(y_{\mathrm{obs}} - h) - \mu(y_{\mathrm{sim}})\|^2$ from the excess parameter risk $\Delta\mathrm{MSE}(h)$ defined in Section 5.

B.1 Two lemmas

Lemma 7 (conditional-mean projection)**.** *Let W be a random element of any measurable space and $\delta \in L^2$. For every measurable h with $\mathbb{E}\|h(W)\|^2 < \infty$,*

$$\mathbb{E}\|\delta - h(W)\|^2 = \mathbb{E}\|\delta - \mathbb{E}[\delta \mid W]\|^2 + \mathbb{E}\|\mathbb{E}[\delta \mid W] - h(W)\|^2.$$

Hence $\mathbb{E}[\delta \mid W]$ is the almost surely unique minimiser, and for σ-fields $\mathcal{G}_1 \subseteq \mathcal{G}_2$, $\mathbb{E}\|\delta - \mathbb{E}[\delta \mid \mathcal{G}_2]\|^2 \leq \mathbb{E}\|\delta - \mathbb{E}[\delta \mid \mathcal{G}_1]\|^2$.

Table 8: Backbone agnosticism. Median gap closure on the flow and diffusion (DDPM) backbones over the same seeds, five per task and three on Bernoulli GLM after its calibration gate. The diffusion backbone carries a number only on cells where it passes simulation-based calibration on well-specified data.

	SLCP		Two Moons		Gaussian Linear		Bernoulli GLM	
Method	Flow	DDPM	Flow	DDPM	Flow	DDPM	Flow	DDPM
NPE (baseline)	0	0	0	0	0	0	0	0
RoPE-EMD†	16.6	11.6	7.8	10.3	13.0	6.3	31.3	33.6
Marginal component only	5.0	2.4	10.9	13.1	17.9	16.0	9.8	8.2
Conditional component only	92.3	73.1	23.2	23.4	76.5	84.2	59.5	65.8
MA-SBI-Z (additive, z only)	103.1	98.6	100.5	99.8	104.0	112.0	98.5	100.2
MA-SBI-R (residual)	99.1	98.8	100.4	99.9	105.8	105.5	100.2	100.3
Oracle (neural)	100	100	100	100	100	100	100	100

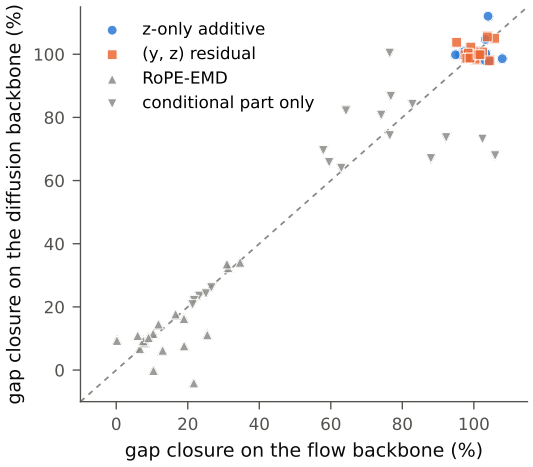


Figure 5: Every cell that passes the calibration gate, gap closure on the flow backbone against the same seed on the diffusion backbone. The two MA-SBI arms sit in the top-right corner on both; the baselines spread along the diagonal.

Proof. Write $\delta - h = (\delta - \mathbb{E}[\delta \mid W]) + (\mathbb{E}[\delta \mid W] - h(W))$. The second summand is $\sigma(W)$-measurable and the first has conditional mean zero given W, so the cross term is $\mathbb{E}[(\mathbb{E}[\delta \mid W] - h(W))^{\top}\mathbb{E}[\delta - \mathbb{E}[\delta \mid W] \mid W]] = 0$. The tower property gives monotonicity in the σ-field. □

An algebraic consequence used repeatedly is

$$C_z = \frac{\mathbb{E}\|\mathbb{E}[\delta \mid z]\|^2}{\mathbb{E}\|\delta\|^2}, \tag{5}$$

which follows from Lemma 7 with $h \equiv 0$.

Lemma 8 (transfer)**.** *Assume (A1′), (A2) and that $\mu(y) = \alpha y$ is the true well-specified posterior mean. Then for every measurable corrector h that is a function of (y_{obs}, z) with $\mathbb{E}\|h\|^2 < \infty$,*

$$D(h) = \alpha^2\,\mathbb{E}\|\delta - h\|^2 \qquad \textit{and} \qquad \Delta\text{MSE}(h) = \alpha^2\,\mathbb{E}\|\delta - h\|^2.$$

Proof. For D, $\mu(y_{\text{obs}} - h) - \mu(y_{\text{sim}}) = \alpha(\delta - h)$ by linearity. For the excess risk, let $u = \mu(y_{\text{sim}}) - \theta$. Then $\mu(y_{\text{obs}} - h) - \theta = u + \alpha(\delta - h)$ and $\Delta\text{MSE}(h) = 2\alpha\mathbb{E}[u^{\top}(\delta - h)] + \alpha^2\mathbb{E}\|\delta - h\|^2$. Now $\mathbb{E}[u \mid y_{\text{sim}}, \delta, z] = \alpha y_{\text{sim}} - \mathbb{E}[\theta \mid y_{\text{sim}}, \delta, z]$, and under (A1′) the pair (δ, z) is independent of θ given y_{sim}, so $\mathbb{E}[\theta \mid y_{\text{sim}}, \delta, z] = \mathbb{E}[\theta \mid y_{\text{sim}}] = \alpha y_{\text{sim}}$ and $\mathbb{E}[u \mid y_{\text{sim}}, \delta, z] = 0$. By the tower property $\mathbb{E}[u \mid y_{\text{obs}}, z] = 0$, since (y_{obs}, z) is a function of $(y_{\text{sim}}, \delta, z)$. Since h is (y_{obs}, z)-measurable, $\mathbb{E}[u^{\top}h] = 0$, and $\mathbb{E}[u^{\top}\delta] = \mathbb{E}[\mathbb{E}[u \mid y_{\text{sim}}, \delta, z]^{\top}\delta] = 0$. □

Table 9: Preconditioned robust NPE under the paper's protocol, median gap closed over five seeds, with preconditioning alone as its ablation. Hyperparameters selected on the well-specified condition; the selected setting still fails the well-specified sanity check on four of five SLCP seeds, which is recorded rather than hidden. The released implementation could not be run without changing the backbone the comparison depends on, so this is a documented reimplementation.

	SLCP		Two Moons	
Method	C2ST	Gap closed (%)	C2ST	Gap closed (%)
NPE	0.841	—	0.979	—
PNPE (preconditioning only)	0.861	-9.3 [-17.0, -8.1]	0.977	0.3 [0.0, 0.4]
PRNPE (Kelly et al., 2026)	0.864	-17.1 [-18.3, -13.9]	0.977	0.1 [0.1, 0.6]
MA-SBI-R (residual)	0.707	100.0 [99.8, 100.7]	0.510	100.1 [100.0, 100.1]
Oracle	0.708	—	0.511	—

Table 10: Light tunnel, leave-one-experiment-out, three seeds. Posterior-mean error in prior-range units against the recorded colour and polariser settings. The side-channel ceiling is low here and the corrector without parameter labels does not beat uncorrected estimation; the calibration-based baseline does.

method	needs θ^*	RMSE	beats NPE	beats RoPE	90% cov.
NPE	no	0.483 ± 0.078	0/15	4/15	0.04
RoPE-EMD†	yes	0.424 ± 0.047	11/15	0/15	0.06
constant shift†	yes	0.525 ± 0.046	7/15	0/15	0.01
MA-SBI-R, MLP†	yes	0.371 ± 0.051	12/15	13/15	0.12
MA-SBI-R, kernel†	yes	0.494 ± 0.055	10/15	1/15	0.01
MA-SBI-R, kernel, bootstrap residuals	no	0.492 ± 0.048	9/15	2/15	0.01
Oracle (simulator input)	no	0.180 ± 0.025	15/15	15/15	0.60

Lemma 8 is what lets Theorem 1(b) cover correctors that depend on y_{obs}. An independence argument covers only $h = h(z)$.

B.2 Proof of Theorem 1

Optimality of $\mathbb{E}[\delta \mid z]$ and the value C_z are Lemma 7 with $W = z$. Under (GL), Lemma 8 gives $\Delta\mathrm{MSE}(0) = \alpha^2\mathbb{E}\|\delta\|^2$ and $\Delta\mathrm{MSE}(h^\star) = \alpha^2\,\mathrm{MMSE}(\delta \mid z)$, whose normalised difference is C_z with the factor α^2 cancelling, which is why the ceiling is a pure δ-space quantity. Taking $W = (y_{\mathrm{obs}}, z)$ gives the second part, and $\sigma(z) \subseteq \sigma(y_{\mathrm{obs}}, z)$ with the monotonicity clause gives $C_{yz} \geq C_z$.

For the mutual-information corollary, take $\delta \sim \mathcal{N}(0, \sigma_\delta^2 I_d)$ jointly Gaussian with z and rotate (δ, z) to canonical variates, with $\rho_k = 0$ for $k > \dim z$. Isotropy is what makes the d canonical directions carry equal marginal variance, so $\mathrm{MMSE}(\delta \mid z) = \sigma_\delta^2 \sum_k (1-\rho_k^2)$ and $I(\delta; z) = -\frac{1}{2}\sum_k \log(1-\rho_k^2)$. Concavity of the logarithm gives $-2I(\delta; z) \leq d\log(d^{-1}\sum_k(1-\rho_k^2))$, that is $1 - C_z \geq e^{-2I/d}$, with equality exactly when all $1-\rho_k^2$ agree. Multiplying by $\alpha^2 d\sigma_\delta^2$ and using $1 - e^{-u} \leq u$ gives the familiar linear form $2\alpha^2\sigma_\delta^2 I(\delta; z)$.

Remark 9 (non-linear posterior-mean map). If μ is only L_μ-Lipschitz then $D(h) = \mathbb{E}\|\mu(y_{\mathrm{obs}} - h) - \mu(y_{\mathrm{sim}})\|^2 \leq L_\mu^2\,\mathbb{E}\|\delta - h\|^2$, since $(y_{\mathrm{obs}} - h) - y_{\mathrm{sim}} = \delta - h$. The companion bound $\mathbb{E}\|\mu(y_{\mathrm{obs}}) - \mu(y_{\mathrm{obs}} - h)\|^2 \leq L_\mu^2\,\mathbb{E}\|h\|^2$ is the one that limits how far a corrector can move the posterior mean, and neither says anything about direction. With a two-sided condition $\ell\|a-b\| \leq \|\mu(a)-\mu(b)\| \leq L_\mu\|a-b\|$ on the relevant set, $D(h^\star)/D(0) \leq (L_\mu^2/\ell^2)(1-C_z)$, so the *displacement* gap closure $1 - D(h^\star)/D(0)$ is at least $1 - (L_\mu^2/\ell^2)(1-C_z)$, which is vacuous unless L_μ/ℓ is near one. The same statement does not carry over to $\Delta\mathrm{MSE}$: the cross term $2\alpha\mathbb{E}[u^\top(\delta - h)]$ of Lemma 8 vanishes only when μ is the true posterior-mean map, and off that case it is unsigned. So C_z is the exact achievable gap closure for a linear posterior-mean map and a diagnostic for a flow backbone.

Remark 10 (C_{yz} is not free). $C_{yz} \geq C_z$ always, but fitting $\mathbb{E}[\delta \mid y_{\mathrm{obs}}, z]$ costs $(d_y + d_z + 1)/K$ times the calibration data by Theorem 2 and, without pairing, has no supervision target unless the

Table 11: The ceilings as a decision on the wind tunnel, per held-out fold and seed. Ceilings are computed from the four training experiments only; the predicted arm is the cheapest family within two points of the best ceiling; the actual arm is the lowest held-out error among constant, z-only and (y, z) correctors.

seed	held-out experiment	C_{const}	C_z	$C_{y,z}$	predicted	actual	agree
0	L_{out}=0.01, O_{dw}=4	60.3	61.4	94.2	(y, z)	(y, z)	✓
0	L_{out}=0.5, O_{dw}=2	60.6	65.9	94.9	(y, z)	(y, z)	✓
0	L_{out}=0.5, O_{dw}=4	56.5	61.6	93.4	(y, z)	(y, z)	✓
0	L_{out}=0.5, O_{dw}=8	57.2	61.9	94.8	(y, z)	(y, z)	✓
0	L_{out}=1, O_{dw}=4	58.9	64.3	94.4	(y, z)	z-only	×
1	L_{out}=0.01, O_{dw}=4	62.0	62.8	95.1	(y, z)	(y, z)	✓
1	L_{out}=0.5, O_{dw}=2	63.1	68.2	96.4	(y, z)	(y, z)	✓
1	L_{out}=0.5, O_{dw}=4	58.3	62.8	95.5	(y, z)	(y, z)	✓
1	L_{out}=0.5, O_{dw}=8	59.5	63.7	95.5	(y, z)	(y, z)	✓
1	L_{out}=1, O_{dw}=4	60.5	65.3	95.7	(y, z)	z-only	×
2	L_{out}=0.01, O_{dw}=4	63.3	64.2	94.4	(y, z)	(y, z)	✓
2	L_{out}=0.5, O_{dw}=2	63.8	68.9	94.9	(y, z)	(y, z)	✓
2	L_{out}=0.5, O_{dw}=4	59.1	63.8	93.7	(y, z)	(y, z)	✓
2	L_{out}=0.5, O_{dw}=8	59.6	64.0	94.1	(y, z)	(y, z)	✓
2	L_{out}=1, O_{dw}=4	62.0	67.1	93.8	(y, z)	z-only	×

Table 12: Wind tunnel, leave-one-experiment-out. Posterior-mean error against the hatch angle recorded with each trajectory, and how often each method beats plain NPE and beats RoPE on a paired per-fold comparison. Methods marked † build their calibration from simulator runs at the true parameter. The bootstrap variant uses none. Coverage is poor for every method because the simulator understates the real sensor noise, which an observation-space shift cannot repair.

method	needs θ^*	RMSE (deg)	beats NPE	beats RoPE	90% cov.
NPE	no	7.27 ± 3.84	$0/15$	$7/15$	0.08
RoPE-EMD†	yes	6.45 ± 1.59	$8/15$	$0/15$	0.10
constant shift†	yes	6.44 ± 2.14	$7/15$	$5/15$	0.07
MA-SBI-R, MLP†	yes	17.97 ± 2.07	$1/15$	$0/15$	0.00
MA-SBI-R, kernel†	yes	6.20 ± 2.04	$10/15$	$12/15$	0.07
MA-SBI-R, kernel, bootstrap residuals	no	6.23 ± 3.65	$15/15$	$10/15$	0.12
Oracle (simulator input)	no	0.05 ± 0.01	$15/15$	$15/15$	1.00

bootstrap surrogate is used. The difference $C_{yz} - C_z$ should be reported as a number rather than argued.

B.3 Proof of Theorem 2

For a fixed design matrix $X \in \mathbb{R}^{n \times p'}$ write $p = \operatorname{rank}(X)$ and let $\hat{\beta}_j = X^+ Y_{\cdot j}$ be the minimum-norm least-squares solution, so the fitted values are HY with $H = XX^+$ the orthogonal projector onto the column space of X and $\operatorname{tr} H = p$. With $\mathbb{E}[E_{\cdot j} E_{\cdot j}^\top] = \sigma^2 I_n$ the in-sample error of column j is $\mathbb{E}\|X(\hat{\beta}_j - \beta_j)\|^2 = \sigma^2 \operatorname{tr} H = \sigma^2 p$. Summing over d output coordinates and dividing by n gives in-sample excess risk $\sigma^2 dp/n$, which is the trace of the hat matrix and not a parameter count. Under a random design with independent sub-Gaussian rows, positive definite second moment of the p retained directions and $n \gtrsim p \log p$, the out-of-sample excess risk is $\sigma^2 dp/n\,(1 + o(1))$ with the correction of order $\sqrt{p \log p / n}$. These are the hypotheses under which the out-of-sample statement of Theorem 2 is read against the in-sample one.

For the additive family the design is one-hot, $X^\top X$ is diagonal with entries n_k, and the out-of-sample risk can be computed directly,

$$\mathbb{E}_z \|\hat{b}_z - b_z\|^2 = \sum_{k=1}^{K} \Pr(z = k) \frac{\sigma^2 d}{n_k} = \sum_{k=1}^{K} p_k \frac{\sigma^2 d}{n p_k} = \frac{\sigma^2 d K}{n},$$

under proportional (stratified) allocation $n_k = n p_k$, so allocating cells in proportion to the context frequencies costs nothing however unbalanced those frequencies are. The qualifier is needed: under independent and identically distributed sampling n_k is random and Jensen's inequality gives

Table 13: Achieved gap closure against each arm's own computed ceiling, in δ-space on SLCP over five seeds, as the misspecification mechanism moves from additive ($\alpha_{\text{mix}} = 0$) to multiplicative ($\alpha_{\text{mix}} = 0.9$). The unrestricted arm is $g(y, z) \to y_{\text{sim}}$ with no residual structure.

		additive		residual		unrestricted	
α_{mix}	n_{cal}	achieved	ceiling	achieved	ceiling	achieved	ceiling
0.0	100	100.0	100.0	100.0	100.0	92.5	100.0
0.0	500	100.0	100.0	100.0	100.0	98.0	100.0
0.0	2000	100.0	100.0	100.0	100.0	99.8	100.0
0.0	5000	100.0	100.0	100.0	100.0	99.9	100.0
0.5	100	32.5	37.6	87.0	100.0	92.2	100.0
0.5	500	37.0	37.6	99.0	100.0	98.6	100.0
0.5	2000	37.4	37.6	99.8	100.0	99.2	100.0
0.5	5000	37.3	37.6	100.0	100.0	99.5	100.0
0.9	100	-0.2	0.8	72.8	100.0	92.8	100.0
0.9	500	0.1	0.8	97.5	100.0	98.9	100.0
0.9	2000	0.5	0.8	99.5	100.0	99.2	100.0
0.9	5000	0.7	0.8	99.2	100.0	99.5	100.0

Table 14: Paraphrase transfer. Gap closure on held-out paraphrases against the number of training paraphrases per context, for the sentence encoder, a hash encoder, one-hot over exact sentence strings, and one-hot given the true context label as a control.

encoder	$k = 1$	$k = 2$	$k = 3$	$k = 5$	$k = 10$
sentence encoder	16.7	20.7	23.3	34.3	34.4
hashed 3-grams	11.9	7.5	-1.3	-9.5	-92.7
one-hot over exact sentences	10.7	12.2	11.7	12.1	12.0
one-hot given the true context	100.0	100.0	100.0	100.0	100.0

$\mathbb{E}[1/n_k] > 1/(np_k)$, so iid sampling costs strictly more than the stratified design. For the residual family $p = \text{rank}([y, z, 1]) \le d_y + d_z + 1$. Setting the two expressions equal gives the multiplier. If z is one-hot with $d_z = K$ the design $[y, z, 1]$ is rank deficient, since the one-hot columns sum to the intercept, and $p \le d_y + K$. More generally, K distinct embedded contexts give a z-block of rank at most K whatever the width of the encoder, so $p \le d_y + \min(d_z, K) + 1$ and a dense high dimensional encoding does not raise the multiplier above what the number of contexts allows.

For a non-linear corrector class the analogue of p is an effective dimension rather than a parameter count and no closed form is available, so Section 6.3 reports the ratio empirically as the horizontal shift between the two learning curves.

B.4 Proof of Theorem 3

One-step recovery. With $r(y) = y - m(y)$,

$$\delta^{\text{boot}} - \delta = y_{\text{sim}} - m(y_{\text{obs}}) = \underbrace{\big(y_{\text{sim}} - m(y_{\text{sim}})\big)}_{r(y_{\text{sim}})} + \underbrace{\big(m(y_{\text{sim}}) - m(y_{\text{obs}})\big)}_{\text{absorbed part of } \delta},$$

and the triangle inequality with $\|m(y_{\text{sim}}) - m(y_{\text{obs}})\| \le L_m \|\delta\|$ gives the stated bound. Under (GL) the decomposition is exact and $L_m = \alpha$ is attained termwise, $r(y_{\text{sim}}) = (1 - \alpha) y_{\text{sim}}$ and $m(y_{\text{sim}}) - m(y_{\text{obs}}) = -\alpha\delta$, but the triangle step itself is tight only when $(1 - \alpha) y_{\text{sim}}$ and $-\alpha\delta$ are aligned, so the bound is not an equality in general.

Gaussian-linear form and pooling. Under (GL) with $f(\theta, \varepsilon) = \theta + \varepsilon$ and $\mathbb{E}\varepsilon = 0$ we have $m(y) = \alpha y$, so $\delta^{\text{boot}} = (1 - \alpha)(y_{\text{sim}} + \delta)$. If $\mathbb{E}[y_{\text{sim}} \mid z] = 0$, which follows from (A1) and the zero-mean prior, then $\mathbb{E}[\delta^{\text{boot}} \mid z] = (1 - \alpha)\mathbb{E}[\delta \mid z]$. Per cell the nuisance term $(1 - \alpha) y_{\text{sim}}$ has per-coordinate variance $(1 - \alpha)^2(\sigma_p^2 + \sigma_s^2) = \sigma_s^4/(\sigma_p^2 + \sigma_s^2)$, so a cell of size n_z estimates the target with per-coordinate standard error $\sigma_s^2/\sqrt{n_z(\sigma_p^2 + \sigma_s^2)}$.

Table 15: Notation.

symbol	meaning
$\theta \in \mathbb{R}^{d_\theta}$, π	parameter and prior
$y_{\text{sim}} = f(\theta, \varepsilon) \in \mathbb{R}^d$	simulator output, ε its internal randomness
$\delta = y_{\text{obs}} - y_{\text{sim}}$	misspecification
$z \in \mathcal{Z}$	side information, encoded by $\phi(z)$
d_y, d_z	rank of the observation block and of the centred embedded calibration design; the encoder's nominal width is reported separately
α_{mix}	interpolation from an additive (0) to a multiplicative (1) misspecification mechanism
K	number of contexts, the indices of the misspecification mechanism
$q(\theta \mid y)$, $\mu(y)$	pre-trained amortised posterior and its mean
$m(y) = \mathbb{E}_{\theta \sim q(\cdot \mid y)} \mathbb{E}_\varepsilon f(\theta, \varepsilon)$	posterior-predictive mean map
$h(z)$, $h(y, z)$	corrector, applied as $q(\theta \mid y_{\text{obs}} - h)$
C_z, C_{yz}	z-only and (y, z)-conditional ceilings
$\alpha = \sigma_p^2/(\sigma_p^2 + \sigma_s^2)$	posterior shrinkage in the Gaussian-linear model
n_{cal}	number of calibration examples

Table 16: Named baselines under the same protocol, median gap closed over five seeds. Each baseline's hyperparameter is chosen on the well-specified condition only, never on the test cells, because the denoising baseline's default slab weight damages a problem with nothing to correct. The full sweep is in Table 17.

Method	SLCP	Two Moons
RNPE (Ward et al., 2022)	5.1	0.7
NPE-MMD (Elsemüller et al., 2025)	−1.9	0.1
MA-SBI-Z	99.2	100.0
MA-SBI-R	100.0	100.1
MA-SBI-Z on the NPE-MMD backbone	110.7	99.5

What the fitted corrector achieves. Dividing Lemma 7 by $\mathbb{E}\|\delta\|^2$ gives, for any z-indexed $\hat{h}$,

$$\mathrm{GC}(\hat{h}) = C_z - \frac{\mathbb{E}\|h^\star(z) - \hat{h}(z)\|^2}{\mathbb{E}\|\delta\|^2}. \tag{6}$$

For $\hat{h} = c\,h^\star$ with $c \in [0, 1]$, Eq. (6) gives $\mathrm{GC}(ch^\star) = C_z[1 - (1 - c)^2]$, and $c = 1 - \alpha$ gives $(1 - \alpha^2)C_z$. The recovered fraction exceeds the naive guess $1 - \alpha$ because gap closure is quadratic and a partially correct shift still removes the cross term.

Iteration. With $T(\delta) = y_{\text{obs}} - m(y_{\text{obs}} - \delta)$ we get under (GL) $T(\delta) = (1 - \alpha)y_{\text{obs}} + \alpha\delta$, an affine contraction with modulus α and unique fixed point $\delta^\star = y_{\text{obs}}$. From $\delta_0 = 0$, $\delta_t = (1 - \alpha^t)y_{\text{obs}}$, so the limit is degenerate, the corrected input being identically zero and the corrected posterior the prior whatever the data. The iteration is therefore not a refinement of the one-step estimate. It is not monotonically worse per sample either: at $d = 1$, $\alpha = 0.7353$, $y_{\text{sim}} = 0.1$ and $\delta = 1.6$, the second iterate is closer to δ than the first. What fails is the object the iteration converges to, not the first few steps. The variant $S(\delta) = (y_{\text{obs}} - \delta) - m(y_{\text{obs}} - \delta)$, which re-applies the residual operator to the already corrected input, is $(1 - \alpha)(y_{\text{obs}} - \delta)$ with fixed point $\frac{1-\alpha}{2-\alpha} y_{\text{obs}}$, a further shrunk value that is also not δ. T fixes y_{obs} for any m with $m(0) = 0$, so the iteration is structurally attracted to the explanation that all of the data is misspecification. The only cure is an external anchor. The exact identity $\delta - [m(y_{\text{obs}}) - m(y_{\text{obs}} - \delta)] = r(y_{\text{obs}}) - r(y_{\text{sim}})$ suggests debiasing the iteration by an estimate $\hat{r}_{\text{ws}}$ of the well-specified predictive bias, but the fixed point of $\delta_{t+1} = y_{\text{obs}} - m(y_{\text{obs}} - \delta_t) - \hat{r}_{\text{ws}}$ is $y_{\text{obs}} - \hat{r}_{\text{ws}}/(1 - \alpha)$ under (GL), which equals δ only if $\hat{r}_{\text{ws}} = r(y_{\text{sim}})$ for that observation. The anchor must therefore be per observation. A context-averaged anchor will not do: under (GL) with a zero-mean prior and z independent of y_{sim}, $\mathbb{E}[r(y_{\text{sim}}) \mid z] = (1 - \alpha)\mathbb{E}[y_{\text{sim}} \mid z] = 0$ in population, so the correction it applies is zero. The pairing-free regime cannot supply a per observation anchor, which is why we do not iterate.

Table 17: Baseline hyperparameter selection, made on the well-specified condition only. * marks a seed where no setting stayed within tolerance of plain NPE, in which case the best of the sweep is reported.

task	seed	NPE	RNPE C2ST by p	sel.	NPE-MMD C2ST by λ	sel.
SLCP	0	0.676	0.02: 0.693, 0.1: 0.743, 0.5: 0.826	0.02^*	0.1: 0.652, 1.0: 0.661, 10.0: 0.653	10.0
SLCP	1	0.718	0.02: 0.730, 0.1: 0.743, 0.5: 0.817	0.02^*	0.1: 0.737, 1.0: 0.732, 10.0: 0.727	10.0
SLCP	2	0.672	0.02: 0.695, 0.1: 0.741, 0.5: 0.832	0.02^*	0.1: 0.669, 1.0: 0.679, 10.0: 0.670	10.0
SLCP	3	0.668	0.02: 0.682, 0.1: 0.753, 0.5: 0.858	0.02^*	0.1: 0.659, 1.0: 0.664, 10.0: 0.667	10.0
SLCP	4	0.726	0.02: 0.727, 0.1: 0.739, 0.5: 0.828	0.02	0.1: 0.701, 1.0: 0.695, 10.0: 0.696	10.0
Two Moons	0	0.500	0.02: 0.503, 0.1: 0.553, 0.5: 0.781	0.02	0.1: 0.501, 1.0: 0.502, 10.0: 0.495	10.0
Two Moons	1	0.500	0.02: 0.507, 0.1: 0.565, 0.5: 0.809	0.02	0.1: 0.497, 1.0: 0.496, 10.0: 0.492	10.0
Two Moons	2	0.503	0.02: 0.507, 0.1: 0.553, 0.5: 0.792	0.02	0.1: 0.527, 1.0: 0.526, 10.0: 0.524	10.0^*
Two Moons	3	0.480	0.02: 0.516, 0.1: 0.552, 0.5: 0.776	0.02^*	0.1: 0.483, 1.0: 0.480, 10.0: 0.483	10.0
Two Moons	4	0.505	0.02: 0.500, 0.1: 0.566, 0.5: 0.806	0.02	0.1: 0.512, 1.0: 0.517, 10.0: 0.509	10.0

Algorithm 1 MA-SBI-R

1: **Input:** frozen amortised posterior q, calibration set, encoder ϕ
2: Estimate C_z and C_{yz} from held-out δ by regression on z and on (y_{obs}, z)
3: **if** $C_z \approx 0$ **then return** $q(\theta \mid y_{\mathrm{obs}})$ ▷ *no side channel value*
4: If δ is not observed, replace it by the pooled bootstrap residual of Eq. (3) or by a distribution-matching fit
5: Fit $\hat{h}$ by Eq. (4); drop the y input when $C_{yz} - C_z$ is small
6: **return** $q\big(\theta \mid y_{\mathrm{obs}} - \hat{h}(y_{\mathrm{obs}}, \phi(z))\big)$

General Lipschitz version. Writing $\hat{h}(z) = \mathbb{E}[\delta^{\mathrm{boot}} \mid z]$ and $e(z) = \|\mathbb{E}[r(y_{\mathrm{sim}}) \mid z]\| + L\,\mathbb{E}[\|\delta\| \mid z]$, Eq. (6) gives $\mathrm{GC}(\hat{h}) \geq C_z - \mathbb{E}_z[e(z)^2]/\mathbb{E}\|\delta\|^2$. Both the cell-mean signed residual and L_m are estimable from simulations alone, which makes this a pre-flight criterion rather than a post-hoc explanation. It is informative only when C_z exceeds roughly L_m^2, since the absorbed term alone contributes L_m^2 to the subtracted quantity: at $L_m = 0.735$ the criterion says nothing unless $C_z > 0.54$.

B.5 Proof of Theorem 4

If $\delta = 0$ almost surely then $\mathbb{E}[\delta \mid z] = 0$ and the objective $\mathbb{E}\|h(z)\|^2$ is uniquely minimised at $h \equiv 0$, so the corrected posterior equals the uncorrected one. The prior does not appear in the objective, so no information about prior misspecification carried by z can enter. When both are wrong, the corrector acts only on the conditioning argument of a fixed $q_{\pi_{\mathrm{sim}}}$, so it moves $q_{\pi_{\mathrm{sim}}}(\theta \mid y_{\mathrm{obs}})$ toward $q_{\pi_{\mathrm{sim}}}(\theta \mid y_{\mathrm{sim}})$ and cannot change π_{sim}.

The finite-n statement needs the mean-well-specified case rather than the literal hypothesis. Suppose only $\mathbb{E}[\delta \mid z] = 0$, with residual variance σ^2 per coordinate. The population-optimal corrector is still $h^\star \equiv 0$, but the fitted one is not exactly zero, with $\mathbb{E}\|\hat{h}\|^2 = \sigma^2 dK/n$ by Theorem 2, so a mean-well-specified simulator incurs excess risk $\alpha^2\sigma^2 dK/n$ from correcting: harmless in population and mildly harmful at small n. Under the literal hypothesis $\delta = 0$ almost surely the residual noise is zero, $\sigma^2 = 0$, and even that term vanishes, which is why in the experiments the fitted norm is zero to three decimals, the calibration targets being identically zero rather than merely mean zero.

B.6 Proof of Theorem 5

Regularity. (R1) $0 < Z(y, z) < \infty$ for almost every y and every z in the support of z. This is the only condition the argument uses.

Well-posedness. Fix (y, z) with $0 < Z(y, z) < \infty$, which holds almost everywhere by (R1). The numerator $p(y \mid \theta, z)\pi(\theta)$ is a fixed function of θ determined by the observed pair and the denominator is a finite positive constant in θ, so $\tilde{q}(\cdot \mid y, z)$ is a uniquely determined density.

Sufficiency. The law of z does not depend on θ, so the joint law of (y, z) at parameter value θ is $P_z(dz)\, p(y \mid \theta, z)\, dy$. Suppose slice injectivity holds for P_z-almost every z and let $\theta_1 \neq \theta_2$ induce the same joint law. Then $p(\cdot \mid \theta_1, z) = p(\cdot \mid \theta_2, z)$ for P_z-almost every z, and any such z outside

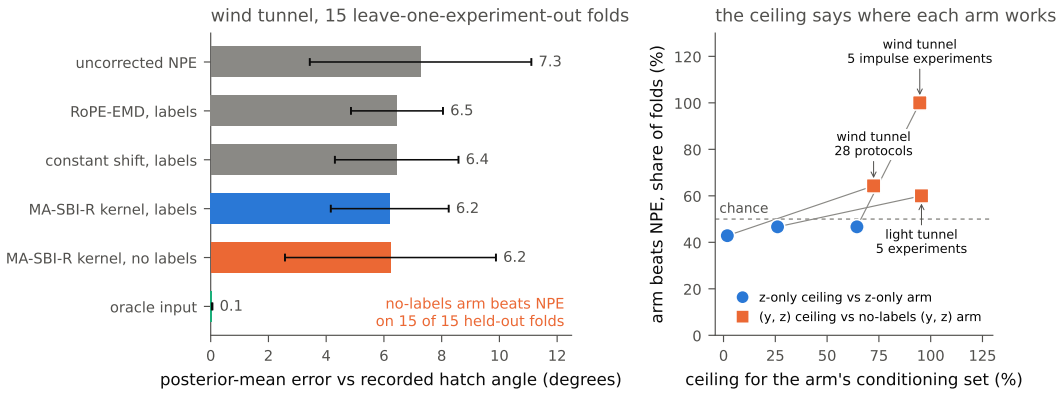


Figure 6: Real data. Left: wind tunnel, posterior-mean error against the recorded hatch angle, mean and standard deviation over 15 held-out folds; orange is the variant that uses no parameter labels. Right: the three physical datasets, each arm's share of folds beating uncorrected NPE against the ceiling computed for that arm's conditioning set. The z-only arm sits at chance wherever its ceiling is low, and the no-labels (y, z) arm wins where the (y, z) ceiling is high.

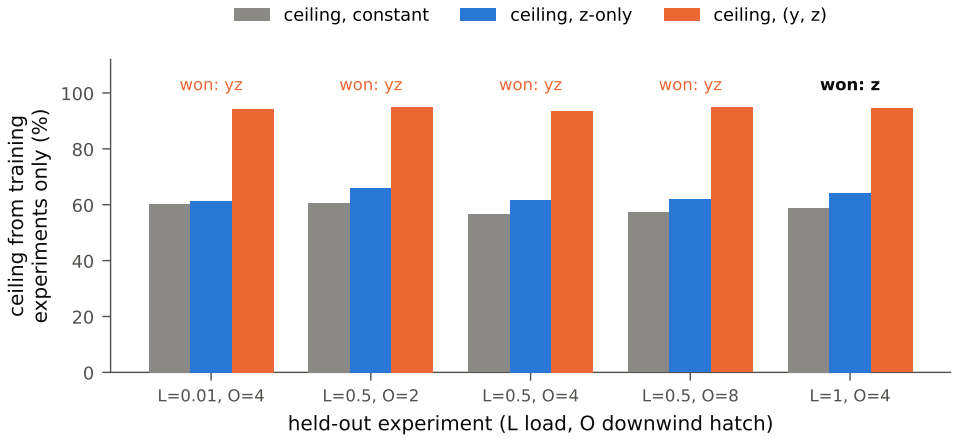


Figure 7: The ceiling as a decision on the wind tunnel, seed 0. For each held-out experiment the three ceilings are computed from the four training experiments alone; the arm they select is the (y, z)-conditioned one on every fold, and the label above each group is the arm that actually won.

the null set contradicts injectivity of $\theta \mapsto p(\cdot \mid \theta, z)$. Hence $\theta_1 = \theta_2$ and θ is identifiable from the joint law of (y, z). Injectivity of the augmented map $(\theta, z) \mapsto (p(\cdot \mid \theta, z), z)$ is the same condition written without the almost-everywhere qualifier, since the second coordinate already separates distinct contexts.

Necessity. Suppose slice injectivity fails on a set of contexts of positive P_z measure, with $\theta_1 \neq \theta_2$ and $p(y \mid \theta_1, z) = p(y \mid \theta_2, z)$ for almost every y and every z in that set. Restricted to it the two parameter values induce identical conditional laws of the observation, the likelihood ratio is constant in y, and no measurable function of (y, z) can separate them at any sample size. A failure on a P_z-null set is invisible in the joint law and does not obstruct identifiability, which is why the positive-measure qualifier is needed.

A criterion stated on the unaugmented map $(\theta, z) \mapsto p(\cdot \mid \theta, z)$ is sufficient but not necessary, because z is observed and a collision across contexts costs nothing.

Remark 11 (z shadowing θ). If $z = g(\theta)$ for injective g then slice injectivity holds vacuously, each slice containing one θ, but the posterior is a point mass determined by z alone and the observation

Table 18: Sample efficiency at matched ceilings, additive mechanism. Calibration size at which each arm first reaches 90 per cent of its own ceiling, and the ratio between them, against the multiplier from the rank of the embedded calibration design. Five seeds. n_{add} and n_{res} are the grid points at which each arm first crosses 90 per cent; the fitted ratio comes from a log-linear interpolation of the two learning curves and therefore differs slightly from the ratio of the tabulated grid points. The multiplier from nominal encoder width would be 18.2 and 98.2.

encoder	nominal d_z	effective d_z	n_{add}	n_{res}	fitted ratio	predicted
hash	64	3	9	25	2.91	3.00
sbert	384	3	8	15	1.79	3.00
sbert para	384	39	74	101	1.36	12.00

Table 19: Stochastic misspecification. Measured posterior-mean gap closure against the closed form $s^2/(s^2 + d\tau^2)$, ten seeds.

τ	predicted	measured	residual (pp)	C2ST gc
0.0000	100.00	100.00 ± 0.00	-0.00	99.6
0.0200	95.74	95.66 ± 0.04	-0.08	—
0.0300	90.91	90.77 ± 0.13	-0.14	82.6
0.0500	78.26	78.02 ± 0.11	-0.24	—
0.0949	49.98	49.35 ± 0.25	-0.63	24.8
0.1500	28.57	27.72 ± 0.43	-0.85	—
0.2000	18.37	17.47 ± 0.17	-0.90	-1.0
0.3000	9.09	8.18 ± 0.35	-0.91	—

Table 20: Continuous z with three deleted bands, evaluated only inside the deleted bands.

$\delta(z)$	encoder	in-distribution	held-out band
linear	sbert	99.7	98.0 ± 0.2
linear	hash	98.4	52.2 ± 4.8
sine	sbert	99.0	90.6 ± 0.8
sine	hash	93.8	-53.4 ± 8.7

channel is bypassed. This is why the exclusion screen, regressing the fitted correction on y and on θ, is a separate precondition rather than a consequence of the theorem. The screen gives evidence against shadowing; it does not verify exogeneity.

B.7 Proof of Theorem 6

Let $P = P_\delta$ and, for ζ in the support of z, let $Q_\zeta = P_{\delta|z=\zeta}$ and $D_\zeta = D_{\mathrm{KL}}(Q_\zeta \| P)$. By Donsker and Varadhan, $\mathbb{E}_{Q_\zeta}[g] \le \log \mathbb{E}_P[e^g] + D_\zeta$ for every measurable g with finite exponential moment. Take $g = t\, u^\top(\delta - \mathbb{E}_P\delta)$ for a unit vector u and $t > 0$; the sub-Gaussian condition gives $\log \mathbb{E}_P[e^g] \le t^2\tau^2/2$, so with $a_\zeta(u) = u^\top(\mathbb{E}[\delta \mid \zeta] - \mathbb{E}[\delta])$, $t\, a_\zeta(u) \le t^2\tau^2/2 + D_\zeta$ for all $t > 0$. Minimising the right side over t gives $a_\zeta(u) \le \sqrt{2\tau^2 D_\zeta}$. Choosing u along $\mathbb{E}[\delta \mid \zeta] - \mathbb{E}[\delta]$ and squaring gives $\|\mathbb{E}[\delta \mid \zeta] - \mathbb{E}[\delta]\|^2 \le 2\tau^2 D_\zeta$, and averaging over z with $\mathbb{E}_z[D_z] = I(\delta; z)$ gives the claim. The L_μ-Lipschitz statement follows pointwise.

Steps one to four are the dual form of the Bobkov and Götze transport inequality, so a sub-Gaussian proxy is equivalent to $W_1(Q, P) \le \tau\sqrt{2D_{\mathrm{KL}}(Q\|P)}$ and Kantorovich duality turns control of W_1 into control of mean shifts. A strong log-concavity route through Bakry and Émery gives the same constant under a strictly stronger hypothesis and is not needed.

The apparent counterexample. Take $\delta = s\, m\mathbf{1}_d + W$ with s uniform on $\{\pm 1\}$ and $W \sim \mathcal{N}(0, \sigma^2 I_d)$, at $d = 5$, $m^2 = 0.144$, $\sigma^2 = 0.016$, where a measured $I(\delta; z) = 0.719$ nats accompanies $\operatorname{tr}\operatorname{Var}(\mathbb{E}[\delta \mid z]) = 0.395$. Since $\mathbb{E}[e^{tu^\top(\delta - \mathbb{E}\delta)}] = \cosh(t\, m\, u^\top\mathbf{1}_d) e^{t^2\sigma^2\|u\|^2/2}$ and $\sup_{\|u\|=1}(u^\top\mathbf{1}_d)^2 = d$, the smallest proxy valid in every direction is $\tau^2_{\mathrm{eff}} = m^2 d + \sigma^2 = 0.736$, not the per-coordinate variance 0.16. The bound then reads $2\tau^2_{\mathrm{eff}} I(\delta; z) = 1.058 \ge 0.395$, a ratio of 0.373. The mechanism the counterexample appealed to is real, a bimodal conditional mean swinging between modes at low divergence cost, but the swing is along $\mathbf{1}_d$, which is exactly the direction in which the honest proxy is d times larger. Bounded noise with $\|\delta\| \le M$ satisfies the condition with $\tau^2 \le M^2$, so the bounded case is subsumed rather than separate.

Table 21: Prior misspecification with exact quadrature posteriors. Fitted correction norm and gap closure under four prior perturbations with a well specified simulator, and under both misspecified.

simulator	test prior	$\|\hat{\delta}\|$	NPE	MA-SBI	sim. oracle
well specified	none	0.001	0.498	0.500	0.498
well specified	mean shift	0.000	0.998	0.998	0.998
well specified	scale3	0.001	0.977	0.977	0.977
well specified	bimodal	0.001	0.990	0.990	0.990
well specified	student t3	0.000	0.696	0.696	0.696
misspecified	none	0.300	0.812	0.499	0.500
misspecified	mean shift	0.300	0.998	0.998	0.998
misspecified	scale3	0.300	0.979	0.974	0.973
misspecified	bimodal	0.300	0.991	0.990	0.990
misspecified	student t3	0.300	0.851	0.678	0.678

Table 22: Negative controls. Two real datasets where the method should not help, and does not. On COVID case counts the correction gives no gain once the case level is made identifiable, and on influenza surveillance the narrative passes the leakage screen but carries no information about the residual, so its ceiling proxy sits at zero. The well-specified cognitive dataset of Section 6.8 returns a fitted correction norm of zero and has no row because there is no criterion to fail.

dataset	what fails	criterion	value
COVID + OxCGRT	no gain once the case level is identifiable	median NLL change	-0.7%
		win rate	0.42
ILINet + FluView	narrative does not index the residual	C_z proxy, held-out season	-0.011
		leakage R^2	0.051

B.8 COROLLARIES

Corollary 12 (uninformative side channel)**.** *If $\delta \perp z$ then $\mathbb{E}[\delta \mid z] = \bar{\delta} := \mathbb{E}[\delta]$, the population-optimal corrector is the constant $\bar{\delta}$, and under (GL) the excess risk after correction is $\alpha^2 \operatorname{tr}\operatorname{Var}(\delta) \le \alpha^2 \mathbb{E}\|\delta\|^2$, so the correction never hurts in population and improves by exactly $\alpha^2\|\bar{\delta}\|^2$. The gap closure is $\|\bar{\delta}\|^2/\mathbb{E}\|\delta\|^2$, which is not zero unless δ is centred.*

The non-negativity here comes from Lemma 7, not from an upper bound. A one-sided bound tending to zero does not by itself say that a correction cannot hurt. At finite n, with $\hat{\bar{\delta}}_n$ the sample mean of n calibration shifts, $\mathbb{E}\|\hat{\bar{\delta}}_n - \bar{\delta}\|^2 = \operatorname{tr}\operatorname{Var}(\delta)/n$, so by Eq. (6) the expected gap closure is $(\|\bar{\delta}\|^2 - \operatorname{tr}\operatorname{Var}(\delta)/n)/\mathbb{E}\|\delta\|^2$, which is negative whenever $n < \operatorname{tr}\operatorname{Var}(\delta)/\|\bar{\delta}\|^2$. That is a variance effect with a closed form, not a threshold in mutual information.

Corollary 13 (oracle recovery)**.** *If $\mathrm{MMSE}(\delta \mid z) \to 0$ then $C_z \to 1$ and, under (GL), the corrected posterior mean attains the well-specified risk. For an estimated corrector, Eq.* (6) *gives $\mathrm{GC}(\hat{h}) = 1 - [\mathrm{MMSE}(\delta \mid z) + \mathbb{E}\|h^\star - \hat{h}\|^2]/\mathbb{E}\|\delta\|^2$, so full recovery needs the estimation error to vanish as well.*

The hypothesis is a statement about the minimum mean squared error, not a limit of mutual information. Letting $I(\delta; z) \to \infty$ makes the exponential bound vacuous rather than tight, and nothing about the achieved reduction follows from it.

Remark 14 (relation to per-index calibration). With one-hot calibration indices as z, the corrector reduces to a per-index shift, which is the correction a calibration-based method recovers from parameter pairs in the same setting. We state this as a reduction and not as a generalisation theorem. In the structural-permutation benchmark the calibration-based method dominates, which is the empirically correct ordering.

B.9 Remarks on the theorems

Theorem 1, the Gaussian clause. The clause recovers the mutual-information bound as a corollary: $1 - e^{-u} \le u$ turns it into the familiar linear form $C_z \le 2I(\delta; z)/d$.

Theorem 2, what counts as d_z. With K distinct sentences the centred design has rank at most $K-1$ whatever the embedding dimension, so a 384-dimensional sentence encoder and a 64-dimensional hash both give an effective d_z of 3 at $K = 4$, and with $d_y = 8$ a multiplier of $(8 + 3 + 1)/4 = 3.0$ rather than the 98.2 and 18.2 their nominal widths suggest. The theorem is exact for least squares and only an upper bound on the multiplier measured for the neural arms of Section 6.3, whose fitted ratios of 2.91, 1.79 and 1.36 all fall below the rank-based 3.0, 3.0 and, at ten paraphrases per context where the rank reaches 39, $(8 + 39 + 1)/4 = 12.0$. The bound loosens as the rank grows. Once $C_{yz} > C_z$ the additive form carries an irreducible bias and the residual form wins past a crossover that is computable from the same held-out data.

Theorem 3, what the deflation costs. At $\alpha = 0.735$ the bootstrap corrector reaches 46 per cent of C_z, at $\alpha = 0.3$ it reaches 91 per cent, and as $\alpha \to 1$ it reaches nothing. Dividing the pooled residual by the known factor $1 - \alpha$ attains C_z in population, at the cost of inflating the cell-mean standard error by $(1 - \alpha)^{-1}$, which is why we do not rescale when cells are small.

Theorem 5, the instrumental variable analogy. A criterion stated on the unaugmented map is sufficient but not necessary. The three preconditions of the method map onto the instrumental variable triple of exogeneity, exclusion and relevance (Angrist et al., 1996): exogeneity is (A1′), exclusion is the screen that regresses the fitted correction on y and on θ, and relevance corresponds to $C_z > 0$. The analogy is expository and no result from that literature is used in any proof.

Theorem 6, a worked example. For the mixture $\delta = s\, m\mathbf{1}_d + W$ with $s \sim \mathrm{Unif}\{\pm 1\}$, the smallest valid proxy is $\tau^2_{\text{eff}} = m^2 d + \sigma^2$, not the per-coordinate variance, and the bound then holds with a ratio of 0.373 rather than being violated by 1.72.

B.10 Numerical checks

Monte Carlo with 4×10^6 draws, $d = 1$, $\sigma_p^2 = 0.25$, $\sigma_s^2 = 0.09$ so $\alpha = 0.7353$, and $K = 4$ equiprobable contexts with means $(-0.4, -0.1, 0.2, 0.5)$ and within-context shift standard deviation 0.2, gives excess-risk gap closure 0.74212 against a predicted $C_z = 0.74218$ at $h^\star(z)$, 0.76933 against a predicted $C_{yz} = 0.76929$ at $h^{\star\star}(y, z)$, and 0.34085 against a predicted $(1 - \alpha^2)C_z = 0.34092$ at $(1 - \alpha)h^\star$. The predicted column is a plug-in evaluated on the same draws, not the closed form. The closed forms for this configuration are $C_z = 0.115/0.155 = 0.74194$, $C_{yz} = 1 - (0.04 \cdot 0.34/0.38)/0.155 = 0.76910$ and $(1 - \alpha^2)C_z = 0.34080$, which the reported pairs match to about 2×10^{-4}, the Monte Carlo scale at this sample size. The second pair is the one that matters, since it confirms that the cross term of Lemma 8 vanishes for a corrector that depends on y_{obs}. Iterating T two hundred times from $y_{\text{obs}} = 1.7$ returns 1.70000 and iterating S returns 0.355814, both as predicted.

C Experimental details

Table 23 lists the configuration and the datasets. Hyperparameters, per-benchmark simulator settings, paraphrase pools, encoder choices and wall-clock costs are collected here and are also emitted by the table-generation script into the supplementary material.

Table 23: Configuration and data sources, read from the result files.

component	setting
Backbone	standardised MAF, 8 transforms, 128 hidden, embedder 32-d
Backbone training	50k simulations, Adam 5e-4, batch 512, early stopping patience 25
Corrector	3-layer MLP, GELU, 128 hidden; 300 epochs, batch 128, Adam 1e-3, early stopping
Side-channel encoder	all-MiniLM-L6-v2, 384-d, unit norm; hashed char 3-grams 64-d as ablation
Misspecification scale	$\|\delta\|$ set to 0.25 of the per-task observation scale
Evaluation	per cell 3 observations $\times$ 4 contexts, C2ST with 5 folds; n_{cal} as stated per table
Benchmark data	sbibm reference posteriors for SLCP, Two Moons, Gaussian Linear, Bernoulli GLM
Physical data	Causal Chamber wind tunnel (wt_intake_impulse_v1, wt_walks_v1) and light tunnel (lt_camera_v1)
Epidemiological data	COVID case counts with OxCGRT policy text; CDC ILINet with FluView narratives
Wind tunnel	400 trajectories per experiment, 5-fold leave-one-experiment-out, kernel corrector temperature 0.1