MaSI: Mechanism-aware Interval Inference with LLM Semantic Covariates

Anonymous authors
Paper under double-blind review

Abstract

Insufficient sample size is a primary obstacle to precise inference, especially when observations pair text covariates with scalar outcomes. Unlike structured covariates with prespecified coordinates and interpretable perturbations, text lacks a canonical low-dimensional representation, and routine edits need not preserve outcome-relevant semantics. Modern large language models can generate source-conditioned texts, creating a new possibility for addressing this difficulty. Yet distribution shift and within-source dependence can make naive use biased and overconfident. We introduce Model-assisted Semantic Inference (MaSI), which uses generated texts to construct a more precise confidence interval for the parameter of interest while controlling the inferential effects of LLM-induced distribution shift under the stated calibration conditions. Specifically, the first set is generated under a scheme that preserves source content while varying expression style; together with real labeled data, it is used to learn outcome-relevant content coordinates and complementary style coordinates. A separate evaluation set constructs a calibrated auxiliary statistic in those coordinates. Density-ratio calibration addresses distribution shift, real-label residual correction preserves the population target, and uncertainty is estimated across original sources. Under explicit representation, calibration, and covariance-control conditions, MaSI provides consistent variance estimation and asymptotically valid confidence intervals. Simulations and empirical-data experiments show shorter intervals with near-nominal coverage in the primary settings, establishing a principled route to more precise inference from scarce text–outcome pairs.

Keywords: prediction-powered inference; uncertainty quantification; content–style representation learning; LLM semantic augmentation

1 Introduction

Many scientific and policy studies use free-form text paired with externally defined outcomes to estimate population summaries. Such summaries support public-health monitoring, social measurement, survey research, and content assessment; because substantive conclusions may hinge on differences comparable to sampling variation, valid confidence intervals are as important as point estimates. Sample scarcity has an additional consequence here: unlike structured covariates with prespecified coordinates and interpretable perturbations, text must first be mapped to a learned representation, while its discrete input space makes augmentations that preserve task-relevant semantics difficult to construct (Feng et al., 2021). We study a setting with only a modest sample of real text–outcome pairs (X, Y) and no separate, much larger pool of genuine unlabeled X from the target population. A given LLM can nevertheless generate source-conditioned variations of each observed text, exposing semantic regularities that one wording leaves implicit and creating a new possibility for addressing this difficulty. These generations are not new population draws. We ask how to use them to construct a more precise confidence interval for the real-population outcome mean while correcting generation shift, accounting for within-source dependence, and retaining nominal coverage.

Model-assisted estimators combine population-level auxiliary information with correction from sampled observations (Breidt & Opsomer, 2017), while semi-supervised estimators exploit additional genuine unlabeled covariates (Zhang & Bradic, 2022; Deng et al., 2024). Prediction-powered inference (PPI) specializes this principle to a labeled sample and a typically larger, independent pool

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.

of genuine unlabeled covariates from the target population (Angelopoulos et al., 2023a; Zrnic & Candès, 2024; Angelopoulos et al., 2023b; Fisch et al., 2024). Predictions on that pool estimate a population predictive mean, while real labels correct prediction error. Here no such pool is observed: every generated text is conditional on one of the same n sources. Generated texts therefore cannot substitute for independent population draws or increase n. The opportunity is instead to extract more outcome-related information from each observed source.

Synthetic-text augmentation can improve predictor training (He et al., 2022; Halterman, 2025), and LLM annotations can support inference when their errors are validated or corrected using sampled human labels (Egami et al., 2023; Gligorić et al., 2025). These workflows address predictor construction or imperfect outcome surrogates. Predictive gains from augmentation do not by themselves provide the correction and variance estimator needed for population inference; annotation-based inference instead begins with genuine target documents and treats the LLM output as a noisy label. We generate covariates but do not treat LLM-generated labels as new outcome observations. Joint calibration of generated (X, Y) pairs is possible in principle, but it requires the generated joint law to cover the real joint law sufficiently well and requires learning the LLM's conditional labeling bias and a high-dimensional joint density ratio from the limited real text–outcome pairs. These requirements may fail when generated labels are nearly deterministic, conditional support is inadequate, or generated records remain dependent within source; increasing the number of generations reduces Monte Carlo error but does not remove systematic label bias that can invalidate confidence-interval coverage. MaSI instead adopts the more parsimonious structure needed to identify the population outcome mean: generated texts provide auxiliary covariate information, a low-dimensional density ratio calibrates the coordinate distributions, and real-label residuals recover the target. The method therefore does not rely on the correctness of the LLM's conditional labeling mechanism and yields valid confidence intervals under weaker and more transparent conditions.

Recent frameworks also assign synthetic records, generated contexts, or AI-derived features explicit inferential roles. Methods for imperfect synthetic data and synthetic-powered inference study settings in which generated records supplement real samples (Byun et al., 2025; Bashari et al., 2025; 2026); context- and generative-augmentation methods use generated information to construct features or records for testing and regression (Ratkovic, 2025; Lu et al., 2026). These methods establish useful roles for synthetic records or generated features, whereas here multiple texts are transformations of one sampled source, may differ systematically from real text, and remain dependent within source. This source-group structure couples three inferential tasks: deciding which semantic variation enters the outcome model, calibrating evaluation texts to real text, and aggregating uncertainty over original sources.

We propose Model-assisted Semantic Inference (MaSI), which assigns two source-conditioned generation sets distinct roles. Representation texts are designed to vary expression while preserving source content and, together with real labels, support coordinates that are stable to expression yet relevant to Y (Wu et al., 2023; Chen et al., 2020; von Kügelgen et al., 2021). Separate evaluation texts are then mapped into this expression-stable, outcome-relevant semantic space, and density-ratio calibration adjusts their coordinate distribution toward that of real text (Shimodaira, 2000; Huang et al., 2007; Kouw & Loog, 2021). MaSI averages their calibrated scores within source and cross-fits the resulting auxiliary statistic with real-label residual correction. Generated texts therefore contribute only through their observed source.

This design addresses three statistical difficulties. First, generation can change expression and outcome-related content simultaneously, so invariance alone may erase useful signal; the representation must suppress expression-specific variation without discarding outcome relevance. Second, evaluation texts need not follow the real-coordinate distribution, so their scores require calibration before averaging. Third, repeated texts from a common source remain dependent, and treating text-level records as independent in variance estimation would create spurious sampling evidence. MaSI meets these requirements through constrained representation learning, density-ratio calibration, and within-source aggregation, while cross-fitting separates nuisance learning from held-out evaluation.

Our contributions are threefold. First, we account for dependence among texts generated from the same source and distributional shift between generated and real text, and develop an estimand-preserving, cross-fitted estimator that exploits their semantic information without an external unlabeled pool or synthetic outcomes. Second, under the stated regularity and covariance-control conditions, we establish an asymptotically linear representation around a deterministic population

limit and consistency of the source-group variance estimator, yielding asymptotically valid $1-\alpha$ confidence intervals; additional results control the error introduced by learning the representation. Third, simulations, controlled text generation, and real-data experiments assess both inferential performance and the proposed mechanism, characterizing when generated information improves precision and how representation shift, calibration, and same-source dependence affect the gain. Together, MaSI extends model-assisted mean inference to data-limited settings without a large independent pool of genuine unlabeled covariates: a given LLM can expose useful semantic structure in each observed source, while real data continue to determine the estimand and independent evidence.

2 Problem Setup

Let $(X, Y) \sim P$ be a real observation, where X contains the text and any source context supplied to the generator and $Y \in \mathbb{R}$ is the externally defined outcome. From n i.i.d. labeled units $\{(X_i, Y_i)\}_{i=1}^n$, we target

$$\mu = \mathbb{E}_P(Y).$$

Here μ is the marginal mean under the real-data law P, not the conditional mean given X. The conditional mean can serve as an ideal outcome predictor, and its population average equals μ by iterated expectation (Angelopoulos et al., 2023a;b). For each unit, two prespecified role-specific generation schemes produce m Q_{style} texts intended to preserve source content while varying expression and m separate Q_{eval} texts that may expose outcome-relevant features not captured by the working coordinates of the original text. All source information supplied to the generator is included in X_i. The generator receives X_i, but not Y_i. The resulting texts can nevertheless remain predictive of Y_i by preserving or exposing outcome-relevant information contained in X_i; they provide model-derived auxiliary structure, not additional labels or independent sampling evidence.

Texts sharing i form one dependent same-source group (prompt group), so the number of independent units remains n, not $n(1+2m)$. The target is defined by P, never by generated records. A separate outcome learner h_η, trained exclusively on real labeled units, predicts Y from the learned coordinates. After fitting, it scores held-out Q_{eval} texts for the auxiliary term, whereas Q_{style} contributes only to representation learning. Real labels also enter residual correction and variance estimation. Figure 1 summarizes this separation; formal design conditions are in Appendix A.

3 MaSI: Representation Learning and Calibrated Inference

MaSI jointly learns a stable, outcome-relevant representation and scoring rule using an outcome loss on real labeled units and representation constraints supplied by their Q_{style} texts. It then applies the fitted rule to Q_{eval} texts to construct a calibrated auxiliary statistic, with uncertainty aggregated over the original sources. The subscripts name these two uses; they do not denote latent content and style variables.

3.1 Learning the Content Representation

Let $\psi : \mathcal{X} \to \mathbb{R}^d$ be a given encoder and $r < d$ the content rank. MaSI seeks an r-dimensional representation that is stable under Q_{style} variation while retaining information about Y. Following the augmentation-relative setup of von Kügelgen et al. (2021), we use *content* operationally for coordinates that vary little under the specified Q_{style} law while retaining outcome relevance; the orthogonal complement serves as a working *style* representation. Together they form the encoded text covariate, but they are defined relative to the generation design rather than as uniquely identified linguistic factors. For a uniformly selected $c \in \{1, \dots, m\}$ and $\widetilde{X}_c^{\text{style}} \sim Q_{\text{style}}(\cdot \mid X)$, define

$$\Delta U_c = \psi(X) - \psi(\widetilde{X}_c^{\text{style}}), \qquad \Sigma_\Delta = \mathbb{E}[\Delta U_c \Delta U_c^\top].$$

Let $I_1, \dots, I_K$ be K folds and I_k^c the training complement of fold k. Write $U_i = \psi(X_i)$ and $\Delta U_{i,c} = U_i - \psi(\widetilde{X}_{i,c}^{\text{style}})$. On I_k^c, MaSI estimates Σ_Δ by averaging $\Delta U_{i,c} \Delta U_{i,c}^\top$ first within each complete source group and then across real units. The bottom-r eigenspace of Σ_Δ defines the generation-induced center $\Pi^\dagger = \text{EigProj}_{\text{bottom},r}(\Sigma_\Delta)$; its empirical counterpart initializes $V_k^{(0)}$, with $\Pi_k^{(0)} = V_k^{(0)} V_k^{(0)\top}$. Thus "spectral" means stability across same-source Q_{style} texts, not maximal marginal variance; Appendix B gives the empirical covariance explicitly.

Starting from $V_k^{(0)}$, MaSI jointly estimates an orthonormal basis $V \in \mathrm{St}(d, r)$ and an outcome learner with parameter $\eta \in \mathbb{R}^{p_{h,k}}$. Let $s_\eta(z)$ denote the output entering the training loss and $h_\eta(z)$ the scalar score used in inference. For continuous outcomes, $s_\eta = h_\eta$ and $\ell\{s_\eta(z), y\} = \{y - h_\eta(z)\}^2$. For categorical values $y_1, \ldots, y_C$, a reference-class parameterization gives probabilities $p_\eta(z)$, with $s_\eta = p_\eta$, $\ell\{p_\eta(z), y_a\} = -\log p_{\eta,a}(z)$, and $h_\eta(z) = \sum_{a=1}^{C} y_a p_{\eta,a}(z)$. Let $g_{q,k}$ denote the population constraints; Appendix C.1 gives the fold-wise criterion and the exact population and empirical constraints. The population outcome risk is $\mathcal{R}^{\mathrm{tr}}_{y,k}(V, \eta) = \mathbb{E}_P[\ell\{s_\eta(V^\top \psi(X)), Y\}]$. Its fold-wise empirical analogue, using constraint proxies $\widetilde{g}_{q,k}$, is

$$\widehat{\mathcal{R}}^{\mathrm{tr}}_{y,k}(V, \eta) = \frac{1}{|I_k^c|} \sum_{i \in I_k^c} \ell\big[s_\eta\{V^\top U_i\}, Y_i\big], \qquad (\widehat{V}_k, \widehat{\eta}_k) = \arg \min_{V \in \mathrm{St}(d,r),\, \eta \in \mathbb{R}^{p_{h,k}}} \widehat{\mathcal{R}}^{\mathrm{tr}}_{y,k}(V, \eta) \quad \text{s.t.} \quad \widetilde{g}_{q,k}(V) \le 0, \quad q = 0, 1, 2, 3. \qquad (1)$$

The joint representation update uses this unpenalized outcome criterion. After fixing $\widehat{V}_k$, we ridge-refit $\widehat{h}_k$ on the real labeled training fold for finite-sample stability; Appendix D includes the penalty in the downstream nuisance bound. With content coordinates $Z_c(V) = V^\top \psi(\cdot)$ and complementary expression coordinates $Z_s(V) = (I_d - VV^\top)\psi(\cdot)$, where I_d is the $d \times d$ identity matrix, the constraints limit instability and real-versus-generated distinguishability in Z_c, outcome predictability from Z_s, and collapse of Z_s. They are evaluated from complete same-source groups and aggregated across real units; Appendix C.1 gives their formal definitions and Appendix C gives the optimization analysis.

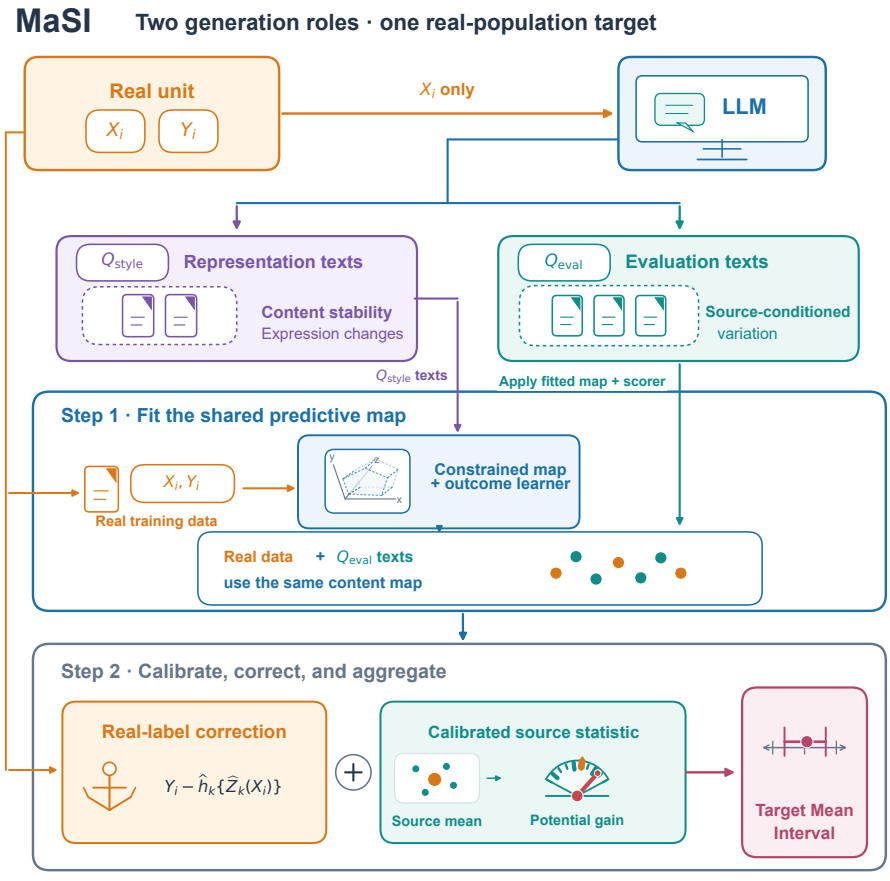

Figure 1: Overview of MaSI. Step 1 uses real labeled units and representation texts to learn an outcome-relevant map stable to expression changes, then scores held-out evaluation texts. Step 2 calibrates and aggregates the auxiliary statistic, adds real-label residual correction, and estimates uncertainty across original units. Only X_i is sent to the LLM.

3.2 Density-ratio Calibration and Same-source Uncertainty

Let $Z_{c,k}(x) = \widehat{V}_k^\top \psi(x)$, with $\widehat{h}_k$ denoting the refitted outcome learner. Conditionally on the training fold, let $P_{k,Z}$ and $Q_{k,Z}$ denote the real and evaluation coordinate laws, and let $w_k^\circ = dP_{k,Z}/dQ_{k,Z}$. Within each outer training fold, MaSI estimates this ratio by prior-corrected logistic classification; source-group-preserving inner folds select shrinkage toward one by balancing content-coordinate and fitted-score moments. Appendix D bounds its prediction-targeted calibration discrepancy and gives sufficient rates for asymptotic negligibility. Although Q_{style} affects $\widehat{V}_k$ through the constraints, the outcome loss and post-representation fit of $\widehat{h}_k$ use only real labeled units; Q_{eval} therefore need not share the Q_{style} distribution. Under overlap, the density ratio calibrates its weighted score mean to that of the same fitted rule under the real-text coordinate law: $Q_{k,Z}(w_k^\circ \widehat{h}_k) = P_{k,Z}(\widehat{h}_k)$. The Q_{style} texts preserve source semantics, whereas Q_{eval} may vary locally around them to provide additional outcome-relevant information. For each held-out real unit, define

$$\widehat{A}_{k,i} = \frac{1}{m}\sum_{c=1}^{m} \widehat{w}_k\{Z_{c,k}(\widetilde{X}_{i,c}^{\mathrm{eval}})\}\widehat{h}_k\{Z_{c,k}(\widetilde{X}_{i,c}^{\mathrm{eval}})\}, \qquad \widehat{B}_{k,i} = Y_i - \widehat{h}_k\{Z_{c,k}(X_i)\}.$$

Writing $\bar{A}_k$ and $\bar{B}_k$ for their held-out fold means and $\pi_k = |I_k|/n$, the estimator is

$$\widehat{\mu} = \sum_{k=1}^{K} \pi_k(\bar{A}_k + \bar{B}_k). \tag{2}$$

The first term is the calibrated within-source auxiliary statistic and the second the real-label residual correction. Because the m scores in $\widehat{A}_{k,i}$ share a source, MaSI averages them first and estimates variance across the n real units. Under the covariance-control regimes in Appendix E, it uses

$$\widehat{\sigma}_{\mathrm{tot}}^2 = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\left\{(\widehat{B}_{k,i} - \bar{B}_k)^2 + (\widehat{A}_{k,i} - \bar{A}_k)^2\right\}. \tag{3}$$

For $0 < \alpha < 1$, MaSI reports $\widehat{C}_{n,\alpha} = [\widehat{\mu} \pm z_{1-\alpha/2}\widehat{\sigma}_{\mathrm{tot}}/\sqrt{n}]$, where z_τ is the standard-normal τ-quantile. Same-source dependence among the m generated scores is retained through $\widehat{A}_{k,i}$; additivity omits only the cross-covariance between the residual and auxiliary contributions. Appendix E proves consistency of Equation (3) under fixed-law exact orthogonality or an n-indexed regime with vanishing average cross-covariance.

4 Theoretical Guarantees

Theorems 1–2 control the learned projection around the generation-induced center $\Pi^\dagger$ and compare it with the semantic oracle $\Pi^\star$; Theorems 3–4 establish a fixed-center expansion and valid source-group confidence intervals. Appendices B–E give the complete conditions and proofs.

4.1 Learning Around a Generation-Stable Center

Let $\Pi^\dagger = \mathrm{EigProj}_{\mathrm{bottom},r}(\Sigma_\Delta)$, $\gamma_\dagger = \lambda_{r+1}(\Sigma_\Delta) - \lambda_r(\Sigma_\Delta) > 0$, and let $t_{n,\delta}$ be the fold-wise concentration radius.

Theorem 1 (Spectral localization of the generation-stable center)**.** *Under Assumption A and the complete-block conditions in Appendix B, if $t_{n,\delta} \le \gamma_\dagger/4$, then with probability at least $1-\delta$,*

$$\max_{1\le k\le K} \|\Pi_k^{(0)} - \Pi^\dagger\|_F \le \frac{2\sqrt{2r}\, t_{n,\delta}}{\gamma_\dagger}. \tag{4}$$

Theorem 2 (Constrained projection control and semantic-oracle comparison)**.** *Let $\widehat{\Pi}_k = \widehat{V}_k\widehat{V}_k^\top$, and let $\bar{c}_{0,n}$ and $\delta_{\mathrm{feas},n}$ denote the stability and feasibility envelopes in Appendix C. Under Assumption B and the learning conditions therein, the first bound holds. If the semantic-perturbation conditions of Lemma B.1 also hold, with separation ν_S and discrepancy τ_e, so does the second:*

$$\begin{aligned}
\max_{1\le k\le K} \|\widehat{\Pi}_k - \Pi^\dagger\|_F &\le \left\{\frac{2(\bar{c}_{0,n} + \delta_{\mathrm{feas},n})}{\gamma_\dagger}\right\}^{1/2}, \\
\max_{1\le k\le K} \|\widehat{\Pi}_k - \Pi^\star\|_F &\le \left\{\frac{2(\bar{c}_{0,n} + \delta_{\mathrm{feas},n})}{\gamma_\dagger}\right\}^{1/2} + \frac{2\sqrt{2r}\tau_e}{\nu_S}.
\end{aligned} \tag{5}$$

The bounds separate local learning error from generation-to-oracle discrepancy; if both vanish, $\widehat{\Pi}_k$ recovers $\Pi^\star$.

4.2 Fixed-center Calibrated Residual Expansion

Choose a basis $V^\dagger$ of $\Pi^\dagger$. Let $P_{Z^\dagger}$ and $Q_{Z^\dagger}$ be the laws of $Z^\dagger = V^{\dagger\top}\psi(X)$ and $V^{\dagger\top}\psi(\widetilde{X}^{\mathrm{eval}})$, respectively. Define $h^\dagger(z) = \mathbb{E}_P(Y \mid Z^\dagger = z)$, $w^\dagger = dP_{Z^\dagger}/dQ_{Z^\dagger}$, $\varphi^\dagger(x) = w^\dagger\{V^{\dagger\top}\psi(x)\}h^\dagger\{V^{\dagger\top}\psi(x)\}$, and $b^\dagger(x,y) = y - h^\dagger\{V^{\dagger\top}\psi(x)\}$. Write $Qg = \mathbb{E}\{g(\widetilde{X}^{\mathrm{eval}})\}$, $B^\dagger = b^\dagger(X,Y)$, $B_i^\dagger = b^\dagger(X_i,Y_i) - \mathbb{E}_P(B^\dagger)$, and

$$A_Q^\dagger = \frac{1}{m}\sum_{c=1}^{m}\{\varphi^\dagger(\widetilde{X}_c^{\mathrm{eval}}) - Q\varphi^\dagger\}$$

let $A_{Q,i}^\dagger$ be its unit-specific copy.

Theorem 3 (Fixed-center expansion under targeted calibration rates)**.** *Under Assumptions A–C, Theorem 2, and the prediction-targeted density-ratio rate and remaining nuisance conditions in Appendix D,*

$$\widehat{\mu} - \mu = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\{B_i^\dagger + A_{Q,i}^\dagger\} + R_n, \qquad \sqrt{n}R_n = o_p(1).$$

Here $\Pi^\dagger$ is the deterministic first-order reference; replacing it by the learned $\widehat{\Pi}_k$, together with other nuisance effects, enters R_n. Density-ratio approximation and estimation errors are also included explicitly in R_n.

4.3 Variance and Confidence-interval Validity

Let $\sigma_P^2 = \mathrm{Var}(B^\dagger)$ and $\sigma_Q^2 = \mathrm{Var}(A_Q^\dagger)$, where σ_Q^2 retains same-source dependence. Exact covariance control holds, for example, when the learned coordinates suffice for the conditional outcome mean or when the conditional mean of the auxiliary group contribution depends on the source only through those coordinates. Appendix E also establishes the indexed regime in which the average cross-covariance vanishes.

Theorem 4 (Additive variance consistency and Wald validity under covariance control)**.** *Under the Appendix E fixed-law exact covariance-control regime, Theorem 3, the stated moment conditions, and $0 < \sigma_P^2 + \sigma_Q^2 < \infty$,*

$$\sqrt{n}(\widehat{\mu} - \mu) \xrightarrow{d} N(0, \sigma_P^2 + \sigma_Q^2), \qquad \widehat{\sigma}_{\mathrm{tot}}^2 \xrightarrow{p} \sigma_P^2 + \sigma_Q^2.$$

Consequently, $\widehat{C}_{n,\alpha}$ has asymptotic coverage $1-\alpha$; Appendix E gives the indexed extension.

Appendix Corollary E.2 gives shorter intervals exactly when $\sigma_P^2 + \sigma_Q^2 < \mathrm{Var}(Y)$. Together, the four results connect representation learning to valid, potentially shorter intervals.

5 Experiments

We report empirical 90% coverage and width relative to Classical. Main comparators are Classical and PPI on raw or fixed PCA coordinates; Appendix F defines other baselines, ablations, and controls. Valid procedures estimate uncertainty across real prompt groups; matched comparisons fix generation count, and same-pool analyses also fix text identity.

5.1 Simulation Experiments: Validity, Baselines, and Mechanism Boundaries

The main simulation uses $n = 200$, five folds, $m = 10$ texts per channel, and 300 Monte Carlo replications across four outcome structures and two evaluation-content retention levels; smaller ρ denotes stronger outcome-relevant perturbation. Methods share real samples, generated covariates, folds, and fold-local tuning. Appendix Table F.1 gives complete definitions; the Monte Carlo standard error near 0.90 coverage is approximately 0.017.

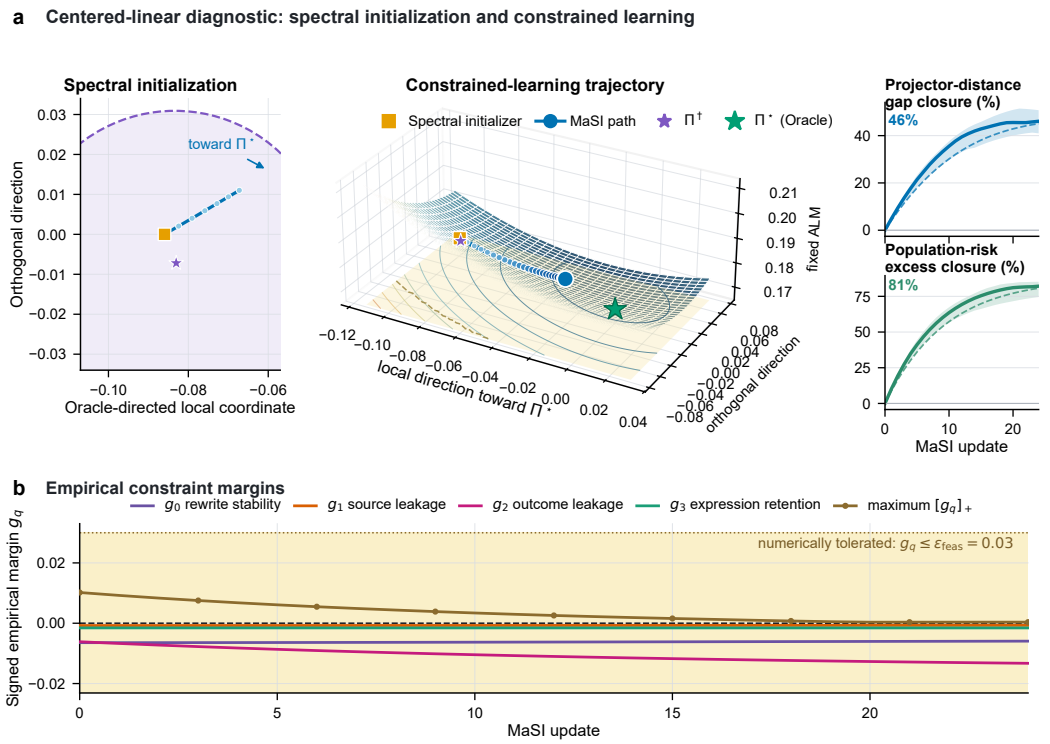

Figure 2: Centered-linear mechanism audit. Panel (a) tracks an outcome-guided, constraint-respecting trajectory from the spectral initializer and retrospectively reports distance to the unavailable simulation oracle, which the optimizer does not use. Solid curves and bands summarize 30-run medians and interquartile ranges; the dashed curve is selected by a prespecified robust-scaled rule. $\Pi^{\dagger}$ and $\Pi^{\star}$ mark the generation-induced center and oracle. Panel (b) reports signed constraint margins; pale yellow denotes $g_q \leq \varepsilon_{\text{audit}} = 0.03$.

Coverage, precision, and problem complexity. Across the eight low-dimensional settings, MaSI attains coverage/width ratio $0.912/0.735$, versus $0.916/0.784$ for raw PPI and $0.913/0.761$ for PCA-5 PPI. Its width ratio matches raw PPI after rounding in both linear settings; relative to raw PPI, the gaps are larger for nonlinear, binary, and ordinal outcomes, especially at $\rho = 0.5$. Against PCA-5 PPI, MaSI is 1.8–8.0% shorter under strong perturbation and at most 2.9% shorter under weak perturbation, at similar coverage. The boundary becomes sharper in a style-dominant nonlinear setting: coverage is comparable for MaSI, raw PPI, and PCA-5 PPI (0.880, 0.870, and 0.860), while MaSI shortens intervals by 21.9% and 22.9% relative to the two PPI methods. Thus the gain reflects the interaction of outcome structure, generation alignment, and expression interference rather than nonlinearity alone (Appendix Tables F.13 and F.14). The independent-text control narrows the average width ratio to 0.585 but covers only 0.782, isolating the cost of treating same-source generations as independent.

Representation and optimization. Across 1,500 fold fits, the spectral initializer remains close to the oracle subspace (mean projector distance 0.040). In a separate 30-run audit, the subsequent constrained fit reduces mean projector distance and excess risk by 46% and 81%, while every path remains within the numerical tolerance (Figure 2; Appendix Table F.4). Together, these results support a sequence of stable initialization, outcome-guided learning, and simultaneous constraint control.

Channel separation and inferential scope. Channel separation is nearly neutral when generation is well aligned, but under stronger perturbation it shortens intervals across all four outcomes at similar coverage. In the style-dominant stress test above, all three representation-based methods fall below nominal coverage, so the comparison describes relative representation performance

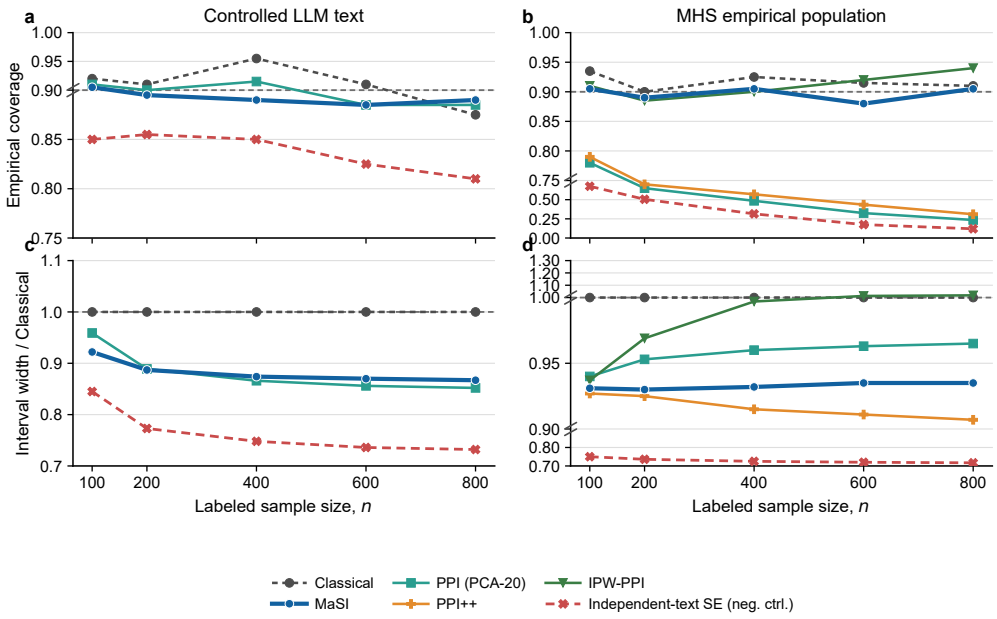

Figure 3: Coverage and width for controlled LLM rendering (left) and the MHS empirical population (right). Top: coverage, with 0.90 marked by the dashed line (Monte Carlo standard error ≈ 0.021 for $R = 200$); bottom: width relative to Classical. The red dashed negative control uses the PCA-20 PPI point estimator but treats same-source texts as independent; valid procedures aggregate by real prompt group. Panels (a), (b), and (d) have marked continuous piecewise-linear ordinates, so distances and slopes are comparable only within segments; panel (c) is linear. Complete comparisons appear in Appendix F.5.

rather than interval validity (Appendix Tables F.15 and F.13). Generation-count analyses show no shortening at $m = 1$ and diminishing gains thereafter, while Appendix Table F.7 reports covariance and ordinal-feasibility diagnostics across n. Appendix Figures F.1 and F.2 examine localization, the fixed-center remainder, coverage, and the relation between outcome alignment and interval shortening; in particular, weak alignment need not shorten intervals. These finite-sample patterns support the predicted roles of localization, outcome alignment, and source-level aggregation.

5.2 Text Experiments: Controlled LLM Rendering and Real-Data Analysis

The controlled study has a known target and LLM-rendered product reviews; MHS (Sachdeva et al., 2022) repeatedly samples from a fixed 2,000-comment empirical population. Both withhold labels from the LLM and aggregate uncertainty by real prompt group. Figure 3 summarizes the primary coverage–width comparisons; complete protocols and results appear in Appendix F.

Controlled rendering. This is a benign regime: MaSI and PCA-20 PPI have essentially the same coverage–width tradeoff, while PCA-5 PPI is slightly shorter at similar coverage. The independent-text control is narrower but undercovers, showing that a fixed representation is empirically competitive here while prompt-group uncertainty remains necessary (Appendix Table F.16).

MHS empirical population: primary comparisons. MaSI maintains coverage of 0.880–0.905 with width ratios of 0.930–0.935. Uncalibrated PCA-20 PPI is shorter but retains persistent bias and sharply deteriorating coverage. Count-matched importance-weighted PPI (IPW-PPI) restores average coverage to 0.907 but yields little shortening (width ratio 0.985); the generated-text PPI++ adaptation, which tunes correction strength without calibrating the Q_{eval}-to-real shift, remains undercovered, and filtered PPI (FPPI) shortens less than MaSI. Complete results appear in Appendix Tables F.19 and F.20. Count matching equalizes generation count, not text identity.

Representation learning and channel roles. Spectral initialization alone retains coverage but yields nearly classical width. The unconstrained outcome-guided fit attains average coverage/width $0.868/0.929$, whereas MaSI attains $0.897/0.933$; the constraints therefore improve coverage with almost no loss of interval shortening. Because all constraints are removed jointly, the ablation supports their collective rather than individual role. With the text pool fixed, the same-pool shuffled-role assignment attains $0.902/0.960$, compared with $0.897/0.933$ for MaSI, providing descriptive evidence that the prespecified channel roles can improve precision at similar coverage (Appendix Tables F.18 and F.20).

Theory-facing diagnostics. Source-classification AUC falls from 0.643 in the raw embedding to 0.495–0.498 in the learned content coordinates, indicating lower real-versus-generated separability under the audited classifier. Permuting Q_{eval} prompt groups reduces held-out predictive fit for every sample-size–learner combination, a pattern consistent with source-linked predictive information rather than generic synthetic variation. As n increases from 100 to 800, the relative residual–synthetic covariance diagnostic decreases from 0.113 to 0.055, indicating weaker residual coupling. Together, these results support the observable empirical counterparts of content stability, source-specific predictive information, and covariance control (Appendix Tables F.11–F.24).

Taken together, MHS separates three empirical regimes: uncalibrated PPI can be short but biased, count-matched IPW-PPI can recover coverage while sacrificing most shortening, and MaSI retains a more favorable coverage–precision balance when generated texts are informative but differ from real text and remain dependent within source.

6 Conclusion and Discussion

MaSI provides a principled route for using source-conditioned LLM generations in real-population mean inference. It separates content-stable representation learning from calibrated auxiliary estimation, while real-label correction anchors the target and prompt-group variance preserves the sampling structure. Under the stated conditions, its variance estimator is consistent and its confidence interval attains asymptotic $1 - \alpha$ coverage. Across the primary simulation, controlled-text, and MHS settings, outcome-relevant, calibratable generation shortens intervals while empirical coverage remains near nominal.

The central distinction is between model-derived information and sampling evidence. In the fixed-law exact covariance-control regime, MaSI is asymptotically more precise exactly when the calibrated signal removes more residual variance than the source-group statistic contributes. More generations can stabilize within-source averaging but do not create independent evidence. Accordingly, MaSI shifts the design question from how much text to generate to which generated variations are informative, calibratable, and compatible with valid uncertainty quantification.

This design addresses the tension between generative fidelity and incremental information. Near-duplicates preserve fidelity but add little beyond a fixed representation; stronger variations may reveal outcome-relevant structure while increasing content or calibration error. The two channels separate these demands: representation texts help learn expression-stable coordinates, evaluation generations contribute local outcome-related information, and real-label correction with source-level uncertainty anchors the estimator. Compact PPI can be competitive when representation and distribution alignment are benign, whereas the full mechanism is more useful in the studied settings where representation shift and same-source dependence coexist. The diagnostics indicate which regime is present.

This perspective motivates generative information design: generators, prompts, and within-source sampling rules should be evaluated as inferential components, not only by linguistic plausibility. Present guarantees assume a fixed generation law, fixed K and m, overlap, nuisance rates, and covariance control; they neither identify internal semantics nor cover adaptive prompting. Interpretation of opaque generators should therefore rest on observable stability, calibration, and source sensitivity. Future work can quantify generator uncertainty, derive weak-overlap sensitivity bounds and covariance-robust intervals, and extend sample-split adaptive generation beyond means. These advances could support public health, social measurement, surveys, and content analysis with scarce text–outcome pairs. MaSI thus offers a testable statistical path from generator-induced semantic variation to auditable precision gains.

AI USE STATEMENT

In this work, `deepseek-ai/DeepSeek-V3` generated the synthetic text covariates used in the controlled-rendering and MHS experiments under prespecified prompts and sampling settings; `Qwen/Qwen3-Embedding-8B` served as the fixed text encoder. OpenAI Codex (GPT-5.6) assisted with manuscript writing and language polishing, including prose refinement and organization, as well as literature search and discovery. The authors checked all AI-assisted text, cited sources, and synthetic-data generation settings. The research questions, methods, theoretical claims, experimental conclusions, and final wording were determined and approved by the authors, who take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Appendices B–E state the theoretical conditions and provide complete proofs, while Appendix F documents the data-generating mechanisms, evaluation populations, fold structure, comparison methods, model and generation settings, diagnostics, and complete reported results. The anonymous repository at `https://anonymous.4open.science/r/MaSI2027` provides the implementation, fixed paper-scale configurations, automated tests, reproduction commands, expected summaries, and runtime guidance for the primary synthetic simulations and mechanism comparisons, together with the audited channel prompt templates used in the controlled-rendering and MHS experiments. This self-contained artifact reproduces the simulation evidence without private data or access to an external model service. Upon acceptance, we will release the remaining releasable analysis code, subject to dataset licenses and model-service terms.

REFERENCES

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

Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. PPI++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023b. doi:10.48550/arXiv.2311.01453.

Meshi Bashari, Roy Maor Lotan, Yonghoon Lee, Edgar Dobriban, and Yaniv Romano. Synthetic-powered predictive inference. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi:10.52202/085713-3925. URL `https://papers.nips.cc/paper_files/paper/2025/hash/aac47a65339d7353562f6365526ee81b-Abstract-Conference.html`.

Meshi Bashari, Yonghoon Lee, Roy Maor Lotan, Edgar Dobriban, and Yaniv Romano. General synthetic-powered inference. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL `https://openreview.net/forum?id=sxLncu2Fhx`.

F. Jay Breidt and Jean D. Opsomer. Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32(2):190–205, 2017. doi:10.1214/16-STS589.

Yewon Byun, Shantanu Gupta, Zachary C. Lipton, Rachel Leah Childers, and Bryan Wilder. Valid inference with imperfect synthetic data. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi:10.52202/085713-5424. URL `https://papers.nips.cc/paper_files/paper/2025/hash/ede6376c80b38796c688a77dce9fb227-Abstract-Conference.html`.

Shuxiao Chen, Edgar Dobriban, and Jane H. Lee. A group-theoretic framework for data augmentation. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/f4573fc71c731d5c362f0d7860945b88-Abstract.html`.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems 23*, pp. 442–450, 2010.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, et al. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. URL https://arxiv.org/abs/2412.19437.

Siyi Deng, Yang Ning, Jiwei Zhao, and Heping Zhang. Optimal and safe estimation for high-dimensional semi-supervised learning. *Journal of the American Statistical Association*, 119(548): 2748–2759, 2024. doi:10.1080/01621459.2023.2277409.

Naoki Egami, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi:10.52202/075280-3000.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 968–988, Online, 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.findings-acl.84. URL https://aclanthology.org/2021.findings-acl.84/.

Adam Fisch, Joshua Maynez, R. Alex Hofer, Bhuwan Dhingra, Amir Globerson, and William W. Cohen. Stratified prediction-powered inference for effective hybrid evaluation of language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi:10.52202/079017-3541.

Kristina Gligorić, Tijana Zrnic, Cinoo Lee, Emmanuel Candès, and Dan Jurafsky. Can unconfident LLM annotations be used for confident conclusions? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3514–3533. Association for Computational Linguistics, 2025. doi:10.18653/v1/2025.naacl-long.179. URL https://aclanthology.org/2025.naacl-long.179/.

Andrew Halterman. Synthetically generated text for supervised text analysis. *Political Analysis*, 33 (3):181–194, 2025. doi:10.1017/pan.2024.31.

Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. Generate, annotate, and learn: NLP with synthetic text. *Transactions of the Association for Computational Linguistics*, 10:826–842, 2022. doi:10.1162/tacl_a_00492. URL https://aclanthology.org/2022.tacl-1.48/.

Jiayuan Huang, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Bernhard Schölkopf. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems 19*, pp. 601–608, 2007.

Wouter M. Kouw and Marco Loog. A review of domain adaptation without target labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):766–785, 2021. doi:10.1109/TPAMI.2019.2945942.

Cheng Lu, Mengxin Wang, Dennis J. Zhang, and Heng Zhang. Generative augmented inference. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=JLvbqyMPl6.

Marc Ratkovic. Large language models for statistical inference: Context augmentation with applications to the two-sample problem and regression. *arXiv preprint arXiv:2506.23862*, 2025. doi:10.48550/arXiv.2506.23862.

Quentin Rebjock and Nicolas Boumal. Fast convergence to non-isolated minima: four equivalent conditions for C^2 functions. *Mathematical Programming*, 213:151–199, 2025. doi:10.1007/s10107-024-02136-6.

Pratik S. Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pp. 83–94, Marseille, France, 2022. European Language Resources Association. URL https://aclanthology.org/2022.nlperspectives-1.11/.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. doi:10.1016/S0378-3758(00)00115-4.

Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. doi:10.1007/s10208-011-9099-z.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16451–16467, 2021.

Chenwei Wu, Holden Lee, and Rong Ge. Connecting pre-trained language model and downstream task via properties of representation. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi:10.52202/075280-2044. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/93712c59f6a81bd92040facf04c8b308-Abstract-Conference.html.

Shirong Xu and Will Wei Sun. Do More Predictions Improve Statistical Inference? Filtered Prediction-Powered Inference. *arXiv preprint arXiv:2602.10464*, 2026. doi:10.48550/arXiv.2602.10464.

Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015. doi:10.1093/biomet/asv008.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025. URL https://arxiv.org/abs/2506.05176.

Yuqian Zhang and Jelena Bradic. High-dimensional semi-supervised learning: in search of optimal inference of the mean. *Biometrika*, 109(2):387–403, 2022. doi:10.1093/biomet/asab042.

Tijana Zrnic and Emmanuel J. Candès. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 2024. doi:10.1073/pnas.2322083121.

A NOTATION, EVALUATION DESIGN, AND PROOF ROADMAP

This appendix fixes notation and records the design conventions used in the proofs. The symbol $\Pi^\star$ denotes an ideal semantic content projector used only for interpretation. The inferential theory is centered at $\Pi^\dagger$, the pseudo-true projector induced by the population paired-difference second-moment matrix of the actual Q_{style} rewrites. For any orthonormal basis $V^\dagger$ of this eigenspace, $Z^\dagger = V^{\dagger\top}\psi(X)$ denotes the fixed-center content coordinate, with the same map applied to generated texts. Fold-specific learned projections are $\widehat{\Pi}_k = \widehat{V}_k\widehat{V}_k^\top$.

Operational meaning of content and style. Following the augmentation-relative perspective of von Kügelgen et al. (2021), we define content and style relative to the specified generation mechanism, not as universal or uniquely identifiable linguistic factors. In MaSI, content denotes coordinates intended to change little under the actual same-source Q_{style} law while retaining outcome relevance; the complement provides working expression coordinates. The subscripts in Q_{style} and Q_{eval} label procedural roles rather than two latent variables: the former supplies texts for representation learning, whereas the latter supplies separate post-fit texts for calibrated auxiliary estimation and may vary both expression and outcome-relevant information. Only the optional semantic model in Lemma B.1 introduces the ideal $\Pi^\star$; all inferential guarantees are centered at $\Pi^\dagger$.

Projection and fixed-center objects.

$\Pi^\star$	Ideal semantic content projector, used only as an interpretive benchmark.
$\Pi^\dagger$	Pseudo-true projector induced by the Q_{style} paired-difference second moment.
$\widehat{\Pi}_k$	Content projection learned on fold k.
$h^\dagger, w^\dagger, \varphi^\dagger, b^\dagger$	Deterministic fixed-center prediction, density-ratio, weighted-prediction, and residual objects at $\Pi^\dagger$.
$h_k^\circ, w_k^\circ, \varphi_k^\circ$	Fold-specific population oracle objects in the learned coordinates.
$\widehat{h}_k, \widehat{w}_k, \widehat{\varphi}_k, \widehat{b}_k$	Fold-specific fitted objects used by the estimator.

Representation and calibration quantities.

$\ell_0^\dagger, \bar{c}_{0,n}$	Minimum population paired-rewrite stability loss and its recentered tolerance $c_{0,n} - \ell_0^\dagger$.
$\underline{\kappa}_k, \underline{\kappa}_k^{\mathrm{nor}}$	Slice-level block curvature and its induced normal-space curvature in the local PL argument.
$\beta_{w,n}, a_{\beta,n}$	Prediction-targeted calibration remainder and density-ratio approximation bias in the prediction direction.
$r_{\gamma,n}, \chi_{w,n}$	Fold-averaged density-ratio parameter error and its tangent coupling with the fitted prediction direction.
$\delta_{h,n}^{\mathrm{ref}}$	Fold-averaged excess-risk envelope for the post-representation outcome refit, including empirical fluctuation, ridge bias, and numerical optimization error.
$a_{\psi,n}, a_{h,\dagger,n}$	Encoder conditional-mean approximation error and, respectively, prediction-class approximation and pseudo-true discrepancy at $\Pi^\dagger$.

Projection, covariance, and precision quantities.

$\delta_{\Pi,n}^\dagger$	Pseudo-true projection error $\max_k \|\widehat{\Pi}_k - \Pi^\dagger\|_F$.
$\delta_{Q,n}^\dagger$	Residual dependence of the synthetic score beyond $Z^\dagger$.
$\mathfrak{r}_{\perp,n}$	Derived bound on residual–synthetic covariance.
$\mathcal{G}_{\mathrm{MA}}$	Model-assisted variance reduction $V_H - V_Q$ in the additive regime.

Roles of the representation constraints.

- g_0 Keeps real texts and paired Q_{style} rewrites close in content coordinates; together with the pseudo-true eigengap, it controls projection error.
- g_1 Limits real-versus-rewrite source information from Q_{style} in the content representation. It does not replace density-ratio calibration for Q_{eval}.
- g_2 Limits restricted outcome-predictive information in the residual expression representation. Through the predictive-leakage bridge, its feasibility contributes to the derived residual–synthetic covariance bound.
- g_3 Retains expression or prompt variation in the residual representation and prevents representation collapse.

These constraints provide complementary control of representation geometry; no individual constraint alone implies valid coverage.

Framework	Connection to MaSI
Model-assisted / PPI / StratPPI (Breidt & Opsomer, 2017; Angelopoulos et al., 2023a;b; Fisch et al., 2024)	Use auxiliary variables, predictions, or autorater labels to improve estimation, with real observations anchoring correction. They do not learn a representation from paired rewrites or model same-source generated texts as source-defined groups.
LLM-assisted surrogate annotation (Egami et al., 2023; Gligorić et al., 2025)	Treats LLM annotations as imperfect proxies for human-coded outcomes and combines them with sampled gold labels for valid downstream inference. MaSI instead uses source-conditioned generations to construct auxiliary covariates without treating them as labeled units; Y is observed only on real units.
Imperfect synthetic-data GMM (Byun et al., 2025)	Combines real and wholly synthetic observations through residual moments. MaSI instead transports generated covariates within source-defined groups.
SPPI / GESPI (Bashari et al., 2025; 2026)	Use synthetic calibration scores or synthetic-data procedures for predictive sets and general base methods. MaSI targets a real-population mean with two role-specific generation channels.
Context augmentation (Ratkovic, 2025)	Generates contexts around observed strings for testing and text-on-text regression. It shares conditional generation but not MaSI's separation of representation and evaluation channels.
Generative augmented inference (Lu et al., 2026)	Uses AI outputs as features in orthogonal moments for human-labeled outcomes. MaSI additionally learns a content map from rewrites, transports evaluation coordinates, and estimates uncertainty by source group.
MaSI	Uses label-free style rewrites for representation learning and evaluation texts for calibrated auxiliary estimation, with real-label residual correction and source-group uncertainty.

Table A.1: Connections to adjacent inference frameworks. Each row identifies the auxiliary object, inferential target, and distinguishing role relative to MaSI.

Design conditions. Real units are i.i.d.; folds are fixed in advance or generated independently; and K and m are fixed. The two generation kernels receive X_i, not Y_i, so their randomization is conditionally independent of Y_i given X_i. Auxiliary generation context is included in X_i. Dependence within a prompt group is unrestricted, whereas complete groups are independent across units. Fold-k nuisance fits and tuning choices are measurable with respect to the training sigma-field $\mathcal{T}_k$; held-out groups enter only the final fold averages. Assumption C supplies the remaining range and moment conditions.

Proof map. Appendix B proves spectral localization; Appendix C develops the conditional surrogate-to-profile and local-ALM projection bounds; Appendix D establishes transport and fixed-

center replacement; and Appendix E proves covariance control, additive variance consistency, and Wald coverage. Appendix F contains the experimental protocols and complete empirical results. The supporting lemmas and full proofs appear in their order of use.

B Paired-rewrite Geometry and Pseudo-true Localization

Let $\mathcal{G}_r := \{\Pi \in \mathbb{R}^{d\times d} : \Pi^\top = \Pi,\ \Pi^2 = \Pi,\ \mathrm{rank}(\Pi) = r\}$ denote the rank-r Grassmann manifold represented by orthogonal projectors.

B.1 Assumption A: Pseudo-true Content Separation

Assumption A (pseudo-true content separation). The population paired-difference second-moment (scatter) matrix $\Sigma_\Delta = \mathbb{E}\left[\frac{1}{m}\sum_{c=1}^m \Delta U_c \Delta U_c^\top\right]$ has a strictly separated bottom-r eigenspace: $\gamma_\dagger := \lambda_{r+1}(\Sigma_\Delta) - \lambda_r(\Sigma_\Delta) \geq \underline{\gamma} > 0$. Its bottom-$r$ spectral projector is denoted by $\Pi^\dagger$. This is the single population-identification condition for the spectral center; inference additionally uses the overlap, nuisance, cluster, and covariance conditions stated in Appendices D–E.

B.2 Generation-induced Center and Semantic Comparison

Lemma B.1 (Generation-induced center and optional semantic comparison)**.** *Under Assumption A,* $\Pi^\dagger = \arg\min_{\Pi\in\mathcal{G}_r} \mathrm{tr}(\Pi\Sigma_\Delta)$ *is unique, and* $\ell_0^\dagger := \min_{\Pi\in\mathcal{G}_r} \mathrm{tr}(\Pi\Sigma_\Delta) = \sum_{j=1}^r \lambda_j(\Sigma_\Delta)$. *If, additionally, there exists an ideal semantic projector* $\Pi^\star = V_\star V_\star^\top$ *and a decomposition* $\Sigma_\Delta = \Sigma_\Delta^0 + R_e$, $\Sigma_\Delta^0 := B_\star \Sigma_{\Delta S} B_\star^\top$, $\lambda_{\min}(\Sigma_{\Delta S}) \geq \nu_S > 0$, $\|R_e\|_{\mathrm{op}} \leq \tau_e \leq \nu_S/4$, *where* $V_\star^\top V_\star = I_r$, $B_\star^\top B_\star = I_{d-r}$, $V_\star^\top B_\star = 0$, *and* $B_\star B_\star^\top = I_d - \Pi^\star$, *then*

$$\|\Pi^\dagger - \Pi^\star\|_F \leq \frac{2\sqrt{2r}\,\tau_e}{\nu_S}. \tag{B.1}$$

Proof. Let $v_1, \ldots, v_d$ be an orthonormal eigenbasis of Σ_Δ with ascending eigenvalues $\lambda_1 \leq \cdots \leq \lambda_d$. For any rank-$r$ orthogonal projector Π, define $p_j = v_j^\top \Pi v_j$, where $0 \leq p_j \leq 1$ and $\sum_{j=1}^d p_j = \mathrm{tr}(\Pi) = r$. Then $\mathrm{tr}(\Pi\Sigma_\Delta) = \sum_{j=1}^d \lambda_j p_j$. By the Rayleigh–Ritz principle, this quantity is minimized by assigning unit mass to the r smallest eigenvalues, giving $\min_{\Pi\in\mathcal{G}_r} \mathrm{tr}(\Pi\Sigma_\Delta) = \sum_{j=1}^r \lambda_j$. Because $\lambda_r < \lambda_{r+1}$, equality requires $p_j = 1$ for $j \leq r$ and $p_j = 0$ for $j > r$. Hence the minimizing range is uniquely $\mathrm{span}(v_1, \ldots, v_r)$, and the unique minimizing projector is $\Pi^\dagger$.

For the optional semantic comparison, under the additional conditions in the lemma statement, Σ_Δ^0 has r zero eigenvalues on $\mathrm{range}(\Pi^\star)$ and all remaining eigenvalues are at least ν_S. Since $\Sigma_\Delta - \Sigma_\Delta^0 = R_e$ and $\|R_e\|_{\mathrm{op}} \leq \tau_e$, the Frobenius-form Davis–Kahan inequality (Yu et al., 2015) gives

$$\|\Pi^\dagger - \Pi^\star\|_F \leq \frac{2\sqrt{2r}\|R_e\|_{\mathrm{op}}}{\nu_S} \leq \frac{2\sqrt{2r}\tau_e}{\nu_S}.$$

This proves (B.1). □

B.3 Complete-block Covariance Concentration

Lemma B.2 (Complete-block covariance concentration)**.** *For fold* k, *let*

$$H_i := \frac{1}{m}\sum_{c=1}^m \Delta U_{i,c} \Delta U_{i,c}^\top, \qquad \widehat{\Sigma}_{\Delta,k} = \frac{1}{|I_k^c|}\sum_{i\in I_k^c} H_i, \qquad E_k = \widehat{\Sigma}_{\Delta,k} - \Sigma_\Delta,$$

and define

$$v_k := \left\| \frac{1}{|I_k^c|}\sum_{i\in I_k^c} \mathbb{E}\{(H_i - \Sigma_\Delta)^2\} \right\|_{\mathrm{op}}.$$

If $\|\Delta U_{i,c}\|_2 \le M_\Delta$ almost surely and the complete blocks are independent across i, then

$$\Pr\{\|E_k\|_{\mathrm{op}} \ge t\} \le 2d \exp\left[-\frac{|I_k^c| t^2}{2v_k + (4/3)M_\Delta^2 t}\right].$$

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

$$\max_{1\le k\le K} \|\widehat{\Sigma}_{\Delta,k} - \Sigma_\Delta\|_{\mathrm{op}} \le t_{n,\delta},$$

where

$$t_{n,\delta} = \max_k \left[\sqrt{\frac{2v_k \log(2dK/\delta)}{|I_k^c|}} + \frac{4M_\Delta^2 \log(2dK/\delta)}{3|I_k^c|}\right].$$

Proof. Because H_i is an average of positive-semidefinite rank-one matrices,

$$\|H_i\|_{\mathrm{op}} \le \frac{1}{m}\sum_{c=1}^m \|\Delta U_{i,c}\Delta U_{i,c}^\top\|_{\mathrm{op}} = \frac{1}{m}\sum_{c=1}^m \|\Delta U_{i,c}\|_2^2 \le M_\Delta^2.$$

Define centered matrices $Y_{k,i} := (H_i - \Sigma_\Delta)/|I_k^c|$ for $i \in I_k^c$. Then $\mathbb{E}(Y_{k,i}) = 0$,

$$\|Y_{k,i}\|_{\mathrm{op}} \le \frac{\|H_i\|_{\mathrm{op}} + \|\Sigma_\Delta\|_{\mathrm{op}}}{|I_k^c|} \le \frac{2M_\Delta^2}{|I_k^c|},$$

and

$$\left\|\sum_{i\in I_k^c} \mathbb{E}(Y_{k,i}^2)\right\|_{\mathrm{op}} = \frac{v_k}{|I_k^c|}.$$

The two-sided self-adjoint matrix Bernstein inequality (Tropp, 2012) therefore yields

$$\Pr\left(\left\|\sum_{i\in I_k^c} Y_{k,i}\right\|_{\mathrm{op}} \ge t\right) \le 2d \exp\left[-\frac{t^2}{2v_k/|I_k^c| + (4/3)M_\Delta^2 t/|I_k^c|}\right],$$

which is the stated fold-wise bound. Setting the right-hand side to δ/K and applying a union bound over k gives $t_{n,\delta}$. The argument treats the entire prompt group H_i as one observation and imposes no independence among rewrites within the same group. □

B.4 Proof of Theorem 1

Proof of Theorem 1. Fix a fold k. Let $V^\dagger$ be an orthonormal basis for the bottom-r eigenspace of Σ_Δ, and let $V_k^{(0)}$ be the bottom-r eigenvectors of $\widehat{\Sigma}_{\Delta,k}$. By Assumption A, the population eigengap separating the target eigenspace from its complement is $\gamma_\dagger > 0$. The Davis–Kahan sin-Θ inequality (Yu et al., 2015) gives

$$\|\sin\Theta(V_k^{(0)}, V^\dagger)\|_F \le \frac{2\sqrt{r}\,\|\widehat{\Sigma}_{\Delta,k} - \Sigma_\Delta\|_{\mathrm{op}}}{\gamma_\dagger}.$$

For two rank-r orthogonal projectors, $\|\Pi_k^{(0)} - \Pi^\dagger\|_F = \sqrt{2}\,\|\sin\Theta(V_k^{(0)}, V^\dagger)\|_F$. Combining the two displays,

$$\|\Pi_k^{(0)} - \Pi^\dagger\|_F \le \frac{2\sqrt{2r}\,\|\widehat{\Sigma}_{\Delta,k} - \Sigma_\Delta\|_{\mathrm{op}}}{\gamma_\dagger}.$$

On the simultaneous event in Lemma B.2, the operator norm in the numerator is at most $t_{n,\delta}$ for every k. Taking the maximum over folds proves the theorem. The condition $t_{n,\delta} \le \gamma_\dagger/4$ guarantees stable ordering of the empirical bottom eigenspace and may be weakened under an equivalent spectral-separation formulation. □

C Riemannian Augmented-Lagrangian Interface and Local Projection Control

C.1 Fold-wise Objects and Actual Empirical Constraints

Fix an outer fold k and write $n_k^{\mathrm{tr}} = |I_k^c|$. The local training space is

$$\Theta_k^\dagger(r_0) = \left\{(V,\eta) : V \in \mathrm{St}(d,r),\ \|VV^\top - \Pi^\dagger\|_F \le r_0,\ \eta \in \mathbb{R}^{p_{h,k}}\right\}.$$

The fitted outcome criterion is

$$\widehat{\mathcal{R}}_{y,k}^{\mathrm{tr}}(V,\eta) = \frac{1}{n_k^{\mathrm{tr}}} \sum_{i \in I_k^c} \ell\big[s_\eta\{V^\top U_i\}, Y_i\big],$$

and $\mathcal{R}_{y,k}^{\mathrm{tr}}(V,\eta)$ denotes its population counterpart. This is the unpenalized criterion used in the joint representation update. After $\widehat{V}_k$ is obtained, the outcome learner is refitted on the same training fold with ridge regularization; this separate refit enters the inference theory through the downstream prediction-error rates. The implementation uses the following explicit differentiable surrogate functionals. Write $Z_i = V^\top U_i$, $\widetilde{Z}_{i,c} = V^\top \widetilde{U}_{i,c}^{\mathrm{style}}$, $S_i = (I_d - VV^\top)U_i$, and $\widetilde{S}_{i,c} = (I_d - VV^\top)\widetilde{U}_{i,c}^{\mathrm{style}}$, and let D_{syn} distinguish real from rewrite coordinates and $D_{\mathrm{style}} = c$ denote the prespecified rewrite-template label. With empirical means and covariance matrices denoted by $\bar{Z}$ and $\widehat{\Sigma}_Z$, define

$$\begin{aligned}
\widehat{\mathcal{S}}_{0,k}(V) &= \frac{1}{n_k^{\mathrm{tr}} m} \sum_{i \in I_k^c} \sum_{c=1}^{m} \|V^\top (U_i - \widetilde{U}_{i,c}^{\mathrm{style}})\|_2^2, \\
\widehat{\mathcal{S}}_{1,k}(V) &= \|\bar{Z}_{\mathrm{real}} - \bar{Z}_{\mathrm{style}}\|_2^2 + \|\widehat{\Sigma}_{Z,\mathrm{real}} - \widehat{\Sigma}_{Z,\mathrm{style}}\|_F^2, \\
\widehat{\mathcal{S}}_{2,k}(V) &= -\frac{1}{2}\log\{1 - \widehat{R}^2_{\mathrm{ridge}}(Y;S)\}, \\
\widehat{\mathcal{S}}_{3,k}(V) &= \frac{\sum_{c=1}^m n_c \|\bar{S}_c - \bar{S}\|_2^2}{\varepsilon_{\mathrm{stab}} + \sum_{i \in I_k^c} \sum_{c=1}^m \|\widetilde{S}_{i,c} - \bar{S}\|_2^2}, \qquad \varepsilon_{\mathrm{stab}} = 10^{-12}.
\end{aligned}$$

Here Z_{style} pools the m rewrites within each complete training block, $n_c = n_k^{\mathrm{tr}}$, and $\widehat{R}^2_{\mathrm{ridge}}$ is the clipped centered ridge coefficient of determination computed from $\{(S_i, Y_i) : i \in I_k^c\}$. Thus $\widehat{\mathcal{S}}_{1,k}$ is the moment-based source-discrepancy surrogate, $\widehat{\mathcal{S}}_{2,k}$ is the Gaussian restricted-information surrogate, and $\widehat{\mathcal{S}}_{3,k}$ is the stabilized between-template Fisher ratio. With implementation thresholds $\widetilde{c}_{q,n}$ fixed by the documented pilot or fold-training protocol before held-out evaluation, the actual evaluated constraints are

$$\widetilde{g}_{0,k} = \widehat{\mathcal{S}}_{0,k} - c_{0,n}, \quad \widetilde{g}_{1,k} = \widehat{\mathcal{S}}_{1,k} - \widetilde{c}_{1,n}, \quad \widetilde{g}_{2,k} = \widehat{\mathcal{S}}_{2,k} - \widetilde{c}_{2,n}, \quad \widetilde{g}_{3,k} = \widetilde{c}_{3,n} - \widehat{\mathcal{S}}_{3,k}.$$

Reported constraint margins are these post-fit evaluated quantities, not held-out selection scores. The theoretical targets for the last three constraints are distinct profiled restricted-prediction functionals. The implementation is covered only under the explicit signed surrogate-to-profile value and gradient bridge stated in Lemma C.1; the bridge is not implied by a small auxiliary-optimizer residual. Specifically,

$$\begin{aligned}
(A_1, Z_1(V)) &= (D_{\mathrm{syn}}, Z_c(V)), & (A_2, Z_2(V)) &= (Y, Z_s(V)), \\
(A_3, Z_3(V)) &= (D_{\mathrm{style}}, Z_s(V)).
\end{aligned}$$

For each q, ℓ_q is a proper prediction loss, $\mathcal{D}_{q,0}$ is the baseline class that does not use $Z_q(V)$, and $\mathcal{D}_q \supseteq \mathcal{D}_{q,0}$ may use it. Categorical modules use clipped probabilities; continuous modules use the bounded or moment-controlled loss class stated in the empirical-process assumptions.

For $q \in \{1,2,3\}$, let $L_{q,0,k}(d_0)$ denote the population auxiliary risk of the no-representation baseline and let $L_{q,1,k}(V,d)$ denote the population auxiliary risk using the representation $Z_q(V)$. Their empirical complete-block versions are denoted by $\widehat{L}_{q,0,k}(d_0)$ and $\widehat{L}_{q,1,k}(V,d)$. Define

$$\begin{aligned}
L^\star_{q,0,k} &= \inf_{d_0 \in \mathcal{D}_{q,0}} L_{q,0,k}(d_0), & L^\star_{q,1,k}(V) &= \inf_{d \in \mathcal{D}_q} L_{q,1,k}(V,d), \\
\widehat{L}^\star_{q,0,k} &= \inf_{d_0 \in \mathcal{D}_{q,0}} \widehat{L}_{q,0,k}(d_0), & \widehat{L}^\star_{q,1,k}(V) &= \inf_{d \in \mathcal{D}_q} \widehat{L}_{q,1,k}(V,d).
\end{aligned}$$

Algorithm 1 Complete fold-wise content representation learning procedure.

1: **for** $k = 1, \ldots, K$ **do**
2: Encode the real and paired-rewrite texts and construct the complete-block differences $\Delta U_{i,c} = U_i - \widetilde{U}_{i,c}$.
3: Compute $\widehat{\Sigma}_{\Delta,k}$ and initialize V with the eigenvectors corresponding to its smallest r eigenvalues.
4: Initialize η, multipliers $\lambda_q \geq 0$, and penalty parameter $\rho_{\mathrm{ALM}} > 0$.
5: **repeat**
6: Form the implemented surrogate augmented-Lagrangian objective using $\widehat{\mathcal{R}}^{\mathrm{tr}}_{y,k}$ and the four evaluated constraints $\widetilde{g}_{q,k}$.
7: Approximately minimize this objective over $V \in \mathrm{St}(d, r)$ and η, then update the multipliers and increase ρ_{ALM} when feasibility does not improve.
8: **until** the Riemannian gradient or objective-change criterion is met, or the iteration budget is reached
9: Record the post-fit residual $\varepsilon_{\mathrm{post},k} = \max_q [\widetilde{g}_{q,k}(V)]_+$.
10: Set $\widehat{V}_k$, then refit $\widehat{h}_k$ on I_k^c at $\widehat{V}_k$ with ridge regularization; set $Z_{c,k}(x) = \widehat{V}_k^\top \psi(x)$ and $Z_{s,k}(x) = (I_d - \widehat{V}_k \widehat{V}_k^\top)\psi(x)$.
11: **end for**

The population and exactly profiled empirical restricted-information functionals are

$$\mathcal{J}_{q,k}(V) = L^\star_{q,0,k} - L^\star_{q,1,k}(V), \qquad \widehat{\mathcal{J}}^{\mathrm{prof}}_{q,k}(V) = \widehat{L}^\star_{q,0,k} - \widehat{L}^\star_{q,1,k}(V).$$

For comparison with the explicit surrogates, let $\widetilde{d}_{q,0,k}$ and $\widetilde{d}_{q,1,k,V}$ be auxiliary predictors returned by a numerical profiling procedure, and set

$$\breve{\mathcal{J}}^{\mathrm{num}}_{q,k}(V) = \widehat{L}_{q,0,k}(\widetilde{d}_{q,0,k}) - \widehat{L}_{q,1,k}(V, \widetilde{d}_{q,1,k,V}).$$

Write $\widehat{L}_{\mathrm{aug,k}}(V) := \widehat{\mathcal{S}}_{0,k}(V)$. Define the exactly profiled empirical constraints

$$\begin{aligned}
\widehat{g}^{\mathrm{prof}}_{0,k}(V) &= \widehat{L}_{\mathrm{aug,k}}(V) - c_{0,n}, \\
\widehat{g}^{\mathrm{prof}}_{1,k}(V) &= \widehat{\mathcal{J}}^{\mathrm{prof}}_{1,k}(V) - c_{1,n}, \\
\widehat{g}^{\mathrm{prof}}_{2,k}(V) &= \widehat{\mathcal{J}}^{\mathrm{prof}}_{2,k}(V) - c_{2,n}, \\
\widehat{g}^{\mathrm{prof}}_{3,k}(V) &= c_{3,n} - \widehat{\mathcal{J}}^{\mathrm{prof}}_{3,k}(V),
\end{aligned}$$

and the numerically profiled constraints

$$\begin{aligned}
\breve{g}^{\mathrm{num}}_{0,k}(V) &= \widehat{g}^{\mathrm{prof}}_{0,k}(V), \\
\breve{g}^{\mathrm{num}}_{1,k}(V) &= \breve{\mathcal{J}}^{\mathrm{num}}_{1,k}(V) - c_{1,n}, \\
\breve{g}^{\mathrm{num}}_{2,k}(V) &= \breve{\mathcal{J}}^{\mathrm{num}}_{2,k}(V) - c_{2,n}, \\
\breve{g}^{\mathrm{num}}_{3,k}(V) &= c_{3,n} - \breve{\mathcal{J}}^{\mathrm{num}}_{3,k}(V).
\end{aligned}$$

The population constraints are

$$\begin{aligned}
g_{0,k}(V) &= L_{\mathrm{aug}}(VV^\top) - c_{0,n}, \\
g_{q,k}(V) &= \mathcal{J}_{q,k}(V) - c_{q,n}, \quad q = 1, 2, \\
g_{3,k}(V) &= c_{3,n} - \mathcal{J}_{3,k}(V).
\end{aligned}$$

Thus $\widetilde{g}_{q,k}$, $\breve{g}^{\mathrm{num}}_{q,k}$, $\widehat{g}^{\mathrm{prof}}_{q,k}$, and $g_{q,k}$ denote, respectively, the implemented surrogate, numerically profiled empirical, exactly profiled empirical, and population constraints; no two levels share a symbol.

The implementation optimizes the explicit surrogate constraints $\widetilde{g}_{q,k}$; numerical and exact proper-loss profiles are theoretical reference layers. Theorem 2 is conditional on the signed implementation-to-profile bridge and the resulting risk and feasibility envelopes. A large $\varepsilon_{\mathrm{post},k}$ therefore removes the local-theory certificate even when the numerical fit itself is retained. The simulation uses 0.03

only as a post-fit diagnostic threshold and reports every prespecified fit, including the ordinal boundary rows. This diagnostic does not estimate the implementation-to-profile bridge. It is not used to choose the reported interval. A switch to the classical sample-mean interval based on the same evaluation data would require an independent-pilot, simultaneous, or selective-inference justification beyond the present theory.

For $\lambda = (\lambda_0, \dots, \lambda_3)^\top \geq 0$ and $\rho_{\mathrm{ALM}} > 0$, define $\phi_{\rho_{\mathrm{ALM}},\lambda_q}(t) = \frac{[\lambda_q + \rho_{\mathrm{ALM}} t]_+^2 - \lambda_q^2}{2\rho_{\mathrm{ALM}}}$. Define the four augmented-Lagrangian objectives by

$$\widehat{A}^{\mathrm{prof}}_{\rho_{\mathrm{ALM}},k}(\theta,\lambda) = \widehat{\mathcal{R}}^{\mathrm{tr}}_{y,k}(\theta) + \sum_{q=0}^{3} \phi_{\rho_{\mathrm{ALM}},\lambda_q}\{\widehat{g}^{\mathrm{prof}}_{q,k}(V)\},$$

$$\breve{A}^{\mathrm{num}}_{\rho_{\mathrm{ALM}},k}(\theta,\lambda) = \widehat{\mathcal{R}}^{\mathrm{tr}}_{y,k}(\theta) + \sum_{q=0}^{3} \phi_{\rho_{\mathrm{ALM}},\lambda_q}\{\breve{g}^{\mathrm{num}}_{q,k}(V)\},$$

$$\widetilde{A}_{\rho_{\mathrm{ALM}},k}(\theta,\lambda) = \widehat{\mathcal{R}}^{\mathrm{tr}}_{y,k}(\theta) + \sum_{q=0}^{3} \phi_{\rho_{\mathrm{ALM}},\lambda_q}\{\widetilde{g}_{q,k}(V)\},$$

$$A_{\rho_{\mathrm{ALM}},k}(\theta,\lambda) = \mathcal{R}^{\mathrm{tr}}_{y,k}(\theta) + \sum_{q=0}^{3} \phi_{\rho_{\mathrm{ALM}},\lambda_q}\{g_{q,k}(V)\}, \qquad \theta = (V,\eta).$$

To define the population target of the training criterion without selecting an arbitrary orientation of the pseudo-true eigenspace, assume that the predictor class is rotation compatible. Define

$$\mathcal{R}^{\mathrm{tr},\dagger}_{y,k} = \inf_{\theta \in \Theta^\dagger_k(r_0)} \mathcal{R}^{\mathrm{tr}}_{y,k}(\theta),$$

and let $\theta^\dagger_k$ be a measurable minimizer. An approximate minimizer may be used by adding its criterion gap to $\alpha_{\mathrm{app},n}$. The comparison quantities below are defined using the same outcome criterion.

C.2 Assumption B: Profile Regularity, Rotation Compatibility, and the Implementation Bridge

Assumption B (profile regularity, rotation compatibility, and implementation bridge). For each fold k, the following common local conditions hold.

(i) *Auxiliary profile regularity.* In a smooth local coordinate chart ϑ of $\mathrm{St}(d,r)$, the population and empirical auxiliary risks are twice continuously differentiable near their selected minimizers. The selected branches $d^\star_{q,k}(\vartheta)$ and $\widehat{d}^\star_{q,k}(\vartheta)$ are locally unique, their auxiliary Hessians satisfy

$$\nabla^2_{dd} L_{q,1,k}\{\vartheta, d^\star_{q,k}(\vartheta)\} \succeq \kappa_{d,q} I, \qquad \nabla^2_{dd} \widehat{L}_{q,a,k} \succeq \widehat{\kappa}_{d,q,a} I,$$

with the lower-curvature constants bounded away from zero, and the mixed derivatives are locally bounded.

(ii) *Rotation compatibility.* Write the finite-dimensional parameter of the fold-k outcome learner as $\eta \in \mathbb{R}^{p_{h,k}}$, so that the prediction rule is $h_\eta(V^\top U)$. For every $R \in \mathbb{O}(r)$ there is a smooth orthogonal action ρ_R on the predictor parameter such that $s_{\rho_R(\eta)}(R^\top z) = s_\eta(z)$ and $h_{\rho_R(\eta)}(R^\top z) = h_\eta(z)$. For second-order learners, the theoretical feature map uses orthonormal symmetric-tensor coordinates $\mathrm{svec}(zz^\top)$, with off-diagonal monomials scaled by $\sqrt{2}$. Orthogonal rotation of z then induces an orthogonal action on the quadratic block. The unscaled monomials in the implementation differ by a fixed invertible diagonal reparameterization, which changes local curvature constants but not the represented functions or the existence of the local PL bound. For multinomial learners, we use a reference-class or centered-logit gauge; the full-softmax implementation is identified with its centered representative because common logit shifts leave both probabilities and loss unchanged. The auxiliary predictor classes are closed under the corresponding orthogonal changes of coordinates. Hence the exactly profiled empirical ALM objective is invariant under $(V,\eta) \mapsto (VR, \rho_R(\eta))$.

(iii) *Implemented-surrogate bridge.* Let $\Gamma_{k,n}$ contain $\widehat{V}_k$ and the projection component of every point on every accepted retraction segment of the fold-(k) algorithm. On an event $\mathcal{E}_{\mathrm{sur},n}$, for $q = 1, 2, 3$,

$$\sup_{V \in \Gamma_{k,n}} |\widetilde{g}_{q,k}(V) - \widehat{g}^{\mathrm{prof}}_{q,k}(V)| \le \eta^{\mathrm{sur,val}}_{q,n}, \tag{C.1}$$

$$\| \operatorname{grad} \widetilde{g}_{q,k}(\widehat{V}_k) - \operatorname{grad} \widehat{g}^{\mathrm{prof}}_{q,k}(\widehat{V}_k)\| \le \eta^{\mathrm{sur,grad}}_{q,n}. \tag{C.2}$$

The signed discrepancies include the implementation–profile threshold difference and the reversed sign of the $q = 3$ constraint; the $q = 0$ discrepancy is zero. Along these paths the implemented and profiled constraint values and gradients are bounded by B_g and $G_{g,q}$, respectively. Accepted segments and the returned point lie within one smooth branch of the clipped R^2 surrogate, and the stabilized Fisher denominator is positive. These bridge envelopes are independent assumptions about the implemented surrogates, not consequences of an auxiliary optimizer's stopping residual.

For the comparison step in Theorem 2, let $V^\dagger$ be any orthonormal basis of $\Pi^\dagger$ and define

$$\eta^0_{k,\dagger} \in \arg\min_{\eta} \mathcal{R}^{\mathrm{tr}}_{y,k}(V^\dagger, \eta), \qquad \theta^{\mathrm{cmp}}_k = (V^\dagger, \eta^0_{k,\dagger}),$$

$$s^{\mathrm{cmp}}_{q,n} := \max_{1\le k\le K} [g_{q,k}(V^\dagger)]_+, \qquad \alpha_{\mathrm{app},n} := \max_{1\le k\le K} \{\mathcal{R}^{\mathrm{tr}}_{y,k}(\theta^{\mathrm{cmp}}_k) - \mathcal{R}^{\mathrm{tr}}_{y,k}(\theta^\dagger_k)\} \ge 0.$$

For the final multipliers define the uniform comparison penalty

$$S^{\mathrm{cmp}}_{\mathrm{ALM},n} := \max_{1\le k\le K} \sum_{q=0}^{3} \phi_{\rho_{\mathrm{ALM},n}, \lambda^{\mathrm{fin}}_{q,k}}(s^{\mathrm{cmp}}_{q,n}).$$

These are definitions, not assumptions.

C.3 Profiled Auxiliary Objectives and Complete-block Empirical Processes

Lemma C.1 (Profile smoothness and implementation-to-profile error)**.** *Let ϑ be the local coordinate of V in Assumption B(i). The population value function*

$$m_{q,k}(\vartheta) = \inf_{d \in \mathcal{D}_q} L_{q,1,k}(\vartheta, d)$$

is twice continuously differentiable and

$$\nabla_\vartheta m_{q,k}(\vartheta) = \nabla_\vartheta L_{q,1,k}\{\vartheta, d^\star_{q,k}(\vartheta)\}, \tag{C.3}$$

$$\nabla^2_{\vartheta\vartheta} m_{q,k}(\vartheta) = L_{\vartheta\vartheta} - L_{\vartheta d} L^{-1}_{dd} L_{d\vartheta}, \tag{C.4}$$

where all derivative blocks are evaluated at $\{\vartheta, d^\star_{q,k}(\vartheta)\}$. The corresponding Riemannian derivatives are obtained by transporting the local-coordinate bilinear form to the tangent bundle. The same formulas hold for the exactly profiled empirical functional.

Let $\mathcal{V}^{\mathrm{alg}}_{k,n}$ denote the finite collection of returned auxiliary-profile evaluation points, including $\widehat{V}_k$. Suppose that at these points

$$\|\nabla_d \widehat{L}_{q,0,k}(\widetilde{d}_{q,0,k})\| \le \varepsilon^{\mathrm{aux}}_{q,0,n}, \qquad \|\nabla_d \widehat{L}_{q,1,k}(V, \widetilde{d}_{q,1,k,V})\| \le \varepsilon^{\mathrm{aux}}_{q,1,n},$$

and suppose the multipliers are bounded by Λ_n. For every auxiliary fit used here, additionally suppose that the returned parameter $\widetilde{d}_{q,a,k,V}$ and its selected exact empirical minimizer $\widehat{d}^\star_{q,a,k,V}$ lie in the same convex neighborhood on the selected local branch, and that

$$\nabla^2_{dd} \widehat{L}_{q,a,k}(V, d) \succeq \widehat{\kappa}_{d,q,a} I$$

throughout the line segment joining them. For the baseline fit $a = 0$, the notation suppresses the irrelevant V. Define

$$\eta^{\mathrm{aux}}_{q,a,n} = \frac{(\varepsilon^{\mathrm{aux}}_{q,a,n})^2}{2\widehat{\kappa}_{d,q,a}}, \qquad \eta^{\mathrm{aux}}_{\nabla g,q,n} = \frac{L_{\vartheta d,q} \varepsilon^{\mathrm{aux}}_{q,1,n}}{\widehat{\kappa}_{d,q,1}}.$$

Then

$$\max_{V\in\mathcal{V}_{k,n}^{\mathrm{alg}}} |\breve{\mathcal{J}}_{q,k}^{\mathrm{num}}(V) - \widehat{\mathcal{J}}_{q,k}^{\mathrm{prof}}(V)| \le \eta_{q,0,n}^{\mathrm{aux}} + \eta_{q,1,n}^{\mathrm{aux}}. \tag{C.5}$$

Consequently,

$$\sup_{V\in\Gamma_{k,n}} |\widetilde{g}_{q,k}(V) - \widehat{g}_{q,k}^{\mathrm{prof}}(V)| \le \eta_{q,n}^{\mathrm{sur,val}}, \qquad \|\operatorname{grad}\widetilde{g}_{q,k}(V) - \operatorname{grad}\widehat{g}_{q,k}^{\mathrm{prof}}(V)\|\big|_{V=\widehat{V}_k} \le \eta_{q,n}^{\mathrm{sur,grad}}. \tag{C.6}$$

Along every accepted segment,

$$|\widetilde{A}_{\rho_{\mathrm{ALM},n},k} - \widehat{A}_{\rho_{\mathrm{ALM},n},k}^{\mathrm{prof}}| \le \eta_{A,n}^{\mathrm{impl}}, \qquad \eta_{A,n}^{\mathrm{impl}} = \sum_{q=1}^{3}(\Lambda_n + \rho_{\mathrm{ALM},n}B_g)\eta_{q,n}^{\mathrm{sur,val}}, \tag{C.7}$$

and, at the returned iterate,

$$\left\|\operatorname{grad}\widetilde{A}_{\rho_{\mathrm{ALM},n},k} - \operatorname{grad}\widehat{A}_{\rho_{\mathrm{ALM},n},k}^{\mathrm{prof}}\right\| \le \eta_{\nabla,n}^{\mathrm{impl}}, \tag{C.8}$$

where

$$\eta_{\nabla,n}^{\mathrm{impl}} = \sum_{q=1}^{3}\left[\rho_{\mathrm{ALM},n}G_{g,q}\eta_{q,n}^{\mathrm{sur,val}} + (\Lambda_n + \rho_{\mathrm{ALM},n}B_g)\eta_{q,n}^{\mathrm{sur,grad}}\right].$$

Proof. Write

$$F_q(\vartheta, d) = \nabla_d L_{q,1,k}(\vartheta, d).$$

At $d = d_{q,k}^\star(\vartheta)$, $F_q\{\vartheta, d_{q,k}^\star(\vartheta)\} = 0$. Assumption B(i) makes $D_d F_q = L_{dd}$ invertible, so the implicit-function theorem gives

$$D_\vartheta d_{q,k}^\star(\vartheta) = -L_{dd}^{-1}L_{d\vartheta}.$$

Differentiate $m_{q,k}(\vartheta) = L_{q,1,k}\{\vartheta, d_{q,k}^\star(\vartheta)\}$:

$$D_\vartheta m_{q,k} = L_\vartheta + L_d D_\vartheta d_{q,k}^\star = L_\vartheta,$$

because $L_d = 0$ at the minimizer. A second derivative gives

$$D_{\vartheta\vartheta}^2 m_{q,k} = L_{\vartheta\vartheta} + L_{\vartheta d}D_\vartheta d_{q,k}^\star = L_{\vartheta\vartheta} - L_{\vartheta d}L_{dd}^{-1}L_{d\vartheta}.$$

This proves (C.3)–(C.4). Since the baseline optimum does not depend on V, the formulas for $\mathcal{J}_{q,k}$ follow with a minus sign. The empirical proof is identical.

Let $\widehat{d}^\star$ denote the selected exact empirical auxiliary minimizer and let $\widetilde{d}$ be the returned auxiliary parameter. The same-branch, segment-wise strong-convexity condition gives

$$\widehat{L}(\widehat{d}^\star) \ge \widehat{L}(\widetilde{d}) + \langle\nabla_d\widehat{L}(\widetilde{d}), \widehat{d}^\star - \widetilde{d}\rangle + \frac{\widehat{\kappa}_d}{2}\|\widehat{d}^\star - \widetilde{d}\|^2.$$

Maximizing the resulting quadratic upper bound over $\widehat{d}^\star - \widetilde{d}$ yields

$$0 \le \widehat{L}(\widetilde{d}) - \widehat{L}(\widehat{d}^\star) \le \frac{\|\nabla_d\widehat{L}(\widetilde{d})\|^2}{2\widehat{\kappa}_d}.$$

Strong monotonicity also gives

$$\|\widetilde{d} - \widehat{d}^\star\| \le \frac{\|\nabla_d\widehat{L}(\widetilde{d})\|}{\widehat{\kappa}_d}.$$

Applying these inequalities to the baseline and representation-dependent auxiliary risks proves (C.5). The mixed-derivative bound separately gives

$$\|\nabla_\vartheta\breve{g}_{q,k}^{\mathrm{num}} - \nabla_\vartheta\widehat{g}_{q,k}^{\mathrm{prof}}\| \le \eta_{\nabla g,q,n}^{\mathrm{aux}}.$$

These two bounds concern only the numerically profiled proper-loss reference. The implementation gap (C.6) is exactly Assumption B(iii), independently of those auxiliary-optimization residuals.

On the bounded local range, $\phi_{\rho,\lambda}$ is $(\Lambda_n + \rho_{\mathrm{ALM},n}B_g)$-Lipschitz, which proves (C.7). Finally, $t \mapsto [\lambda + \rho t]_+$ is ρ-Lipschitz. Applying the product rule to each augmented constraint term,

$$\begin{aligned}&\|\operatorname{grad}(\phi_{\rho_{\mathrm{ALM},n},\lambda_q}\circ\widetilde{g}_{q,k}) - \operatorname{grad}(\phi_{\rho_{\mathrm{ALM},n},\lambda_q}\circ\widehat{g}_{q,k}^{\mathrm{prof}})\| \\ &\quad\le \rho_{\mathrm{ALM},n}G_{g,q}|\widetilde{g}_{q,k} - \widehat{g}_{q,k}^{\mathrm{prof}}| + (\Lambda_n + \rho_{\mathrm{ALM},n}B_g)\|\nabla\widetilde{g}_{q,k} - \nabla\widehat{g}_{q,k}^{\mathrm{prof}}\|.\end{aligned}$$

Summing over $q = 1, 2, 3$ proves (C.8). □

Lemma C.2 (Uniform complete-block concentration and bridge transfer). *Let $\mathcal{A} = \{y, 0\} \cup \{(q, a) : q = 1, 2, 3,\ a = 0, 1\}$ index the main risk, augmentation loss, and six auxiliary risks. For $a \in \mathcal{A}$, suppose the corresponding complete-block function class $\mathcal{F}_{a,k}$ is nonnegative and has oscillation envelope B_a, so $0 \le f(C) \le B_a$ for every block C and every $f \in \mathcal{F}_{a,k}$. Let its Rademacher complexity be*

$$\mathfrak{R}_{a,k} = \mathbb{E}_{C,\sigma}\left[\sup_{f\in\mathcal{F}_{a,k}}\left|\frac{1}{n_k^{\mathrm{tr}}}\sum_{i\in I_k^c}\sigma_i f(C_{a,k,i})\right|\right].$$

For $0 < \delta < 1$, set

$$\xi_{a,n}(\delta) = \max_k\left[2\mathfrak{R}_{a,k} + B_a\sqrt{\frac{\log(16K/\delta)}{2n_k^{\mathrm{tr}}}}\right].$$

Then, with probability at least $1-\delta$, simultaneously over all folds,

$$\sup_{\theta\in\Theta_k^\dagger(r_0)}|\widehat{\mathcal{R}}_{y,k}^{\mathrm{tr}}(\theta) - \mathcal{R}_{y,k}^{\mathrm{tr}}(\theta)| \le \xi_{y,n}, \tag{C.9}$$

$$\sup_V|\widehat{g}_{0,k}^{\mathrm{prof}}(V) - g_{0,k}(V)| = \sup_V|\widetilde{g}_{0,k}(V) - g_{0,k}(V)| \le \xi_{0,n}, \tag{C.10}$$

$$\sup_V|\widehat{g}_{q,k}^{\mathrm{prof}}(V) - g_{q,k}(V)| \le \xi_{g,q,n}^{\mathrm{prof}}, \qquad q = 1, 2, 3, \tag{C.11}$$

$$\sup_{V\in\Gamma_{k,n}}|\widetilde{g}_{q,k}(V) - g_{q,k}(V)| \le \xi_{g,q,n}^{\mathrm{total}}, \qquad q = 1, 2, 3, \tag{C.12}$$

where

$$\xi_{g,q,n}^{\mathrm{prof}} = \xi_{q,0,n} + \xi_{q,1,n}, \tag{C.13}$$

$$\xi_{g,q,n}^{\mathrm{total}} = \xi_{q,0,n} + \xi_{q,1,n} + \eta_{q,n}^{\mathrm{sur,val}}. \tag{C.14}$$

Thus the profile empirical-process error and the implementation-to-profile error are kept separate.

Proof. Fix one class $\mathcal{F}_{a,k}$. Let

$$Z_{a,k}^+ = \sup_{f\in\mathcal{F}_{a,k}}\{P_{a,k}f - \widehat{P}_{a,k}f\}.$$

Replacing one complete block changes $\widehat{P}_{a,k}f$ by at most B_a/n_k^{tr} uniformly in f. McDiarmid's inequality therefore gives

$$\Pr\{Z_{a,k}^+ - \mathbb{E}Z_{a,k}^+ \ge t\} \le \exp\left(-\frac{2n_k^{\mathrm{tr}}t^2}{B_a^2}\right).$$

Let $C'_{a,k,i}$ be an independent ghost block. By symmetrization,

$$\begin{aligned}\mathbb{E}Z_{a,k}^+ &\le \mathbb{E}_{C,C'}\sup_f\left|\frac{1}{n_k^{\mathrm{tr}}}\sum_i\{f(C'_{a,k,i}) - f(C_{a,k,i})\}\right| \\ &= \mathbb{E}_{C,C',\sigma}\sup_f\left|\frac{1}{n_k^{\mathrm{tr}}}\sum_i\sigma_i\{f(C'_{a,k,i}) - f(C_{a,k,i})\}\right| \\ &\le 2\mathfrak{R}_{a,k}.\end{aligned}$$

The reverse deviation has the same bound. Hence

$$\Pr\left[\sup_{f\in\mathcal{F}_{a,k}}|\widehat{P}_{a,k}f - P_{a,k}f| > 2\mathfrak{R}_{a,k} + t\right] \le 2\exp\left(-\frac{2n_k^{\mathrm{tr}}t^2}{B_a^2}\right).$$

Taking

$$t = B_a\sqrt{\frac{\log(16K/\delta)}{2n_k^{\mathrm{tr}}}}$$

and applying a union bound over the eight classes and K folds proves the simultaneous risk bounds. In particular, the augmentation envelope is $B_0 = M_\Delta^2$, rather than one.

For $q \in \{1, 2, 3\}$, the elementary infimum perturbation inequality gives

$$\begin{aligned}\sup_V |\widehat{\mathcal{J}}_{q,k}^{\mathrm{prof}}(V) - \mathcal{J}_{q,k}(V)| &\le \sup_{d_0} |\widehat{L}_{q,0,k}(d_0) - L_{q,0,k}(d_0)| \\ &\quad + \sup_{V,d} |\widehat{L}_{q,1,k}(V,d) - L_{q,1,k}(V,d)| \\ &\le \xi_{q,0,n} + \xi_{q,1,n}.\end{aligned}$$

Indeed, if $\sup_f |\widehat{L}(f) - L(f)| \le \xi$, then both

$$\inf_f \widehat{L}(f) \le \inf_f L(f) + \xi \quad \text{and} \quad \inf_f L(f) \le \inf_f \widehat{L}(f) + \xi,$$

so the two infima differ by at most ξ. Finally, Lemma C.1 gives

$$\sup_{V \in \Gamma_{k,n}} |\widetilde{g}_{q,k}(V) - \widehat{g}_{q,k}^{\mathrm{prof}}(V)| \le \eta_{q,n}^{\mathrm{sur,val}}.$$

The first profile inequality yields (C.11)–(C.13). Combining it with the implementation gap (C.6) yields (C.12)–(C.14). The threshold discrepancy is already included in the signed bridge envelope. □

C.4 Morse–Bott Geometry and a Derived Local PL Inequality

Lemma C.3 (Local curvature of the unpenalized outcome loss)**.** *Conditional on the fold-k training information, suppose that the outcome learner has a finite-dimensional feature map $\Phi_{k,i}(V) = \Phi_k(V^\top U_i)$, including an intercept coordinate when present; write $\Phi_k(V) = \Phi_k(V^\top U)$ for a generic draw. Suppose the loss is convex and twice continuously differentiable in its linear predictor. For a scalar linear predictor, define the empirical and population outcome Hessians on the local tube $\mathcal{U}_{k,n}$ by*

$$\begin{aligned}\widehat{H}_{h,k}(V,\eta) &:= \nabla_{\eta\eta}^2 \widehat{\mathcal{R}}_{y,k}^{\mathrm{tr}}(V,\eta) = \frac{1}{n_k^{\mathrm{tr}}} \sum_{i \in I_k^c} \omega_{k,i}(V,\eta) \Phi_{k,i}(V) \Phi_{k,i}(V)^\top, \\ H_{h,k}(V,\eta) &:= \nabla_{\eta\eta}^2 \mathcal{R}_{y,k}^{\mathrm{tr}}(V,\eta) = \mathbb{E}_P\left[\omega_k(V,\eta) \Phi_k(V) \Phi_k(V)^\top\right],\end{aligned} \tag{C.15}$$

where $\omega_{k,i}(V,\eta) \ge 0$. Suppose that, for some $\kappa_{h,k} > 0$,

$$\inf_{(V,\eta) \in \mathcal{U}_{k,n}} \lambda_{\min}\{H_{h,k}(V,\eta)\} \ge 2\kappa_{h,k}, \qquad \sup_{(V,\eta) \in \mathcal{U}_{k,n}} \|\widehat{H}_{h,k}(V,\eta) - H_{h,k}(V,\eta)\|_{\mathrm{op}} \le \kappa_{h,k}. \tag{C.16}$$

Then, uniformly on $\mathcal{U}_{k,n}$,

$$H_{hh,k} = \nabla_{\eta\eta}^2 F_{k,n}^{\mathrm{prof}}(V,\eta) = \widehat{H}_{h,k}(V,\eta) \succeq \kappa_{h,k} I.$$

For squared loss, a sufficient empirical condition is $\lambda_{\min}\{(n_k^{\mathrm{tr}})^{-1} \sum_i \Phi_{k,i}(V) \Phi_{k,i}(V)^\top\} \ge c_{\Phi,k} > 0$. For binary cross-entropy, the same condition together with, uniformly over all i and $(V,\eta) \in \mathcal{U}_{k,n}$, $\epsilon \le p_{k,i}(V,\eta) \le 1 - \epsilon$ gives curvature at least $\epsilon(1-\epsilon)c_{\Phi,k}$. For multinomial cross-entropy, the corresponding conclusion holds under an identifiable reference-class parameterization, the uniform bound $\min_a p_{k,i,a}(V,\eta) \ge \epsilon$ for all i and $(V,\eta) \in \mathcal{U}_{k,n}$, and the same Gram-matrix condition for the expanded features.

Proof. For squared loss with $h_\eta(V^\top U_i) = \eta^\top \Phi_{k,i}(V)$, direct differentiation gives

$$\begin{aligned}\nabla_\eta \widehat{\mathcal{R}}_{y,k}^{\mathrm{tr}}(V,\eta) &= -\frac{2}{n_k^{\mathrm{tr}}} \sum_{i \in I_k^c} \{Y_i - \eta^\top \Phi_{k,i}(V)\} \Phi_{k,i}(V), \\ \nabla_{\eta\eta}^2 \widehat{\mathcal{R}}_{y,k}^{\mathrm{tr}}(V,\eta) &= \frac{2}{n_k^{\mathrm{tr}}} \sum_{i \in I_k^c} \Phi_{k,i}(V) \Phi_{k,i}(V)^\top.\end{aligned}$$

Thus $\omega_{k,i} = 2$, and the empirical Gram lower bound yields $\widehat{H}_{h,k}(V,\eta) \succeq 2c_{\Phi,k}I$.

For binary cross-entropy, write

$$p_{k,i} = \{1 + \exp[-\eta^\top \Phi_{k,i}(V)]\}^{-1}, \qquad \ell_i(\eta) = -Y_i \log p_{k,i} - (1 - Y_i)\log(1 - p_{k,i}).$$

Since $\nabla_\eta p_{k,i} = p_{k,i}(1 - p_{k,i})\Phi_{k,i}(V)$,

$$\nabla_\eta \ell_i(\eta) = (p_{k,i} - Y_i)\Phi_{k,i}(V),$$
$$\nabla^2_{\eta\eta} \ell_i(\eta) = p_{k,i}(1 - p_{k,i})\Phi_{k,i}(V)\Phi_{k,i}(V)^\top.$$

If $p_{k,i} \in [\epsilon, 1 - \epsilon]$, summing the last display gives

$$\widehat{H}_{h,k}(V,\eta) \succeq \epsilon(1-\epsilon)\left\{\frac{1}{n_k^{\mathrm{tr}}} \sum_{i \in I_k^c} \Phi_{k,i}(V)\Phi_{k,i}(V)^\top\right\} \succeq \epsilon(1-\epsilon)c_{\Phi,k}I.$$

For C-class cross-entropy, use class C as the reference and write $p_{k,i,-C}$ for the first $C-1$ probabilities. The Hessian contribution is

$$W_{k,i} \otimes \Phi_{k,i}(V)\Phi_{k,i}(V)^\top, \qquad W_{k,i} = \mathrm{Diag}(p_{k,i,-C}) - p_{k,i,-C}p_{k,i,-C}^\top.$$

For any $a \in \mathbb{R}^{C-1}$, set $a_C = 0$. The covariance identity and the terms involving the reference class give

$$\begin{aligned} a^\top W_{k,i} a &= \frac{1}{2}\sum_{j=1}^{C}\sum_{\ell=1}^{C} p_{k,i,j}p_{k,i,\ell}(a_j - a_\ell)^2 \\ &\geq p_{k,i,C}\sum_{j=1}^{C-1} p_{k,i,j}a_j^2 \geq \epsilon^2\|a\|^2. \end{aligned}$$

Hence $W_{k,i} \succeq \epsilon^2 I_{C-1}$, and the expanded-feature Gram bound yields

$$\widehat{H}_{h,k}(V,\eta) = \frac{1}{n_k^{\mathrm{tr}}}\sum_{i \in I_k^c} W_{k,i} \otimes \Phi_{k,i}(V)\Phi_{k,i}(V)^\top \succeq \epsilon^2 c_{\Phi,k}I.$$

Polynomial features are covered by taking Φ_k to be the corresponding finite expansion.

Finally, Weyl's inequality and (C.16) imply, uniformly on $\mathcal{U}_{k,n}$,

$$\begin{aligned} \lambda_{\min}\{\widehat{H}_{h,k}(V,\eta)\} &\geq \lambda_{\min}\{H_{h,k}(V,\eta)\} - \|\widehat{H}_{h,k}(V,\eta) - H_{h,k}(V,\eta)\|_{\mathrm{op}} \\ &\geq 2\kappa_{h,k} - \kappa_{h,k} = \kappa_{h,k}. \end{aligned}$$

The profiled constraints depend on V but not on the main outcome parameter η; consequently their $\eta\eta$ Hessian blocks vanish, and the predictor block of $F_{k,n}^{\mathrm{prof}}$ equals $\widehat{H}_{h,k}$. This proves the stated local curvature bound. □

Lemma C.4 (Local Procrustes slice for the rotation orbit)**.** *Let $\theta_k^{\mathrm{loc}} = (V_k^{\mathrm{loc}}, \eta_k^{\mathrm{loc}})$, and suppose the rotation compatibility in Assumption B(ii) holds. In a sufficiently small neighborhood of V_k^{loc}, every rotation orbit has a unique representative in the local slice*

$$\mathcal{C}_k := \{(V,\eta) : (V_k^{\mathrm{loc}})^\top V \text{ is symmetric positive definite}\}.$$

The slice representative is obtained by a smooth map

$$(V,\eta) \mapsto \{VR(V), \rho_{R(V)}(\eta)\},$$

with $R(V) \in \mathbb{O}(r)$, and it leaves $F_{k,n}^{\mathrm{prof}}$ unchanged. Moreover, every tangent vector $\xi = (\xi_V, \xi_\eta)$ to $\mathcal{C}_k$ at θ_k^{loc} satisfies

$$(V_k^{\mathrm{loc}})^\top \xi_V = 0.$$

Thus the V-component of a slice tangent is a horizontal Grassmann direction, and normal-space curvature can be checked on the slice and then lifted to the full rotation-invariant neighborhood.

Proof. For V close to V_k^{loc}, the matrix $A(V) = (V_k^{\mathrm{loc}})^\top V$ is nonsingular. Let $A(V) = U(V)H(V)$ be its polar decomposition, where $U(V) \in \mathbb{O}(r)$ and $H(V)$ is symmetric positive definite. Set $R(V) = U(V)^\top$. Then

$$(V_k^{\mathrm{loc}})^\top V R(V) = U(V)H(V)U(V)^\top$$

is symmetric positive definite, so the transformed point lies in $\mathcal{C}_k$. The local uniqueness and smoothness follow from the local uniqueness and smoothness of the polar decomposition near nonsingular matrices. Rotation compatibility gives

$$F_{k,n}^{\mathrm{prof}}\{VR(V), \rho_{R(V)}(\eta)\} = F_{k,n}^{\mathrm{prof}}(V, \eta).$$

Finally, if $V(t)$ is a differentiable curve in $\mathcal{C}_k$ with $V(0) = V_k^{\mathrm{loc}}$ and derivative ξ_V, then $(V_k^{\mathrm{loc}})^\top \xi_V$ is symmetric by the slice condition. The Stiefel tangent condition gives

$$(V_k^{\mathrm{loc}})^\top \xi_V + \xi_V^\top V_k^{\mathrm{loc}} = 0,$$

so the same matrix is also skew-symmetric. Hence $(V_k^{\mathrm{loc}})^\top \xi_V = 0$. This proves the horizontal-slice claim and the quotient-to-slice reduction. □

Proposition C.1 (Complete local curvature and derived PL inequality)**.** *Fix a fold k and let*

$$F_{k,n}^{\mathrm{prof}}(V, \eta) = \widehat{A}_{\rho_{\mathrm{ALM},n},k}^{\mathrm{prof}}\{(V, \eta), \lambda_k^{\mathrm{fin}}\}$$

be the exactly profiled empirical ALM objective on $\mathcal{M}_k = \mathrm{St}(d, r) \times \mathbb{R}^{p_{h,k}}$. Let $\theta_k^{\mathrm{loc}} = (V_k^{\mathrm{loc}}, \eta_k^{\mathrm{loc}})$ be a local minimizer and let

$$\mathcal{S}_k = \{(V_k^{\mathrm{loc}} R, \rho_R(\eta_k^{\mathrm{loc}})) : R \in \mathbb{O}(r)\}$$

be its rotation orbit. Suppose the following conditions hold on a tubular neighborhood $\mathcal{U}_{k,n}$ of $\mathcal{S}_k$. Shrink the tube, if necessary, so that its local minimum is attained exactly on this orbit, and define

$$(F_{k,n}^{\mathrm{prof}})_{\mathrm{loc}}^\star := inf_{\theta \in \mathcal{U}_{k,n}} F_{k,n}^{\mathrm{prof}}(\theta) = F_{k,n}^{\mathrm{prof}}(\theta_k^{\mathrm{loc}}).$$

Normal-space curvature is evaluated through the local Procrustes slice $\mathcal{C}_k$ of Lemma C.4; in this slice, the V-component of every tangent representative is horizontal.

- *(a) The active set of the positive-part ALM penalties is constant on $\mathcal{U}_{k,n}$, so $F_{k,n}^{\mathrm{prof}}$ is C^2 there. The only local nonidentifiability is the orthogonal rotation in Assumption B(ii), so the local minimizer set of the same value is exactly $\mathcal{S}_k$.*
- *(b) Let $V_{\Delta,k}$ span the bottom-r eigenspace of $\widehat{\Sigma}_{\Delta,k}$ and let $\widehat{\gamma}_{\Delta,k} = \lambda_{r+1}(\widehat{\Sigma}_{\Delta,k}) - \lambda_r(\widehat{\Sigma}_{\Delta,k}) > 0$. The Riemannian Hessian of $\widehat{L}_{\mathrm{aug,k}}$ is locally Lipschitz with constant $L_{\Delta,k}$, and $d_{\mathcal{G}}(V_k^{\mathrm{loc}}, V_{\Delta,k}) \le r_{V,k}$, with $\gamma_{0,k} := 2\widehat{\gamma}_{\Delta,k} - L_{\Delta,k} r_{V,k} > 0$. Let $a_{0,k} = [\lambda_{0,k}^{\mathrm{fin}} + \rho_{\mathrm{ALM},n} \widehat{g}_{0,k}^{\mathrm{prof}}(V_k^{\mathrm{loc}})]_+$. The total negative pure-V curvature contributed by the main risk and the remaining profiled constraint terms is bounded above by $\eta_{V,k} \ge 0$. Define $\kappa_{V,k} = a_{0,k}\gamma_{0,k} - \eta_{V,k} > 0$. If the zeroth constraint is locally inactive, the same conclusion may be used with $\kappa_{V,k}$ replaced by any directly verified lower bound for the complete pure-V Hessian block.*
- *(c) The outcome-predictor block obeys $H_{hh,k} := \nabla_{\eta\eta}^2 F_{k,n}^{\mathrm{prof}}(\theta_k^{\mathrm{loc}}) \succeq \kappa_{h,k} I$, with $\kappa_{h,k} > 0$. The lower bound in condition (c) is supplied by Lemma C.3.*
- *(d) For tangent directions in the local Procrustes slice $\mathcal{C}_k$, the mixed block satisfies $\|H_{Vh,k}\|_{\mathrm{op}} \le c_k$ and $c_k^2 < \kappa_{V,k}\kappa_{h,k}$.*
- *(e) The full Riemannian Hessian has local operator-norm upper bound $\overline{\kappa}_k < \infty$ and is locally Lipschitz with constant $L_{H,k} < \infty$.*

Define

$$\underline{\kappa}_k = \frac{\kappa_{V,k} + \kappa_{h,k} - \sqrt{(\kappa_{V,k} - \kappa_{h,k})^2 + 4c_k^2}}{2} > 0. \tag{C.17}$$

Let $T_k = T_{\theta_k^{\mathrm{loc}}}\mathcal{S}_k$ and $C_k = T_{\theta_k^{\mathrm{loc}}}\mathcal{C}_k$. Since $T_{\theta_k^{\mathrm{loc}}}\mathcal{M}_k = T_k \oplus C_k$, finite-dimensional norm equivalence gives a constant $C_{\mathrm{sl},k} \geq 1$ such that, for every $\xi \in N_{\theta_k^{\mathrm{loc}}}\mathcal{S}_k$, its unique decomposition $\xi = t+c$, $t \in T_k$, $c \in C_k$, satisfies $\|c\| \geq C_{\mathrm{sl},k}^{-1}\|\xi\|$. Define the induced normal-space curvature by

$$\underline{\kappa}_k^{\mathrm{nor}} = \underline{\kappa}_k / C_{\mathrm{sl},k}^2. \tag{C.18}$$

Then

$$\ker\{\mathrm{Hess}\, F_{k,n}^{\mathrm{prof}}(\theta_k^{\mathrm{loc}})\} = T_{\theta_k^{\mathrm{loc}}}\mathcal{S}_k,$$

and, for every normal vector $\xi \in N_{\theta_k^{\mathrm{loc}}}\mathcal{S}_k$,

$$\langle \xi, \mathrm{Hess}\, F_{k,n}^{\mathrm{prof}}(\theta_k^{\mathrm{loc}})[\xi]\rangle \geq \underline{\kappa}_k^{\mathrm{nor}}\|\xi\|^2. \tag{C.19}$$

Consequently, $F_{k,n}^{\mathrm{prof}}$ is Morse–Bott around $\mathcal{S}_k$. On a possibly smaller tube satisfying $\mathrm{dist}(\theta, \mathcal{S}_k) \leq \underline{\kappa}_k^{\mathrm{nor}}/L_{H,k}$, it obeys the local PL inequality

$$F_{k,n}^{\mathrm{prof}}(\theta) - (F_{k,n}^{\mathrm{prof}})_{\mathrm{loc}}^{\star} \leq \frac{\|\operatorname{grad} F_{k,n}^{\mathrm{prof}}(\theta)\|^2}{2\mu_{\mathrm{PL},k}}, \qquad \mu_{\mathrm{PL},k} = \frac{(\underline{\kappa}_k^{\mathrm{nor}})^2}{4\overline{\kappa}_k}. \tag{C.20}$$

Proof. We prove the result in five steps.

Step 1: rotation directions are Hessian-null directions. By Assumption B(ii),

$$F_{k,n}^{\mathrm{prof}}(VR, \rho_R(\eta)) = F_{k,n}^{\mathrm{prof}}(V, \eta) \qquad (R \in \mathbb{O}(r)).$$

Let $R(t) = \exp(t\Omega)$ for a skew-symmetric matrix Ω and define

$$\theta(t) = \{V_k^{\mathrm{loc}}R(t), \rho_{R(t)}(\eta_k^{\mathrm{loc}})\}.$$

Then $\theta(t) \in \mathcal{S}_k$ and $F_{k,n}^{\mathrm{prof}}\{\theta(t)\}$ is constant. Hence its first and second derivatives at zero vanish. Because θ_k^{loc} is a local minimizer, its Riemannian Hessian is positive semidefinite. For a positive-semidefinite self-adjoint operator, a zero quadratic form implies that the vector lies in the kernel. Therefore

$$T_{\theta_k^{\mathrm{loc}}}\mathcal{S}_k \subseteq \ker\{\mathrm{Hess}\, F_{k,n}^{\mathrm{prof}}(\theta_k^{\mathrm{loc}})\}.$$

Step 2: the paired-difference eigengap supplies pure projection curvature. By Lemma C.4, the V-component of every slice tangent is horizontal after choosing the local representative of the rotation orbit. Write

$$\widehat{\Sigma}_{\Delta,k}V_{\Delta,k} = V_{\Delta,k}\Lambda_1, \qquad \widehat{\Sigma}_{\Delta,k}V_{\Delta,k,\perp} = V_{\Delta,k,\perp}\Lambda_2.$$

A horizontal Grassmann tangent vector at $V_{\Delta,k}$ has the form $\xi_V = V_{\Delta,k,\perp}K$. Consider the second-order Stiefel curve

$$V(t) = V_{\Delta,k} + tV_{\Delta,k,\perp}K - \frac{t^2}{2}V_{\Delta,k}K^\top K + O(t^3).$$

It satisfies $V(t)^\top V(t) = I_r + O(t^3)$. Since

$$\widehat{L}_{\mathrm{aug,k}}(V) = \mathrm{tr}(V^\top \widehat{\Sigma}_{\Delta,k}V),$$

direct expansion gives

$$\frac{d^2}{dt^2}\widehat{L}_{\mathrm{aug,k}}\{V(t)\}\Big|_{t=0} = 2\,\mathrm{tr}\{K^\top(\Lambda_2 - \Lambda_1)K\} \geq 2\widehat{\gamma}_{\Delta,k}\|K\|_F^2.$$

By local Hessian Lipschitz continuity and $d_{\mathcal{G}}(V_k^{\mathrm{loc}}, V_{\Delta,k}) \leq r_{V,k}$,

$$\langle \xi_V, \mathrm{Hess}\, \widehat{g}_{0,k}^{\mathrm{prof}}(V_k^{\mathrm{loc}})[\xi_V]\rangle \geq \gamma_{0,k}\|\xi_V\|^2.$$

For the active zeroth ALM term, the chain rule gives

$$\begin{aligned}
&\langle \xi_V, \mathrm{Hess}(\phi_{\rho_{\mathrm{ALM},n},\lambda_{0,k}^{\mathrm{fin}}} \circ \widehat{g}_{0,k}^{\mathrm{prof}})(V_k^{\mathrm{loc}})[\xi_V]\rangle \\
&\quad = \rho_{\mathrm{ALM},n}\langle \operatorname{grad} \widehat{g}_{0,k}^{\mathrm{prof}}, \xi_V\rangle^2 + a_{0,k}\langle \xi_V, \mathrm{Hess}\, \widehat{g}_{0,k}^{\mathrm{prof}}[\xi_V]\rangle \\
&\quad \geq a_{0,k}\gamma_{0,k}\|\xi_V\|^2.
\end{aligned}$$

Subtracting the allowed negative curvature $\eta_{V,k}$ from all other pure-V terms yields

$$\langle \xi_V, H_{VV,k}\xi_V \rangle \geq \kappa_{V,k}\|\xi_V\|^2. \tag{C.21}$$

The profiled auxiliary Hessian is already included here through the Schur-complement identity in Lemma C.1, $L_{VV} - L_{Vd}L_{dd}^{-1}L_{dV}$.

Step 3: the full block Hessian is positive on the local slice. Write a slice tangent vector as $c = (c_V, c_h) \in C_k$. By (C.21), condition (c), and the mixed-block bound,

$$\langle c, H_k c\rangle \geq \kappa_{V,k}\|c_V\|^2 + \kappa_{h,k}\|c_h\|^2 - 2c_k\|c_V\|\|c_h\|.$$

The right-hand side is the quadratic form of

$$M_k = \begin{pmatrix} \kappa_{V,k} & -c_k \\ -c_k & \kappa_{h,k} \end{pmatrix}$$

in the two nonnegative coordinates $(\|c_V\|, \|c_h\|)$. The smallest eigenvalue of M_k is exactly $\underline{\kappa}_k$ in (C.17). The condition $c_k^2 < \kappa_{V,k}\kappa_{h,k}$ makes it strictly positive. Thus

$$\langle c, H_k c\rangle \geq \underline{\kappa}_k\|c\|^2 \qquad (c \in C_k).$$

Step 4: slice curvature transfers to the orthogonal normal space. Let $\xi \in N_{\theta_k^{\mathrm{loc}}}\mathcal{S}_k$. Since $T_{\theta_k^{\mathrm{loc}}}\mathcal{M}_k = T_k \oplus C_k$, write uniquely $\xi = t + c$, with $t \in T_k$ and $c \in C_k$. By the definition of $C_{\mathrm{sl},k}$, $\|c\| \geq C_{\mathrm{sl},k}^{-1}\|\xi\|$. Step 1 gives $H_k t = 0$, and the Riemannian Hessian is self-adjoint. Therefore

$$\begin{aligned} \langle \xi, H_k\xi\rangle &= \langle t + c, H_k(t + c)\rangle \\ &= \langle c, H_k c\rangle \\ &\geq \underline{\kappa}_k\|c\|^2 \\ &\geq \underline{\kappa}_k^{\mathrm{nor}}\|\xi\|^2. \end{aligned}$$

This proves (C.19). Step 1 also gives $T_{\theta_k^{\mathrm{loc}}}\mathcal{S}_k \subseteq \ker H_k$, and the last display shows that no nonzero orthogonal normal vector belongs to the kernel. Since tangent and normal spaces form an orthogonal decomposition of $T_{\theta_k^{\mathrm{loc}}}\mathcal{M}_k$, we obtain

$$\ker H_k = T_{\theta_k^{\mathrm{loc}}}\mathcal{S}_k.$$

Together with the local description of the minimizer set in condition (a), this is the Morse–Bott property.

For $R \in \mathbb{O}(r)$, define the group action

$$G_R(V, \eta) = (VR, \rho_R(\eta)).$$

By Assumption B(ii), G_R is an isometry of the product manifold and

$$F_{k,n}^{\mathrm{prof}} \circ G_R = F_{k,n}^{\mathrm{prof}}.$$

The Riemannian gradient and Hessian are therefore equivariant:

$$\operatorname{grad} F_{k,n}^{\mathrm{prof}}(G_R\theta) = DG_R(\theta)\operatorname{grad} F_{k,n}^{\mathrm{prof}}(\theta),$$

and

$$\operatorname{Hess} F_{k,n}^{\mathrm{prof}}(G_R\theta)\{DG_R(\theta)\xi\} = DG_R(\theta)\operatorname{Hess} F_{k,n}^{\mathrm{prof}}(\theta)[\xi].$$

Because $DG_R(\theta)$ is an isometry, the Hessian kernel, the normal-space eigenvalues, and the curvature lower bound $\underline{\kappa}_k^{\mathrm{nor}}$ are constant along the orbit $\mathcal{S}_k$. Hence (C.19) holds at every $\bar{\theta} \in \mathcal{S}_k$.

Step 5: Morse–Bott curvature implies a local PL inequality. Let θ lie in a sufficiently small tubular neighborhood and let $\bar{\theta} = \mathrm{Proj}_{\mathcal{S}_k}(\theta)$ be its unique nearest point on $\mathcal{S}_k$. Write $\theta = \mathrm{Exp}_{\bar{\theta}}(\xi)$ with $\xi \in N_{\bar{\theta}}\mathcal{S}_k$. Along the minimizing geodesic, the Hessian upper bound yields

$$F_{k,n}^{\mathrm{prof}}(\theta) - (F_{k,n}^{\mathrm{prof}})_{\mathrm{loc}}^{\star} \leq \frac{\overline{\kappa}_k}{2}\|\xi\|^2.$$

Parallel-transporting the gradient back to $\bar{\theta}$ and using Hessian Lipschitz continuity gives

$$\left\| \mathrm{PT}_{\theta \to \bar{\theta}} \operatorname{grad} F_{k,n}^{\mathrm{prof}}(\theta) - \operatorname{Hess} F_{k,n}^{\mathrm{prof}}(\bar{\theta})[\xi] \right\| \le \frac{L_{H,k}}{2} \|\xi\|^2.$$

If $\|\xi\| \le \underline{\kappa}_k^{\mathrm{nor}} / L_{H,k}$, then

$$\|\operatorname{grad} F_{k,n}^{\mathrm{prof}}(\theta)\| \ge \underline{\kappa}_k^{\mathrm{nor}} \|\xi\| - \frac{L_{H,k}}{2} \|\xi\|^2 \ge \frac{\underline{\kappa}_k^{\mathrm{nor}}}{2} \|\xi\|.$$

Hence

$$\|\xi\|^2 \le \frac{4}{(\underline{\kappa}_k^{\mathrm{nor}})^2} \|\operatorname{grad} F_{k,n}^{\mathrm{prof}}(\theta)\|^2.$$

Substitution into the objective upper bound gives

$$F_{k,n}^{\mathrm{prof}}(\theta) - (F_{k,n}^{\mathrm{prof}})_{\mathrm{loc}}^{\star} \le \frac{2\overline{\kappa}_k}{(\underline{\kappa}_k^{\mathrm{nor}})^2} \|\operatorname{grad} F_{k,n}^{\mathrm{prof}}(\theta)\|^2 = \frac{1}{2\mu_{\mathrm{PL},k}} \|\operatorname{grad} F_{k,n}^{\mathrm{prof}}(\theta)\|^2,$$

which is (C.20). The equivalence between local Morse–Bott geometry and local PL for C^2 manifold objectives is also consistent with Rebjock & Boumal (2025). □

For the fixed number of folds define the uniform local PL constant

$$\mu_{\mathrm{PL},n} := \min_{1 \le k \le K} \mu_{\mathrm{PL},k} = \min_{1 \le k \le K} \frac{(\underline{\kappa}_k^{\mathrm{nor}})^2}{4\overline{\kappa}_k}. \tag{C.22}$$

Under the uniform curvature and Hessian-upper-bound conditions of Proposition C.1, $\mu_{\mathrm{PL},n} > 0$. All later appearances of $\mu_{\mathrm{PL},n}$ refer to (C.22).

Lemma C.5 (Outer–inner accepted-path retention in a common local tube)**.** *Fix a fold k. Let $t = 0, \dots, T_k$ index the outer augmented-Lagrangian iterations, and let $s = 0, \dots, S_{k,t}$ index the accepted inner iterates within outer iteration t. Write*

$$F_{k,t}^{\mathrm{prof}}(\theta) = \widehat{A}_{\rho_{k,t},k}^{\mathrm{prof}}(\theta, \lambda_{k,t}), \qquad \widetilde{F}_{k,t}(\theta) = \widetilde{A}_{\rho_{k,t},k}(\theta, \lambda_{k,t}).$$

Assume that the local block-Hessian conditions of Proposition C.1 hold uniformly for $F_{k,t}^{\mathrm{prof}}$ on one common tubular neighborhood $\mathcal{U}_{k,n}$ for every $t \le T_k$. Let

$$F_{k,t}^{\min} = \inf_{\theta \in \mathcal{U}_{k,n}} F_{k,t}^{\mathrm{prof}}(\theta),$$

and define the outer-iteration boundary gaps

$$\Delta_{k,t}^{\mathrm{basin}} := \inf_{\theta \in \partial \mathcal{U}_{k,n}} \{F_{k,t}^{\mathrm{prof}}(\theta) - F_{k,t}^{\min}\}, \qquad \Delta_k^{\mathrm{basin}} := \min_{0 \le t \le T_k} \Delta_{k,t}^{\mathrm{basin}} > 0. \tag{C.23}$$

For an accepted inner update, write

$$\gamma_{k,t,s}(\tau) = R_{\theta_{k,t,s}}(\tau \alpha_{k,t,s} \zeta_{k,t,s}), \qquad 0 \le \tau \le 1, \qquad \theta_{k,t,s+1} = \gamma_{k,t,s}(1),$$

and assume the accepted retraction segment obeys

$$\sup_{0 \le \tau \le 1} \widetilde{F}_{k,t}\{\gamma_{k,t,s}(\tau)\} \le \widetilde{F}_{k,t}(\theta_{k,t,s}) + \varepsilon_{\mathrm{path},k,t,s}. \tag{C.24}$$

Along the same accepted segment, suppose the implementation-to-profile error obeys

$$\sup_{0 \le \tau \le 1} |\widetilde{F}_{k,t}\{\gamma_{k,t,s}(\tau)\} - F_{k,t}^{\mathrm{prof}}\{\gamma_{k,t,s}(\tau)\}| \le \eta_{A,k,t,s}^{\mathrm{impl}}. \tag{C.25}$$

This pathwise condition is required on the entire accepted retraction segment, not only at its endpoints; it is the segmentwise form of Assumption B(iii). The next outer iteration is warm-started at the preceding returned parameter,

$$\theta_{k,t+1,0} = \theta_{k,t,S_{k,t}}.$$

To quantify the change induced by the multiplier and penalty update, define

$$d_{k,t}^{\mathrm{out}} := \sup_{\theta \in \mathcal{U}_{k,n}} |F_{k,t+1}^{\mathrm{prof}}(\theta) - F_{k,t}^{\mathrm{prof}}(\theta)|, \qquad 0 \le t < T_k. \tag{C.26}$$

This drift is directly controlled by the recorded multiplier and penalty updates. Specifically, suppose

$$|\widehat{g}_{q,k}^{\mathrm{prof}}(V)| \le B_g, \qquad 0 \le \lambda_{q,k,t} \le \Lambda, \qquad 0 < \rho_{\min} \le \rho_{k,t} \le \rho_{\max}.$$

Define

$$L_\lambda = \max\{B_g, \Lambda/\rho_{\min}\}, \qquad L_\rho = \frac{1}{2}\max\{B_g^2, \Lambda^2/\rho_{\min}^2\}.$$

Then

$$d_{k,t}^{\mathrm{out}} \le L_\lambda \|\lambda_{k,t+1} - \lambda_{k,t}\|_1 + 4L_\rho |\rho_{k,t+1} - \rho_{k,t}|. \tag{C.27}$$

Indeed, for a scalar constraint value g, the positive-part ALM penalty satisfies

$$\phi_{\rho,\lambda}(g) = \begin{cases} \lambda g + \rho g^2/2, & \lambda + \rho g > 0, \\ -\lambda^2/(2\rho), & \lambda + \rho g \le 0. \end{cases}$$

Its partial derivatives agree at the boundary and obey

$$|\partial_\lambda \phi_{\rho,\lambda}(g)| \le L_\lambda, \qquad |\partial_\rho \phi_{\rho,\lambda}(g)| \le L_\rho.$$

The mean-value inequality, followed by summation over the four constraints, proves (C.27)*. If $\theta_{k,0,0} \in \mathcal{U}_{k,n}$ and*

$$\begin{aligned} &F_{k,0}^{\mathrm{prof}}(\theta_{k,0,0}) - F_{k,0}^{\min} \\ &\quad + \sum_{t=0}^{T_k} \sum_{s=0}^{S_{k,t}-1} \{\varepsilon_{\mathrm{path},k,t,s} + 2\eta_{A,k,t,s}^{\mathrm{impl}}\} + 2\sum_{t=0}^{T_k-1} d_{k,t}^{\mathrm{out}} < \Delta_k^{\mathrm{basin}}, \end{aligned} \tag{C.28}$$

then every accepted retraction segment in every outer iteration lies in $\mathcal{U}_{k,n}$. In particular, the final output $\widehat{\theta}_k = \theta_{k,T_k,S_{k,T_k}}$ lies in the local PL tube.

The initial gap in (C.28) *is linked to Theorem 1 as follows. Let $\theta_{k,0}^{\mathrm{loc}} = (V_{k,0}^{\mathrm{loc}}, \eta_{k,0}^{\mathrm{loc}})$ be a point in the initial local minimizer orbit of $F_{k,0}^{\mathrm{prof}}$. Suppose the selected predictor branch $\eta_{k,0}^{\circ}(V)$ is locally Lipschitz,*

$$\|\eta_{k,0}^{\circ}(V_1) - \eta_{k,0}^{\circ}(V_2)\| \le L_{\eta,k} d_{\mathcal{G}}(V_1, V_2),$$

the initialization uses $\theta_{k,0,0} = (V_k^{(0)}, \eta_{k,0}^{\circ}(V_k^{(0)}))$, and

$$d_{\mathcal{G}}(V_{k,0}^{\mathrm{loc}}, V^\dagger) \le r_{\mathrm{loc},k,n}.$$

If the local product-manifold Hessian is bounded above by $\overline{\kappa}_{k,0}$ and, locally, $d_{\mathcal{G}}(V_1, V_2) \le c_\Pi \|V_1 V_1^\top - V_2 V_2^\top\|_F$, then

$$\begin{aligned} &F_{k,0}^{\mathrm{prof}}(\theta_{k,0,0}) - F_{k,0}^{\min} \\ &\quad \le \frac{\overline{\kappa}_{k,0}}{2}(1 + L_{\eta,k}^2)\left\{c_\Pi \|\Pi_k^{(0)} - \Pi^\dagger\|_F + r_{\mathrm{loc},k,n}\right\}^2. \end{aligned} \tag{C.29}$$

Thus Theorem 1, together with a local-minimizer displacement $r_{\mathrm{loc},k,n}$, provides a primitive route to the initial term in (C.28)*. If another predictor initialization is used, the initial objective gap can instead be evaluated directly from the recorded training objective.*

Proof. We first record the effect of an outer update on normalized suboptimality. By (C.26),

$$F_{k,t+1}^{\mathrm{prof}}(\theta) \le F_{k,t}^{\mathrm{prof}}(\theta) + d_{k,t}^{\mathrm{out}} \quad (\theta \in \mathcal{U}_{k,n}).$$

Taking infima over $\theta \in \mathcal{U}_{k,n}$ in both directions gives

$$|F_{k,t+1}^{\min} - F_{k,t}^{\min}| \le d_{k,t}^{\mathrm{out}}.$$

Consequently, for every $\theta \in \mathcal{U}_{k,n}$,

$$F_{k,t+1}^{\mathrm{prof}}(\theta) - F_{k,t+1}^{\min} \le F_{k,t}^{\mathrm{prof}}(\theta) - F_{k,t}^{\min} + 2d_{k,t}^{\mathrm{out}}. \tag{C.30}$$

We now prove retention by induction over the lexicographic order of (t, s). Suppose all preceding accepted segments lie in $\mathcal{U}_{k,n}$. If the current segment first left the tube, continuity would give a smallest $\tau_\star \in (0, 1]$ such that $\gamma_{k,t,s}(\tau_\star) \in \partial\mathcal{U}_{k,n}$. By (C.25) and (C.24),

$$\begin{aligned} F_{k,t}^{\mathrm{prof}}\{\gamma_{k,t,s}(\tau_\star)\} &\le \widetilde{F}_{k,t}\{\gamma_{k,t,s}(\tau_\star)\} + \eta_{A,k,t,s}^{\mathrm{impl}} \\ &\le \widetilde{F}_{k,t}(\theta_{k,t,s}) + \varepsilon_{\mathrm{path},k,t,s} + \eta_{A,k,t,s}^{\mathrm{impl}} \\ &\le F_{k,t}^{\mathrm{prof}}(\theta_{k,t,s}) + \varepsilon_{\mathrm{path},k,t,s} + 2\eta_{A,k,t,s}^{\mathrm{impl}}. \end{aligned}$$

Repeated use of this inequality within an outer iteration and of (C.30) between outer iterations yields

$$\begin{aligned} &F_{k,t}^{\mathrm{prof}}\{\gamma_{k,t,s}(\tau_\star)\} - F_{k,t}^{\mathrm{min}} \\ &\le F_{k,0}^{\mathrm{prof}}(\theta_{k,0,0}) - F_{k,0}^{\mathrm{min}} + \sum_{\ell=0}^{t-1}\sum_{j=0}^{S_{k,\ell}-1}\{\varepsilon_{\mathrm{path},k,\ell,j} + 2\eta_{A,k,\ell,j}^{\mathrm{impl}}\} \\ &\quad + \sum_{j=0}^{s}\{\varepsilon_{\mathrm{path},k,t,j} + 2\eta_{A,k,t,j}^{\mathrm{impl}}\} + 2\sum_{\ell=0}^{t-1} d_{k,\ell}^{\mathrm{out}} \\ &< \Delta_k^{\mathrm{basin}} \le \Delta_{k,t}^{\mathrm{basin}}, \end{aligned}$$

where the strict inequality is (C.28). This contradicts the definition of $\Delta_{k,t}^{\mathrm{basin}}$ because the alleged first exit point lies on $\partial\mathcal{U}_{k,n}$. Hence the current path remains in the tube. Evaluating the same inequalities at $\tau = 1$ establishes the induction step for the next inner iterate. At the end of outer iteration t, inequality (C.30) controls the same warm-started parameter under $F_{k,t+1}^{\mathrm{prof}}$. Thus induction continues across the outer update. All accepted paths, and therefore the final output, remain in $\mathcal{U}_{k,n}$.

For (C.29), choose $\theta_{k,0}^{\mathrm{loc}}$ in the initial minimizer orbit. The triangle inequality and the local projector-to-Grassmann metric equivalence give

$$\begin{aligned} d_{\mathcal{G}}(V_k^{(0)}, V_{k,0}^{\mathrm{loc}}) &\le d_{\mathcal{G}}(V_k^{(0)}, V^\dagger) + d_{\mathcal{G}}(V^\dagger, V_{k,0}^{\mathrm{loc}}) \\ &\le c_\Pi \|\Pi_k^{(0)} - \Pi^\dagger\|_F + r_{\mathrm{loc},k,n}. \end{aligned}$$

The Lipschitz predictor branch then implies

$$d_{\mathcal{M}}\{\theta_{k,0,0}, \theta_{k,0}^{\mathrm{loc}}\}^2 \le (1 + L_{\eta,k}^2) d_{\mathcal{G}}(V_k^{(0)}, V_{k,0}^{\mathrm{loc}})^2.$$

Since the gradient vanishes on the local minimizer orbit, a second-order Taylor upper bound with Hessian norm at most $\overline{\kappa}_{k,0}$ gives

$$F_{k,0}^{\mathrm{prof}}(\theta_{k,0,0}) - F_{k,0}^{\mathrm{min}} \le \frac{\overline{\kappa}_{k,0}}{2} d_{\mathcal{M}}\{\theta_{k,0,0}, \theta_{k,0}^{\mathrm{loc}}\}^2.$$

Combining the last three displays proves (C.29). □

C.5 ALM Risk and Feasibility Interface

Proposition C.2 (Complete ALM risk and feasibility interface)**.** *Suppose the outer–inner accepted-path conditions of Lemma C.5 hold; hence every accepted path and the returned point $\widehat{\theta}_k$ lie in the common local tube of Proposition C.1. Impose the same-basin comparison condition*

$$\theta_k^{\mathrm{cmp}} \in \mathcal{U}_{k,n}, \qquad (F_{k,n}^{\mathrm{prof}})_{\mathrm{loc}}^\star \le F_{k,n}^{\mathrm{prof}}(\theta_k^{\mathrm{cmp}}),$$

where the second inequality is automatic under the local-infimum definition in Proposition C.1. Suppose also that Assumption B(iii) and Lemma C.1 provide the implementation-to-profile value and gradient envelopes, and

$$0 \le \mathcal{R}_{y,k}^{\mathrm{tr}}(\theta) \le M_y, \qquad \max_q\{|g_{q,k}(V)|, |\widehat{g}_{q,k}^{\mathrm{prof}}(V)|, |\widetilde{g}_{q,k}(V)|\} \le B_g$$

on that region. Let $0 \le \lambda_{q,k}^{\mathrm{fin}} \le \Lambda_n$, and suppose

$$\|\operatorname{grad} \widetilde{A}_{\rho_{\mathrm{ALM},n},k}(\widehat{\theta}_k, \lambda_k^{\mathrm{fin}})\| \le \epsilon_{\mathrm{in},n}.$$

When the multiplier-stopping bound is used, let

$$\lambda_{q,k}^{\mathrm{fin}} = [\lambda_{q,k}^{\mathrm{prev}} + \rho_{\mathrm{ALM},n}\widetilde{g}_{q,k}(\widehat{V}_k)]_+, \qquad \|\lambda_k^{\mathrm{fin}} - \lambda_k^{\mathrm{prev}}\|_\infty \le \delta_{\lambda,n}.$$

On the simultaneous event of Lemma C.2, set

$$\xi_{g,0,n}^{\mathrm{prof}} = \xi_{g,0,n}^{\mathrm{total}} := \xi_{0,n}, \qquad \xi_{\mathrm{ALM},n} = \xi_{y,n} + \sum_{q=0}^{3}(\Lambda_n + \rho_{\mathrm{ALM},n}B_g)\xi_{g,q,n}^{\mathrm{prof}}.$$

Let

$$\epsilon_{\mathrm{PL},n} = \frac{(\epsilon_{\mathrm{in},n} + \eta_{\nabla,n}^{\mathrm{impl}})^2}{2\mu_{\mathrm{PL},n}}, \qquad r_{\mathrm{ALM},n} = 2\xi_{\mathrm{ALM},n} + \epsilon_{\mathrm{PL},n}.$$

Then, for each fold k,

$$\mathcal{R}_{y,k}^{\mathrm{tr}}(\widehat{\theta}_k) - \mathcal{R}_{y,k}^{\mathrm{tr}}(\theta_k^\dagger) \le \delta_{\mathrm{risk},k,n}^{\mathrm{ALM}}, \tag{C.31}$$

where

$$\delta_{\mathrm{risk},k,n}^{\mathrm{ALM}} = r_{\mathrm{ALM},n} + S_{\mathrm{ALM},n}^{\mathrm{cmp}} + \frac{\|\lambda_k^{\mathrm{fin}}\|_2^2}{2\rho_{\mathrm{ALM},n}} + \alpha_{\mathrm{app},n}, \qquad \delta_{\mathrm{risk},n}^{\mathrm{ALM}} := \max_{1\le k\le K}\delta_{\mathrm{risk},k,n}^{\mathrm{ALM}}.$$

For every $q = 0, 1, 2, 3$,

$$[g_{q,k}(\widehat{V}_k)]_+ \le \left[\frac{2}{\rho_{\mathrm{ALM},n}}\left\{M_y + r_{\mathrm{ALM},n} + S_{\mathrm{ALM},n}^{\mathrm{cmp}} + \frac{\|\lambda_k^{\mathrm{fin}}\|_2^2}{2\rho_{\mathrm{ALM},n}}\right\}\right]^{1/2}. \tag{C.32}$$

If the multiplier-stopping condition is used,

$$[g_{q,k}(\widehat{V}_k)]_+ \le \frac{\delta_{\lambda,n}}{\rho_{\mathrm{ALM},n}} + \xi_{g,q,n}^{\mathrm{total}}. \tag{C.33}$$

Thus $\delta_{\mathrm{feas},n}$ may be taken as the maximum over folds and constraints of the smaller bound in (C.32) *and* (C.33).

Proof. For $\lambda \ge 0$ and $\rho > 0$,

$$\phi_{\rho,\lambda}(t) = \begin{cases} \lambda t + \rho t^2/2, & \lambda + \rho t > 0, \\ -\lambda^2/(2\rho), & \lambda + \rho t \le 0. \end{cases}$$

Hence $\phi_{\rho,\lambda}$ is nondecreasing and

$$\phi_{\rho,\lambda}(t) \ge -\frac{\lambda^2}{2\rho}, \qquad \frac{\rho}{2}[t]_+^2 \le \phi_{\rho,\lambda}(t) + \frac{\lambda^2}{2\rho}. \tag{C.34}$$

On $[-B_g, B_g]$,

$$|\phi_{\rho,\lambda}(a) - \phi_{\rho,\lambda}(b)| \le (\lambda + \rho B_g)|a - b|. \tag{C.35}$$

Lemma C.2 and the boundedness conditions in the proposition give, uniformly over the local region,

$$\begin{aligned} &|\widehat{A}_{\rho_{\mathrm{ALM},n},k}^{\mathrm{prof}}(\theta, \lambda_k^{\mathrm{fin}}) - A_{\rho_{\mathrm{ALM},n},k}(\theta, \lambda_k^{\mathrm{fin}})| \\ &\le |\widehat{\mathcal{R}}_{y,k}^{\mathrm{tr}}(\theta) - \mathcal{R}_{y,k}^{\mathrm{tr}}(\theta)| \\ &\quad + \sum_{q=0}^{3}|\phi_{\rho_{\mathrm{ALM},n},\lambda_{q,k}^{\mathrm{fin}}}\{\widehat{g}_{q,k}^{\mathrm{prof}}(V)\} - \phi_{\rho_{\mathrm{ALM},n},\lambda_{q,k}^{\mathrm{fin}}}\{g_{q,k}(V)\}| \\ &\le \xi_{\mathrm{ALM},n}. \end{aligned} \tag{C.36}$$

By the numerical stopping rule,

$$\|\operatorname{grad}\widetilde{A}_{\rho_{\mathrm{ALM},n},k}(\widehat{\theta}_k, \lambda_k^{\mathrm{fin}})\| \le \epsilon_{\mathrm{in},n},$$

and Lemma C.1 gives

$$\|\operatorname{grad}\widehat{A}_{\rho_{\mathrm{ALM},n},k}^{\mathrm{prof}}(\widehat{\theta}_k, \lambda_k^{\mathrm{fin}})\| \le \epsilon_{\mathrm{in},n} + \eta_{\nabla,n}^{\mathrm{impl}}.$$

Proposition C.1 and equation (C.20) therefore yield

$$\widehat{A}^{\mathrm{prof}}_{\rho_{\mathrm{ALM},n},k}(\widehat{\theta}_k,\lambda_k^{\mathrm{fin}}) - (F^{\mathrm{prof}}_{k,n})^\star_{\mathrm{loc}} \le \frac{\|\operatorname{grad}\widehat{A}^{\mathrm{prof}}_{\rho_{\mathrm{ALM},n},k}(\widehat{\theta}_k,\lambda_k^{\mathrm{fin}})\|^2}{2\mu_{\mathrm{PL},n}} \le \epsilon_{\mathrm{PL},n}. \tag{C.37}$$

Using (C.36)–(C.37), the same-basin inequality, and the explicit comparator θ_k^{cmp},

$$\begin{aligned} A_{\rho_{\mathrm{ALM},n},k}(\widehat{\theta}_k,\lambda_k^{\mathrm{fin}}) &\le \widehat{A}^{\mathrm{prof}}_{\rho_{\mathrm{ALM},n},k}(\widehat{\theta}_k,\lambda_k^{\mathrm{fin}}) + \xi_{\mathrm{ALM},n} \\ &\le \widehat{A}^{\mathrm{prof}}_{\rho_{\mathrm{ALM},n},k}(\theta_k^{\mathrm{cmp}},\lambda_k^{\mathrm{fin}}) + \epsilon_{\mathrm{PL},n} + \xi_{\mathrm{ALM},n} \\ &\le A_{\rho_{\mathrm{ALM},n},k}(\theta_k^{\mathrm{cmp}},\lambda_k^{\mathrm{fin}}) + 2\xi_{\mathrm{ALM},n} + \epsilon_{\mathrm{PL},n} \\ &= A_{\rho_{\mathrm{ALM},n},k}(\theta_k^{\mathrm{cmp}},\lambda_k^{\mathrm{fin}}) + r_{\mathrm{ALM},n}. \end{aligned} \tag{C.38}$$

Expanding,

$$\mathcal{R}^{\mathrm{tr}}_{y,k}(\widehat{\theta}_k) - \mathcal{R}^{\mathrm{tr}}_{y,k}(\theta_k^{\mathrm{cmp}}) \le r_{\mathrm{ALM},n} + \sum_{q=0}^{3}\phi_{\rho_{\mathrm{ALM},n},\lambda^{\mathrm{fin}}_{q,k}}\{g_{q,k}(V^\dagger)\} - \sum_{q=0}^{3}\phi_{\rho_{\mathrm{ALM},n},\lambda^{\mathrm{fin}}_{q,k}}\{g_{q,k}(\widehat{V}_k)\}.$$

Since $g_{q,k}(V^\dagger) \le s^{\mathrm{cmp}}_{q,n}$ and the penalty is nondecreasing,

$$\sum_q \phi_{\rho_{\mathrm{ALM},n},\lambda^{\mathrm{fin}}_{q,k}}\{g_{q,k}(V^\dagger)\} \le S^{\mathrm{cmp}}_{\mathrm{ALM},n}.$$

The first inequality in (C.34) gives

$$-\sum_q \phi_{\rho_{\mathrm{ALM},n},\lambda^{\mathrm{fin}}_{q,k}}\{g_{q,k}(\widehat{V}_k)\} \le \frac{\|\lambda_k^{\mathrm{fin}}\|_2^2}{2\rho_{\mathrm{ALM},n}}.$$

Therefore

$$\mathcal{R}^{\mathrm{tr}}_{y,k}(\widehat{\theta}_k) - \mathcal{R}^{\mathrm{tr}}_{y,k}(\theta_k^{\mathrm{cmp}}) \le r_{\mathrm{ALM},n} + S^{\mathrm{cmp}}_{\mathrm{ALM},n} + \frac{\|\lambda_k^{\mathrm{fin}}\|_2^2}{2\rho_{\mathrm{ALM},n}}.$$

Adding the comparison bound

$$\mathcal{R}^{\mathrm{tr}}_{y,k}(\theta_k^{\mathrm{cmp}}) - \mathcal{R}^{\mathrm{tr}}_{y,k}(\theta_k^{\dagger}) \le \alpha_{\mathrm{app},n}$$

proves (C.31).

For feasibility, (C.38), $\mathcal{R}^{\mathrm{tr}}_{y,k}(\widehat{\theta}_k) \ge 0$, and $\mathcal{R}^{\mathrm{tr}}_{y,k}(\theta_k^{\mathrm{cmp}}) \le M_y$ imply

$$\sum_{p=0}^{3}\phi_{\rho_{\mathrm{ALM},n},\lambda^{\mathrm{fin}}_{p,k}}\{g_{p,k}(\widehat{V}_k)\} \le M_y + S^{\mathrm{cmp}}_{\mathrm{ALM},n} + r_{\mathrm{ALM},n}.$$

Fix q, move all other penalty terms to the right, and use $\phi_{\rho,\lambda}(t) \ge -\lambda^2/(2\rho)$:

$$\phi_{\rho_{\mathrm{ALM},n},\lambda^{\mathrm{fin}}_{q,k}}\{g_{q,k}(\widehat{V}_k)\} \le M_y + S^{\mathrm{cmp}}_{\mathrm{ALM},n} + r_{\mathrm{ALM},n} + \sum_{p\ne q}\frac{(\lambda^{\mathrm{fin}}_{p,k})^2}{2\rho_{\mathrm{ALM},n}}.$$

The second inequality in (C.34) then gives

$$\frac{\rho_{\mathrm{ALM},n}}{2}[g_{q,k}(\widehat{V}_k)]_+^2 \le M_y + S^{\mathrm{cmp}}_{\mathrm{ALM},n} + r_{\mathrm{ALM},n} + \frac{\|\lambda_k^{\mathrm{fin}}\|_2^2}{2\rho_{\mathrm{ALM},n}},$$

which proves (C.32).

For the multiplier bound, if $\widetilde{g}_{q,k}(\widehat{V}_k) \le 0$, its positive part is zero. If it is positive, the update is untruncated and

$$\widetilde{g}_{q,k}(\widehat{V}_k) = \frac{\lambda^{\mathrm{fin}}_{q,k} - \lambda^{\mathrm{prev}}_{q,k}}{\rho_{\mathrm{ALM},n}} \le \frac{\delta_{\lambda,n}}{\rho_{\mathrm{ALM},n}}.$$

The positive-part map is one-Lipschitz. At the returned point, Lemma C.2 gives

$$|\widetilde{g}_{q,k}(\widehat{V}_k) - g_{q,k}(\widehat{V}_k)| \le \xi^{\mathrm{total}}_{g,q,n}.$$

Hence (C.33) follows. □

C.6 Projection Control and Semantic Comparison

Proposition C.3 (Projection control and semantic comparison)**.** *Let*

$$\ell_0^\dagger = L_{\mathrm{aug}}(\Pi^\dagger) = \sum_{j=1}^{r} \lambda_j(\Sigma_\Delta), \qquad c_{0,n} = \ell_0^\dagger + \bar{c}_{0,n}.$$

For every rank-r projector Π,

$$L_{\mathrm{aug}}(\Pi) - \ell_0^\dagger \geq \frac{\gamma_\dagger}{2}\|\Pi - \Pi^\dagger\|_F^2. \tag{C.39}$$

Consequently, if $[g_{0,k}(\widehat{V}_k)]_+ \leq \delta_{\mathrm{feas},n}$, then $\bar{c}_{0,n} + \delta_{\mathrm{feas},n} \geq 0$ automatically and

$$\|\widehat{\Pi}_k - \Pi^\dagger\|_F^2 \leq \frac{2}{\gamma_\dagger}(\bar{c}_{0,n} + \delta_{\mathrm{feas},n}). \tag{C.40}$$

If the optional semantic conditions in Lemma B.1 hold, then

$$\|\widehat{\Pi}_k - \Pi^\star\|_F \leq \left\{\frac{2(\bar{c}_{0,n} + \delta_{\mathrm{feas},n})}{\gamma_\dagger}\right\}^{1/2} + \frac{2\sqrt{2r}\tau_e}{\nu_S}.$$

Proof. Let $v_1, \ldots, v_d$ be an eigenbasis of Σ_Δ with ascending eigenvalues $\lambda_1 \leq \cdots \leq \lambda_d$, and write $p_j = v_j^\top \Pi v_j$. Because Π is a rank-r orthogonal projector, $0 \leq p_j \leq 1$ and $\sum_{j=1}^d p_j = r$. Since $\Pi^\dagger$ projects onto $v_1, \ldots, v_r$,

$$L_{\mathrm{aug}}(\Pi) - \ell_0^\dagger = \sum_{j=1}^{r} \lambda_j(p_j - 1) + \sum_{j=r+1}^{d} \lambda_j p_j.$$

Let

$$s = \sum_{j=r+1}^{d} p_j = \sum_{j=1}^{r}(1 - p_j),$$

where the equality follows from $\sum_j p_j = r$. Using $\lambda_j \leq \lambda_r$ for $j \leq r$ and $\lambda_j \geq \lambda_{r+1}$ for $j > r$,

$$\begin{aligned} L_{\mathrm{aug}}(\Pi) - \ell_0^\dagger &\geq -\lambda_r \sum_{j=1}^{r}(1 - p_j) + \lambda_{r+1} \sum_{j=r+1}^{d} p_j \\ &= (\lambda_{r+1} - \lambda_r)s = \gamma_\dagger s. \end{aligned}$$

For two rank-r projectors,

$$\begin{aligned} \|\Pi - \Pi^\dagger\|_F^2 &= \mathrm{tr}(\Pi) + \mathrm{tr}(\Pi^\dagger) - 2\,\mathrm{tr}(\Pi\Pi^\dagger) \\ &= 2\{r - \mathrm{tr}(\Pi\Pi^\dagger)\} = 2\,\mathrm{tr}\{\Pi(I - \Pi^\dagger)\} = 2s. \end{aligned}$$

This proves (C.39).

Next, $g_{0,k}(\widehat{V}_k) = L_{\mathrm{aug}}(\widehat{\Pi}_k) - c_{0,n}$ and $[g_{0,k}(\widehat{V}_k)]_+ \leq \delta_{\mathrm{feas},n}$ imply

$$\begin{aligned} L_{\mathrm{aug}}(\widehat{\Pi}_k) - \ell_0^\dagger &\leq c_{0,n} + \delta_{\mathrm{feas},n} - \ell_0^\dagger \\ &= \bar{c}_{0,n} + \delta_{\mathrm{feas},n}. \end{aligned}$$

Because $L_{\mathrm{aug}}(\widehat{\Pi}_k) \geq \ell_0^\dagger$, feasibility also implies $\bar{c}_{0,n} + \delta_{\mathrm{feas},n} \geq 0$; no sign condition on $\bar{c}_{0,n}$ is required separately. Combining the upper and lower bounds proves (C.40). The semantic comparison follows from the triangle inequality and Lemma B.1. □

C.7 Proof of Theorem 2

Proof of Theorem 2. Proposition C.2 supplies the population g_0-feasibility envelope. Proposition C.3 then gives the first projection bound fold by fold; taking the maximum over the fixed number of folds yields the displayed form. Under Lemma B.1, the triangle inequality gives the second bound.

For $q = 1, 2$, $g_{q,k}(V) = \mathcal{J}_{q,k}(V) - c_{q,n}$, so

$$\mathcal{J}_{q,k}(\widehat{V}_k) - c_{q,n} \le [g_{q,k}(\widehat{V}_k)]_+ \le \delta_{\mathrm{feas},n}.$$

For $q = 3$, $g_{3,k}(V) = c_{3,n} - \mathcal{J}_{3,k}(V)$, hence

$$c_{3,n} - \mathcal{J}_{3,k}(\widehat{V}_k) \le [g_{3,k}(\widehat{V}_k)]_+ \le \delta_{\mathrm{feas},n}.$$

These inequalities are auxiliary consequences of simultaneous feasibility, not additional claims of Theorem 2. No marginal-to-conditional Shannon-information bridge is used. □

Because each $\mathcal{J}_{q,k}(V)$ is invariant to rotations $V \mapsto VR$ with $R \in \mathbb{O}(r)$, it induces a function on $\mathcal{G}_r$; we write this quotient function as $\mathcal{J}_{q,k}(\Pi)$ and use its Grassmann gradient grad_Π. Define the conditional outcome increment

$$\mathcal{K}_{2,k}(\Pi) = \inf_{a_c \in \mathcal{A}_c} \mathbb{E}\ell_2\{Y, a_c(Z_c(\Pi))\} - \inf_{a_{cs} \in \mathcal{A}_{cs}} \mathbb{E}\ell_2\{Y, a_{cs}(Z_c(\Pi), Z_s(\Pi))\}.$$

Here $\mathcal{A}_c \subseteq \mathcal{A}_{cs}$, with the smaller class embedded by ignoring Z_s, so $\mathcal{K}_{2,k}(\Pi) \ge 0$.

Proposition C.4 (Marginal-to-conditional g_2 bridge under additive visibility)**.** *Fix fold k. For every $\Pi \in \mathcal{B}(\Pi^\dagger, r_0)$, write $C_\Pi = Z_c(\Pi)$, $S_\Pi = Z_s(\Pi)$, and $\mu_Y = \mathbb{E}Y$. Assume $\mathbb{E}(Y^2) < \infty$, squared loss, and the following deliberately restricted structure.*

1. *The no-representation class for $\mathcal{J}_{2,k}$ has population optimum μ_Y, and the represented class attains, or is $L_2(P)$-dense at, $\mathbb{E}(Y \mid S_\Pi)$. The classes $\mathcal{A}_c$ and $\mathcal{A}_{cs}$ likewise attain, or are dense at, $\mathbb{E}(Y \mid C_\Pi)$ and $\mathbb{E}(Y \mid C_\Pi, S_\Pi)$.*
2. *The pair (C_Π, S_Π) generates the same sigma-field as U, and there exist centered $f_\Pi(C_\Pi), g_\Pi(S_\Pi) \in L_2(P)$ such that*

$$\mathbb{E}(Y \mid U) = \mathbb{E}(Y \mid C_\Pi, S_\Pi) = \mu_Y + f_\Pi(C_\Pi) + g_\Pi(S_\Pi). \tag{C.41}$$

3. *For a constant $\kappa > 0$, uniform over the local region,*

$$\|g_\Pi(S_\Pi) + \mathbb{E}\{f_\Pi(C_\Pi) \mid S_\Pi\}\|_2 \ge \kappa \|g_\Pi(S_\Pi)\|_2. \tag{C.42}$$

Then

$$\mathcal{J}_{2,k}(\Pi) = \|g_\Pi(S_\Pi) + \mathbb{E}\{f_\Pi(C_\Pi) \mid S_\Pi\}\|_2^2, \tag{C.43}$$

$$\mathcal{K}_{2,k}(\Pi) = \|g_\Pi(S_\Pi) - \mathbb{E}\{g_\Pi(S_\Pi) \mid C_\Pi\}\|_2^2, \tag{C.44}$$

and

$$\mathcal{K}_{2,k}(\Pi) \le \kappa^{-2} \mathcal{J}_{2,k}(\Pi). \tag{C.45}$$

Moreover,

$$\|\mathbb{E}(Y \mid U) - \mathbb{E}(Y \mid C_\Pi)\|_2^2 = \mathcal{K}_{2,k}(\Pi).$$

Thus, in this special model, (C.46) *and* (C.47) *below hold with $C_{\mathrm{leak}} = \kappa^{-2}$, $\eta_{\mathrm{leak},n} = 0$, and $C_{Y,\ell} = 1$.*

Proof. For squared loss, conditional expectation is the $L_2(P)$ projection. The class assumptions therefore give

$$\mathcal{J}_{2,k}(\Pi) = \mathrm{Var}\{\mathbb{E}(Y \mid S_\Pi)\}.$$

Taking the conditional expectation of (C.41) given S_Π, and using the centering of f_Π and g_Π, yields

$$\mathbb{E}(Y \mid S_\Pi) - \mu_Y = g_\Pi(S_\Pi) + \mathbb{E}\{f_\Pi(C_\Pi) \mid S_\Pi\},$$

which proves (C.43). The nested square-loss risks defining $\mathcal{K}_{2,k}$ similarly satisfy

$$\mathcal{K}_{2,k}(\Pi) = \|\mathbb{E}(Y \mid C_\Pi, S_\Pi) - \mathbb{E}(Y \mid C_\Pi)\|_2^2.$$

By (C.41), the difference inside this norm is

$$g_\Pi(S_\Pi) - \mathbb{E}\{g_\Pi(S_\Pi) \mid C_\Pi\},$$

which proves (C.44). Conditional expectation is an $L_2(P)$ orthogonal projection, so

$$\mathcal{K}_{2,k}(\Pi) \le \|g_\Pi(S_\Pi)\|_2^2.$$

Combining this contraction with the visibility condition (C.42) and (C.43) gives

$$\mathcal{K}_{2,k}(\Pi) \le \|g_\Pi\|_2^2 \le \kappa^{-2}\mathcal{J}_{2,k}(\Pi),$$

proving (C.45). Finally, (C_Π, S_Π) and U generate the same sigma-field, so the preceding nested-risk identity is exactly the displayed conditional-mean identity. This proves all claims. □

Remark C.1 (Interpretable no-suppression conditions). Let $q_\Pi(S_\Pi) = \mathbb{E}\{f_\Pi(C_\Pi) \mid S_\Pi\}$. A sufficient strict no-suppression condition is $\langle g_\Pi, q_\Pi\rangle_{L_2(P)} \ge 0$, which permits $\kappa = 1$. More generally, the uniform finite-cancellation condition $\|q_\Pi\|_2 \le (1-\kappa)\|g_\Pi\|_2$, for $\kappa \in (0,1]$, implies (C.42) by the reverse triangle inequality. Additivity alone does not imply either condition.

Remark C.2 (XOR and suppressor counterexamples). Some interaction and visibility restriction is indispensable for a bridge of this form, although the preceding conditions need not be minimal. First, let $C, S \overset{\text{iid}}{\sim} \text{Bernoulli}(1/2)$ and $Y = C \oplus S$. Then $\mathbb{E}(Y \mid S) = \mathbb{E}(Y \mid C) = 1/2$, hence $\mathcal{J}_2 = 0$, whereas $\mathbb{E}(Y \mid C, S) = Y$ and $\mathcal{K}_2 = \mathbb{E}(Y - 1/2)^2 = 1/4$. This XOR example violates the additive conditional-mean structure.

Additivity by itself is still insufficient. Let $S, U_0 \overset{\text{iid}}{\sim} N(0,1)$, $C = S + U_0$, and $Y = U_0 = C - S$. Then $\mathbb{E}(Y \mid C, S) = C - S = f(C) + g(S)$ is additive, but $\mathbb{E}(Y \mid S) = 0$, so $\mathcal{J}_2 = 0$. Since $\mathbb{E}(Y \mid C) = C/2$,

$$\mathcal{K}_2 = \mathbb{E}\{(C-S) - C/2\}^2 = \tfrac{1}{2} > 0.$$

Here $\mathbb{E}\{f(C) \mid S\} = S = -g(S)$, so exact marginal suppression violates (C.42). Thus no finite universal constant can bound conditional leakage by marginal g_2 without an interaction restriction and a visibility condition.

Proposition C.5 (g_2 feasibility controls pseudo-true outcome leakage under a marginal-to-conditional bridge)**.** *Assume that $\mathcal{J}_{2,k}$ is locally Lipschitz around $\Pi^\dagger$ with constant*

$$L_{J,2} = \sup_{\Pi \in \mathcal{B}(\Pi^\dagger, r_0)} \|\operatorname{grad}_\Pi \mathcal{J}_{2,k}(\Pi)\| < \infty.$$

Suppose, only for this proposition, that on the local region

$$\mathcal{K}_{2,k}(\Pi) \le C_{\text{leak}}\mathcal{J}_{2,k}(\Pi) + \eta_{\text{leak},n}, \qquad \eta_{\text{leak},n} \to 0, \tag{C.46}$$

and that

$$\|\mathbb{E}(Y \mid U) - \mathbb{E}(Y \mid Z_c(\Pi))\|_{L_2(P)}^2 \le C_{Y,\ell}\mathcal{K}_{2,k}(\Pi). \tag{C.47}$$

Proposition C.4 verifies both conditions with $C_{\text{leak}} = \kappa^{-2}$, $\eta_{\text{leak},n} = 0$, and $C_{Y,\ell} = 1$ in its restricted squared-loss model; outside such a structure, (C.46) *remains an explicit assumption. For squared loss with rich prediction classes, $C_{Y,\ell} = 1$; for bounded categorical outcomes under log loss,* (C.47) *follows from conditional Pinsker. Then Theorem 2 implies*

$$\mathcal{J}_{2,k}(\Pi^\dagger) \le c_{2,n} + \delta_{\text{feas},n} + L_{J,2}\delta_{\Pi,n}^\dagger, \tag{C.48}$$

and, with

$$a_{\psi,n} := \|\mathbb{E}(Y \mid X) - \mathbb{E}(Y \mid U)\|_{L_2(P)},$$

$$\|\mathbb{E}(Y \mid X) - \mathbb{E}(Y \mid Z^\dagger)\|_{L_2(P)} \le a_{\psi,n} + \Big[C_{Y,\ell}C_{\text{leak}}\{c_{2,n} + \delta_{\text{feas},n} + L_{J,2}\delta_{\Pi,n}^\dagger\} + C_{Y,\ell}\eta_{\text{leak},n}\Big]^{1/2}. \tag{C.49}$$

Here $\delta_{\text{feas},n}$ is the population-profile feasibility envelope from Theorem 2: the multiplier route includes $\xi_{g,2,n}^{\text{prof}} + \eta_{2,n}^{\text{sur,val}}$, while the objective route includes the PL term induced by $\eta_{\nabla,n}^{\text{impl}}$. Therefore any asymptotic use of this proposition requires these bridge contributions, together with $c_{2,n} + L_{J,2}\delta_{\Pi,n}^\dagger + \eta_{\text{leak},n}$, to vanish at the rate imposed downstream.

Proof. The mean-value inequality on the local projection manifold gives

$$|\mathcal{J}_{2,k}(\Pi^\dagger) - \mathcal{J}_{2,k}(\widehat{\Pi}_k)| \le L_{J,2}\|\widehat{\Pi}_k - \Pi^\dagger\|_F.$$

Theorem 2 gives

$$\mathcal{J}_{2,k}(\widehat{\Pi}_k) \le c_{2,n} + \delta_{\mathrm{feas},n}, \qquad \|\widehat{\Pi}_k - \Pi^\dagger\|_F \le \delta^\dagger_{\Pi,n},$$

which proves (C.48). Next,

$$\mathbb{E}(Y \mid X) - \mathbb{E}(Y \mid Z^\dagger) = \{\mathbb{E}(Y \mid X) - \mathbb{E}(Y \mid U)\} + \{\mathbb{E}(Y \mid U) - \mathbb{E}(Y \mid Z^\dagger)\}.$$

Minkowski's inequality, (C.47), and (C.46) yield

$$\begin{aligned}\|\mathbb{E}(Y \mid X) - \mathbb{E}(Y \mid Z^\dagger)\|_2 &\le a_{\psi,n} + C_{Y,\ell}^{1/2}\mathcal{K}_{2,k}(\Pi^\dagger)^{1/2} \\ &\le a_{\psi,n} + \{C_{Y,\ell}C_{\mathrm{leak}}\mathcal{J}_{2,k}(\Pi^\dagger) + C_{Y,\ell}\eta_{\mathrm{leak},n}\}^{1/2}.\end{aligned}$$

Substituting (C.48) proves (C.49). □

D Density-ratio Calibration and Fixed-center Expansion

D.1 Notation and Sampling Structure

For fold k, let $\mathcal{T}_k$ be the sigma-field generated by the training observations, their generated training groups, all fitted nuisance functions, and all training-side tuning decisions. Define

$$\mathbb{P}_{k,n}f = \frac{1}{|I_k|}\sum_{i\in I_k} f(X_i, Y_i), \qquad \mathbb{Q}_{k,n}g = \frac{1}{|I_k|}\sum_{i\in I_k}\frac{1}{m}\sum_{c=1}^{m} g(\widetilde{X}^{\mathrm{eval}}_{i,c}).$$

Conditional on $\mathcal{T}_k$, the fold-k nuisance functions are fixed and the held-out complete clusters are independent. For the corresponding conditional population laws, write $P_k f = \mathbb{E}\{f(X,Y) \mid \mathcal{T}_k\}$ and $Q_k g = \mathbb{E}\{g(\widetilde{X}^{\mathrm{eval}}) \mid \mathcal{T}_k\}$. These operators are used only for fold-specific population comparisons; the reported estimator and variance remain based on complete held-out groups. The Q_{eval} generator for a held-out complete cluster is conditionally label-free: given the theoretical covariate X_i and external generation randomness, it does not use Y_i. When the prompt is built from auxiliary source context rather than only the displayed text, that context is part of X_i.

Let $\widehat{Z}_k(x) = \widehat{V}_k^\top\psi(x)$, $P_{k,Z} = \mathcal{L}\{\widehat{Z}_k(X) \mid \mathcal{T}_k\}$, and $Q_{k,Z} = \mathcal{L}\{\widehat{Z}_k(\widetilde{X}^{\mathrm{eval}}) \mid \mathcal{T}_k\}$. Define $w_k^\circ = \frac{dP_{k,Z}}{dQ_{k,Z}}$, $\eta_k^\circ \in \arg\min_\eta \mathcal{R}^{\mathrm{tr}}_{y,k}(\widehat{V}_k, \eta)$, and $h_k^\circ = h_{\eta_k^\circ}$. Thus η_k°, equivalently $s_{\eta_k^\circ}$, minimizes the unpenalized population outcome risk at the learned projection, while $h_k^\circ = h_{\eta_k^\circ}$ is its associated scalar inference score. For a categorical proper-loss learner, h_k° is the mean induced by the oracle class-probability vector. Set $\varphi_k^\circ(x) = w_k^\circ\{\widehat{Z}_k(x)\}\widehat{h}_k\{\widehat{Z}_k(x)\}$, $\widehat{\varphi}_k(x) = \widehat{w}_k\{\widehat{Z}_k(x)\}\widehat{h}_k\{\widehat{Z}_k(x)\}$, and $\widehat{b}_k(x,y) = y - \widehat{h}_k\{\widehat{Z}_k(x)\}$. The deterministic fixed center $(\Pi^\dagger, h^\dagger, w^\dagger, \varphi^\dagger, b^\dagger)$ is defined in Section 4; the definitions above agree with its learned fold-wise analogues.

Define

$$\Delta_{b,n} = \sum_k \pi_k\|\widehat{b}_k - b^\dagger\|^2_{L_2(P)}, \qquad \Delta_{\varphi,n} = \sum_k \pi_k\|\widehat{\varphi}_k - \varphi^\dagger\|^2_{L_2(Q)}.$$

D.2 Assumption C: Overlap and Local Nuisance Stability

On a geodesically convex neighborhood $\mathcal{B}_\Pi$ of $\Pi^\dagger$, choose a smooth local section $V_k(\Pi) \in \mathrm{St}(d,r)$ such that $V_k(\Pi)V_k(\Pi)^\top = \Pi$, and define $Z_{k,\Pi}(x) = V_k(\Pi)^\top\psi(x)$, $P_{k,\Pi} = \mathcal{L}\{Z_{k,\Pi}(X) \mid \mathcal{T}_k\}$, and $Q_{k,\Pi} = \mathcal{L}\{Z_{k,\Pi}(\widetilde{X}^{\mathrm{eval}}) \mid \mathcal{T}_k\}$, with $w^\circ_{k,\Pi} = dP_{k,\Pi}/dQ_{k,\Pi}$ whenever this derivative exists.

Assumption C (overlap and local nuisance stability). For every fold and every $\Pi \in \mathcal{B}_\Pi$, $P_{k,\Pi} \ll Q_{k,\Pi}$, and $P_{Z^\dagger} \ll Q_{Z^\dagger}$. In particular, $P_{k,Z} \ll Q_{k,Z}$ at the learned projection on the localization event. The bounded-overlap condition gives $0 \le w^\dagger \le M_w$ and $\sup_{k,\Pi\in\mathcal{B}_\Pi}\|w^\circ_{k,\Pi}\|_\infty \le M_w$, while the fitted nuisance functions satisfy, with probability tending to one, $0 \le \widehat{w}_k \le M_{\widehat{w}}$ and $|\widehat{h}_k| \le M_h$. These envelopes may be enforced by the clipping already used in computation, or may follow from

normalized bounded features together with compact parameter sets; the theory does not require an additional post-fitting operation when the latter route is used. The fold-specific population oracle predictors obey, with probability tending to one,

$$\sup_{1\le k\le K}\left\{\|h_k^\circ\|_{L_2(P_{k,Z})}+\|h_k^\circ\|_{L_2(Q_{k,Z})}\right\}<\infty.$$

For $h^\dagger(z)=\mathbb{E}(Y\mid Z^\dagger=z)$, $\|h^\dagger\|^2_{L_2(P)}\le\mathbb{E}(Y^2)$. The displayed bound for h_k° is automatic when the oracle parameters belong to the same compact parameter set used in computation and the feature map has a bounded second moment. Using the preceding local section, define the population outcome branch by

$$\eta^\circ_{k,\Pi}\in\arg\min_\eta \mathcal{R}^{\mathrm{tr}}_{y,k}\{V_k(\Pi),\eta\},\qquad \mathcal{H}_k(\Pi)(x)=h_{\eta^\circ_{k,\Pi}}\{Z_{k,\Pi}(x)\}.$$

Define

$$\mathcal{F}_k(\Pi)(x)=w^\circ_{k,\Pi}\{Z_{k,\Pi}(x)\}\mathcal{H}_k(\Pi)(x).$$

Rotation compatibility makes the resulting prediction and weighted-prediction functions invariant to the choice of local section. The selected branches are locally unique and the two maps are Fréchet differentiable with uniformly bounded derivatives. The approximation term $a_{h,\dagger,n}$ therefore includes prediction-class approximation and any local pseudo-true discrepancy between the training oracle at $\Pi^\dagger$ and $h^\dagger(z)=\mathbb{E}(Y\mid Z^\dagger=z)$.

Define the explicit prediction-class approximation error

$$a^2_{h,\dagger,n}:=\sum_{k=1}^K\pi_k\|\mathcal{H}_k(\Pi^\dagger)-h^\dagger(Z^\dagger)\|^2_{L_2(P)}. \tag{D.1}$$

Correct specification is the special case $a_{h,\dagger,n}=0$; the theory retains the approximation term explicitly.

D.3 Derived Ratio Bound

Proposition D.1 (A sufficient density-ratio L_2 consistency condition)**.** *Let $\mathcal{L}_{w,k}(\Pi,w)$ and $\widehat{\mathcal{L}}_{w,k}(\Pi,w)$ be the population and empirical ratio criteria and let $w^\circ_{k,\Pi}=dP_{k,\Pi}/dQ_{k,\Pi}$. Suppose, for this proposition, that uniformly over $\Pi\in\mathcal{B}_\Pi$,*

$$\mathcal{L}_{w,k}(\Pi,w)-\mathcal{L}_{w,k}(\Pi,w^\circ_{k,\Pi})\ge\kappa_w\|w-w^\circ_{k,\Pi}\|^2_{L_2(Q_{k,\Pi})}.$$

Assume that the joint complete-block loss class indexed by (Π,w) is nonnegative with oscillation envelope B_w, meaning that every block-level loss in the class lies in $[0,B_w]$, and has Rademacher complexity at most $r_{w,n}$, that for every local Π there exists $w^{\mathrm{app}}_{k,\Pi}\in\mathcal{W}_k$ satisfying

$$\sup_{\Pi\in\mathcal{B}_\Pi}\{\mathcal{L}_{w,k}(\Pi,w^{\mathrm{app}}_{k,\Pi})-\mathcal{L}_{w,k}(\Pi,w^\circ_{k,\Pi})\}\le a^2_{w,n},$$

and that at the learned projection

$$\widehat{\mathcal{L}}_{w,k}(\widehat{\Pi}_k,\widehat{w}_k)\le\inf_{w\in\mathcal{W}_k}\widehat{\mathcal{L}}_{w,k}(\widehat{\Pi}_k,w)+\eta^{\mathrm{opt}}_{w,n}.$$

Then, with probability at least $1-\delta$ simultaneously over the fixed number of folds,

$$\sup_{\Pi\in\mathcal{B}_\Pi,w\in\mathcal{W}_k}|\widehat{\mathcal{L}}_{w,k}(\Pi,w)-\mathcal{L}_{w,k}(\Pi,w)|\le\xi_{w,n}:=2r_{w,n}+B_w\sqrt{\frac{\log(2K/\delta)}{2n^{\mathrm{tr}}_k}},$$

and, with $w^\circ_k=w^\circ_{k,\widehat{\Pi}_k}$,

$$\|\widehat{w}_k-w^\circ_k\|^2_{L_2(Q_{k,Z})}\le\frac{2\xi_{w,n}+a^2_{w,n}+\eta^{\mathrm{opt}}_{w,n}}{\kappa_w}. \tag{D.2}$$

Consequently,

$$\varepsilon_{w,n}:=\left\{\sum_k\pi_k\|\widehat{w}_k-w^\circ_k\|^2_{L_2(Q_{k,Z})}\right\}^{1/2}=O_p\{\xi^{1/2}_{w,n}+a_{w,n}+(\eta^{\mathrm{opt}}_{w,n})^{1/2}\}.$$

This proposition is one sufficient learner-level route to $\varepsilon_{w,n}$; Proposition D.2 handles the prediction-targeted calibration rate.

Proof. Treat each complete real/synthetic training group as one independent block. Replacing one block changes the empirical criterion by at most B_w/n_k^{tr}. McDiarmid's inequality and ghost-sample symmetrization therefore give the displayed uniform bound; the supremum includes Π, so the event may be evaluated at the random $\widehat{\Pi}_k$.

On that event,

$$\widehat{\mathcal{L}}_{w,k}(\widehat{\Pi}_k, \widehat{w}_k) \le \widehat{\mathcal{L}}_{w,k}(\widehat{\Pi}_k, w^{\mathrm{app}}_{k,\widehat{\Pi}_k}) + \eta^{\mathrm{opt}}_{w,n}.$$

Add and subtract population criteria twice to obtain

$$\mathcal{L}_{w,k}(\widehat{\Pi}_k, \widehat{w}_k) \le \mathcal{L}_{w,k}(\widehat{\Pi}_k, w^{\mathrm{app}}_{k,\widehat{\Pi}_k}) + 2\xi_{w,n} + \eta^{\mathrm{opt}}_{w,n}.$$

Subtract $\mathcal{L}_{w,k}(\widehat{\Pi}_k, w_k^\circ)$, use the approximation bound, and apply population curvature:

$$\kappa_w \|\widehat{w}_k - w_k^\circ\|^2_{L_2(Q_{k,Z})} \le 2\xi_{w,n} + a^2_{w,n} + \eta^{\mathrm{opt}}_{w,n}.$$

Averaging over the fixed number of folds proves the final rate. □

D.4 Ratio-tangent Control of the Prediction-targeted Calibration Remainder

Define the fold-level and aggregate prediction-targeted calibration discrepancies by

$$\beta_{w,k,n} := |Q\{\widehat{\varphi}_k - \varphi_k^\circ\}| = \left|Q_k(\widehat{w}_k\widehat{h}_k) - P_k(\widehat{h}_k)\right|, \qquad \beta_{w,n} = \sum_{k=1}^K \pi_k \beta_{w,k,n}. \tag{D.3}$$

Proposition D.2 (Prediction-targeted calibration from ratio-tangent near-orthogonality)**.** *For fold k, let $w_k^\circ = dP_{k,Z}/dQ_{k,Z}$ denote the exact population ratio. Suppose the implemented ratio family is $\{w_{k,\gamma} : \gamma \in \Gamma_k\}$, the fitted ratio is $\widehat{w}_k = w_{k,\widehat{\gamma}_k}$, and γ_k° is the pseudo-true parameter selected by the population ratio criterion. Define the prediction-direction model-approximation bias*

$$a_{\beta,k,n} := \left|Q_{k,Z}\left[\{w_{k,\gamma_k^\circ} - w_k^\circ\}\widehat{h}_k\right]\right|. \tag{D.4}$$

If $a^{\mathrm{mod}}_{w,k,n} = \|w_{k,\gamma_k^\circ} - w_k^\circ\|_{L_2(Q_{k,Z})}$, then

$$a_{\beta,k,n} \le \|\widehat{h}_k\|_{L_2(Q_{k,Z})} a^{\mathrm{mod}}_{w,k,n} \le M_h a^{\mathrm{mod}}_{w,k,n}. \tag{D.5}$$

Assume that $\gamma \mapsto w_{k,\gamma}$ is twice Fréchet differentiable into $L_2(Q_{k,Z})$, with

$$\|\dot{w}_{k,\gamma_k^\circ}[u]\|_{L_2(Q_{k,Z})} \le L_{w,1,k}\|u\|,$$

$$\|w_{k,\gamma} - w_{k,\gamma_k^\circ} - \dot{w}_{k,\gamma_k^\circ}[\gamma - \gamma_k^\circ]\|_{L_2(Q_{k,Z})} \le L_{w,2,k}\|\gamma - \gamma_k^\circ\|^2.$$

If hard clipping is used, these local statements are required only on the interior event where clipping is inactive; alternatively use the derivative of the smooth clipping map actually implemented.

Define

$$r_{\gamma,k,n} = \|\widehat{\gamma}_k - \gamma_k^\circ\|,$$

$$\chi_{w,k,n} = \sup_{\|u\|\le 1} \left|Q_{k,Z}\{\dot{w}_{k,\gamma_k^\circ}[u]\widehat{h}_k\}\right|. \tag{D.6}$$

Then

$$\beta_{w,k,n} \le a_{\beta,k,n} + \chi_{w,k,n} r_{\gamma,k,n} + L_{w,2,k}\|\widehat{h}_k\|_{L_2(Q_{k,Z})} r^2_{\gamma,k,n}. \tag{D.7}$$

Let

$$a_{\beta,n} = \sum_k \pi_k a_{\beta,k,n}, \qquad r^2_{\gamma,n} = \sum_k \pi_k r^2_{\gamma,k,n}, \qquad \chi^2_{w,n} = \sum_k \pi_k \chi^2_{w,k,n}.$$

Under $|\widehat{h}_k| \le M_h$ and uniform second-derivative bounds,

$$\beta_{w,n} \le a_{\beta,n} + \chi_{w,n} r_{\gamma,n} + C_w M_h r^2_{\gamma,n}. \tag{D.8}$$

Thus

$$\sqrt{n}\{a_{\beta,n} + \chi_{w,n} r_{\gamma,n} + r^2_{\gamma,n}\} = o_p(1) \tag{D.9}$$

implies $\sqrt{n}\,\beta_{w,n} = o_p(1)$.

Proof. Decompose the exact ratio error as

$$\widehat{w}_k - w_k^\circ = \{w_{k,\widehat{\gamma}_k} - w_{k,\gamma_k^\circ}\} + \{w_{k,\gamma_k^\circ} - w_k^\circ\}.$$

Multiplying by $\widehat{h}_k$, taking $Q_{k,Z}$-expectations, and using the triangle inequality gives

$$\beta_{w,k,n} \le a_{\beta,k,n} + \left| Q_{k,Z}\{(w_{k,\widehat{\gamma}_k} - w_{k,\gamma_k^\circ})\widehat{h}_k\} \right|.$$

Let $\delta\gamma_k = \widehat{\gamma}_k - \gamma_k^\circ$. Taylor expansion in $L_2(Q_{k,Z})$ gives

$$w_{k,\widehat{\gamma}_k} - w_{k,\gamma_k^\circ} = \dot{w}_{k,\gamma_k^\circ}[\delta\gamma_k] + R_{w,k}, \qquad \|R_{w,k}\|_2 \le L_{w,2,k} r_{\gamma,k,n}^2.$$

Therefore

$$\begin{aligned}\beta_{w,k,n} &\le a_{\beta,k,n} + \left| Q_{k,Z}\{\dot{w}_{k,\gamma_k^\circ}[\delta\gamma_k]\widehat{h}_k\} \right| + \|R_{w,k}\|_2 \|\widehat{h}_k\|_2 \\ &\le a_{\beta,k,n} + \chi_{w,k,n} r_{\gamma,k,n} + L_{w,2,k}\|\widehat{h}_k\|_2 r_{\gamma,k,n}^2,\end{aligned}$$

which proves (D.7). Cauchy–Schwarz gives

$$\sum_k \pi_k \chi_{w,k,n} r_{\gamma,k,n} \le \chi_{w,n} r_{\gamma,n},$$

and the quadratic terms average to at most a constant times $r_{\gamma,n}^2$. This proves (D.8); multiplying by $\sqrt{n}$ proves the last assertion. □

Corollary D.1 (Regular finite-dimensional ratio learner)**.** *If $r_{\gamma,n} = O_p(n^{-1/2})$, $\chi_{w,n} = o_p(1)$, and $\sqrt{n}\, a_{\beta,n} = o_p(1)$, then $\sqrt{n}\,\beta_{w,n} = o_p(1)$. Under exact ratio-model specification, $a_{\beta,n} = 0$. Indeed,*

$$\sqrt{n}\,\chi_{w,n} r_{\gamma,n} = O_p(\chi_{w,n}) = o_p(1), \qquad \sqrt{n}\, r_{\gamma,n}^2 = O_p(n^{-1/2}) = o_p(1).$$

If, in addition, $\widehat{h}_k \to h_k^\circ$ in $L_2(Q_{k,Z})$, exact population tangent orthogonality

$$Q_{k,Z}\{\dot{w}_{k,\gamma_k^\circ}[u] h_k^\circ\} = 0 \qquad \textit{for all } u$$

implies

$$\begin{aligned}\chi_{w,k,n} &= \sup_{\|u\|\le 1} \left| Q_{k,Z}\{\dot{w}_{k,\gamma_k^\circ}[u]\widehat{h}_k\} \right| \\ &= \sup_{\|u\|\le 1} \left| Q_{k,Z}\{\dot{w}_{k,\gamma_k^\circ}[u](\widehat{h}_k - h_k^\circ)\} \right| \\ &\le L_{w,1,k}\|\widehat{h}_k - h_k^\circ\|_{L_2(Q_{k,Z})}.\end{aligned}$$

Thus $\chi_{w,n} = o_p(1)$.

Remark D.1 (Ratio-tangent control is a separate interface). Freezing Q_{style}, Q_{eval}, and the encoder fixes the external mechanism, but the target ratio $w_k^\circ = dP_{k,Z}/dQ_{k,Z}$ still depends on the target law, learned coordinates, and ratio family. Consequently, neither pretraining nor g_2 feasibility implies $\chi_{w,n} = o_p(1)$. Exact ratio specification sets $a_{\beta,n} = 0$ but does not set $\chi_{w,n} = 0$; the tangent condition above must be verified separately.

D.5 Exact Decomposition and Prediction-Targeted Calibration

Lemma D.1 (Exact bridge and fixed-center decomposition)**.** *The estimator satisfies*

$$\widehat{\mu} - \mu = \Gamma_{n,\dagger} + R_{c,n} + R_{w,n},$$

where

$$\Gamma_{n,\dagger} = \sum_k \pi_k\{(\mathbb{Q}_{k,n} - Q)\varphi^\dagger + (\mathbb{P}_{k,n} - P)b^\dagger\},$$

$$R_{c,n} = \sum_k \pi_k\{(\mathbb{Q}_{k,n} - Q)(\widehat{\varphi}_k - \varphi^\dagger) + (\mathbb{P}_{k,n} - P)(\widehat{b}_k - b^\dagger)\},$$

$$R_{w,n} = \sum_k \pi_k Q(\widehat{\varphi}_k - \varphi_k^\circ), \qquad |R_{w,n}| \le \beta_{w,n}.$$

Moreover,

$$Q(\widehat{\varphi}_k - \varphi_k^\circ) = Q_{k,Z}\{(\widehat{w}_k - w_k^\circ)\widehat{h}_k\} = Q_k(\widehat{w}_k\widehat{h}_k) - P_k(\widehat{h}_k).$$

Proof. For one fold,

$$\mathbb{Q}_{k,n}\widehat{\varphi}_k + \mathbb{P}_{k,n}\widehat{b}_k - \mu = (\mathbb{Q}_{k,n} - Q)\widehat{\varphi}_k + (\mathbb{P}_{k,n} - P)\widehat{b}_k + Q\widehat{\varphi}_k + P\widehat{b}_k - \mu.$$

The exact ratio identity gives $Q\varphi_k^\circ + P\widehat{b}_k = \mu$. Hence the population remainder is $Q(\widehat{\varphi}_k - \varphi_k^\circ)$. Add and subtract $(\varphi^\dagger, b^\dagger)$ in the empirical-process terms, multiply by π_k, and sum over folds. The calibration identity follows from $Q_{k,Z}(w_k^\circ \widehat{h}_k) = P_{k,Z}(\widehat{h}_k)$. The triangle inequality gives $|R_{w,n}| \le \beta_{w,n}$. □

D.6 Derived Prediction and Center Stability

Proposition D.3 (Post-representation outcome refit and fixed-center stability)**.** *Let η_k° be the unpenalized population minimizer defined above and $h_k^\circ = h_{\eta_k^\circ}$ its associated scalar score. Let $\widehat{\eta}_k^{\mathrm{ref}}$ denote the coefficient returned by the post-representation outcome refit, so that $\widehat{h}_k = h_{\widehat{\eta}_k^{\mathrm{ref}}}$. Suppose this refit obeys*

$$\widehat{\mathcal{R}}^{\mathrm{tr}}_{y,k}(\widehat{V}_k, \widehat{\eta}_k^{\mathrm{ref}}) + \lambda_{h,k,n} J_k(\widehat{\eta}_k^{\mathrm{ref}}) \le \inf_{\eta} \left\{ \widehat{\mathcal{R}}^{\mathrm{tr}}_{y,k}(\widehat{V}_k, \eta) + \lambda_{h,k,n} J_k(\eta) \right\} + \varepsilon^{\mathrm{opt}}_{h,k,n}, \tag{D.10}$$

where $J_k \ge 0$, $\lambda_{h,k,n} \ge 0$, and $\varepsilon^{\mathrm{opt}}_{h,k,n} \ge 0$. This includes an exact ridge refit with $J_k(\eta) = \|\eta_{-0}\|^2/2$, where the intercept coordinate is excluded. Relative to the unpenalized population oracle, the rates below require $\sum_{k=1}^K \pi_k \lambda_{h,k,n} J_k(\eta_k^\circ) = o_p(1)$; a nonvanishing penalty on the mean-risk scale instead targets a penalized population oracle. Define

$$\Delta^{\mathrm{emp}}_{h,k,n} = \sup_{\eta \in \mathcal{H}_{k,n}} \left| \widehat{\mathcal{R}}^{\mathrm{tr}}_{y,k}(\widehat{V}_k, \eta) - \mathcal{R}^{\mathrm{tr}}_{y,k}(\widehat{V}_k, \eta) \right|, \quad \delta^{\mathrm{ref}}_{h,k,n} = 2\Delta^{\mathrm{emp}}_{h,k,n} + \lambda_{h,k,n} J_k(\eta_k^\circ) + \varepsilon^{\mathrm{opt}}_{h,k,n},$$

on a convex parameter set $\mathcal{H}_{k,n}$ containing the full line segment between η_k° and $\widehat{\eta}_k^{\mathrm{ref}}$. Assume that $\mathcal{R}^{\mathrm{tr}}_{y,k}(\widehat{V}_k, \cdot)$ is $\kappa^{\mathrm{ref}}_{h,k}$-strongly convex on this set and that the prediction map is locally Lipschitz in its finite-dimensional parameter:

$$\|h_{\eta_1} - h_{\eta_2}\|_{L_2(P_{k,Z})} \le L_{h,P,k}\|\eta_1 - \eta_2\|, \qquad \|h_{\eta_1} - h_{\eta_2}\|_{L_2(Q_{k,Z})} \le L_{h,Q,k}\|\eta_1 - \eta_2\|.$$

For a feature-linear learner $h_\eta(z) = g\{\eta^\top \Phi_k(z)\}$ with $\sup_t |g'(t)| \le L_g$, one may take

$$L_{h,P,k} = L_g\{P_{k,Z}\|\Phi_k(Z)\|^2\}^{1/2}, \qquad L_{h,Q,k} = L_g\{Q_{k,Z}\|\Phi_k(Z)\|^2\}^{1/2}.$$

Then

$$\|\widehat{\eta}_k^{\mathrm{ref}} - \eta_k^\circ\|^2 \le \frac{2\delta^{\mathrm{ref}}_{h,k,n}}{\kappa^{\mathrm{ref}}_{h,k}}, \tag{D.11}$$

$$\begin{aligned} e^2_{h,P,n} &:= \sum_k \pi_k \|\widehat{h}_k - h_k^\circ\|^2_{L_2(P_{k,Z})} \le C_{h,P}\delta^{\mathrm{ref}}_{h,n}, \\ e^2_{h,Q,n} &:= \sum_k \pi_k \|\widehat{h}_k - h_k^\circ\|^2_{L_2(Q_{k,Z})} \le C_{h,Q}\delta^{\mathrm{ref}}_{h,n}, \end{aligned} \tag{D.12}$$

where

$$\delta^{\mathrm{ref}}_{h,n} = \sum_k \pi_k \delta^{\mathrm{ref}}_{h,k,n}, \qquad C_{h,P} = 2\max_k L^2_{h,P,k}/\min_k \kappa^{\mathrm{ref}}_{h,k}, \qquad C_{h,Q} = 2\max_k L^2_{h,Q,k}/\min_k \kappa^{\mathrm{ref}}_{h,k}.$$

The center-replacement bounds become

$$\Delta^{1/2}_{b,n} \le C_b\{e_{h,P,n} + \delta^\dagger_{\Pi,n} + a_{h,\dagger,n}\}, \tag{D.13}$$

$$\Delta^{1/2}_{\varphi,n} \le C_\varphi\{\varepsilon_{w,n} + e_{h,Q,n} + \delta^\dagger_{\Pi,n} + a_{h,\dagger,n}\}. \tag{D.14}$$

Proof. Apply (D.10) with $\eta = \eta_k^\circ$. Because $J_k \ge 0$,

$$\widehat{\mathcal{R}}^{\mathrm{tr}}_{y,k}(\widehat{V}_k, \widehat{\eta}_k^{\mathrm{ref}}) \le \widehat{\mathcal{R}}^{\mathrm{tr}}_{y,k}(\widehat{V}_k, \eta_k^\circ) + \lambda_{h,k,n} J_k(\eta_k^\circ) + \varepsilon^{\mathrm{opt}}_{h,k,n}.$$

Adding and subtracting the two corresponding population risks and using the definition of $\Delta^{\mathrm{emp}}_{h,k,n}$ gives

$$\begin{aligned}\mathcal{R}^{\mathrm{tr}}_{y,k}(\widehat{V}_k, \widehat{\eta}^{\mathrm{ref}}_k) - \mathcal{R}^{\mathrm{tr}}_{y,k}(\widehat{V}_k, \eta^{\circ}_k) \\ \le 2\Delta^{\mathrm{emp}}_{h,k,n} + \lambda_{h,k,n} J_k(\eta^{\circ}_k) + \varepsilon^{\mathrm{opt}}_{h,k,n} = \delta^{\mathrm{ref}}_{h,k,n}.\end{aligned}$$

Strong convexity and the population optimality of η°_k imply

$$\mathcal{R}^{\mathrm{tr}}_{y,k}(\widehat{V}_k, \eta) - \mathcal{R}^{\mathrm{tr}}_{y,k}(\widehat{V}_k, \eta^{\circ}_k) \ge \frac{\kappa^{\mathrm{ref}}_{h,k}}{2} \|\eta - \eta^{\circ}_k\|^2.$$

Taking $\eta = \widehat{\eta}^{\mathrm{ref}}_k$ and combining the last two displays proves (D.11). The two parameter-to-prediction Lipschitz inequalities then give (D.12) after fold averaging.

For the real side, add and subtract the population outcome minimizer at the learned projection and the oracle branch at $\Pi^{\dagger}$:

$$\widehat{b}_k - b^{\dagger} = \{h^{\circ}_k(\widehat{Z}_k) - \widehat{h}_k(\widehat{Z}_k)\} + \{\mathcal{H}_k(\Pi^{\dagger}) - \mathcal{H}_k(\widehat{\Pi}_k)\} + \{h^{\dagger}(Z^{\dagger}) - \mathcal{H}_k(\Pi^{\dagger})\}.$$

Minkowski's inequality, local Fréchet smoothness, and the definition of $a_{h,\dagger,n}$ yield (D.13).

For the synthetic side,

$$\begin{aligned}\widehat{\varphi}_k - \varphi^{\dagger} = (\widehat{w}_k - w^{\circ}_k)\widehat{h}_k + w^{\circ}_k(\widehat{h}_k - h^{\circ}_k) \\ + \{\mathcal{F}_k(\widehat{\Pi}_k) - \mathcal{F}_k(\Pi^{\dagger})\} + w^{\dagger}\{\mathcal{H}_k(\Pi^{\dagger}) - h^{\dagger}(Z^{\dagger})\}.\end{aligned}$$

The first term is bounded by $M_h \|\widehat{w}_k - w^{\circ}_k\|_{L_2(Q)}$. The second term is bounded by $M_w e_{h,Q,k}$ or, using transport,

$$\|w^{\circ}_k(\widehat{h}_k - h^{\circ}_k)\|^2_{L_2(Q)} \le M_w \|\widehat{h}_k - h^{\circ}_k\|^2_{L_2(P)}.$$

The third term is controlled by local Fréchet smoothness, and the last by M_w times the prediction-class approximation error. Fold averaging and Minkowski prove (D.14). □

Lemma D.2 (Cross-fitted center replacement)**.** *Under the standing held-out complete-cluster design,*

$$\sqrt{n} R_{c,n} = O_p(\Delta^{1/2}_{\varphi,n} + \Delta^{1/2}_{b,n}).$$

Proof. Write

$$U_{k,n} = (\mathbb{Q}_{k,n} - Q)(\widehat{\varphi}_k - \varphi^{\dagger}), \qquad V_{k,n} = (\mathbb{P}_{k,n} - P)(\widehat{b}_k - b^{\dagger}).$$

Then $R_{c,n} = \sum_k \pi_k (U_{k,n} + V_{k,n})$. Conditional on $\mathcal{T}_k$, the nuisance functions are fixed. Jensen's inequality within a prompt group gives

$$\mathbb{E}(U^2_{k,n} \mid \mathcal{T}_k) \le \frac{1}{|I_k|} \|\widehat{\varphi}_k - \varphi^{\dagger}\|^2_{L_2(Q)},$$

and ordinary held-out independence gives

$$\mathbb{E}(V^2_{k,n} \mid \mathcal{T}_k) \le \frac{1}{|I_k|} \|\widehat{b}_k - b^{\dagger}\|^2_{L_2(P)}.$$

For any $M > 0$, conditional Markov's inequality yields

$$\Pr\left\{ \sqrt{|I_k|}\, |U_{k,n}| > M \|\widehat{\varphi}_k - \varphi^{\dagger}\|_{L_2(Q)} \,\middle|\, \mathcal{T}_k \right\} \le M^{-2},$$

and

$$\Pr\left\{ \sqrt{|I_k|}\, |V_{k,n}| > M \|\widehat{b}_k - b^{\dagger}\|_{L_2(P)} \,\middle|\, \mathcal{T}_k \right\} \le M^{-2}.$$

Since K is fixed, a union bound over the folds gives an event with probability at least $1 - 2K/M^2$ on which

$$\begin{aligned}\sqrt{n} \left| \sum_k \pi_k U_{k,n} \right| &\le M\sqrt{n} \sum_k \frac{\pi_k}{\sqrt{|I_k|}} \|\widehat{\varphi}_k - \varphi^{\dagger}\|_{L_2(Q)} \\ &= M \sum_k \sqrt{\pi_k} \|\widehat{\varphi}_k - \varphi^{\dagger}\|_{L_2(Q)} \\ &\le M\sqrt{K}\, \Delta^{1/2}_{\varphi,n},\end{aligned}$$

and similarly

$$\sqrt{n}\left|\sum_k \pi_k V_{k,n}\right| \le M\sqrt{K}\,\Delta_{b,n}^{1/2}.$$

Letting M increase after n gives

$$\sqrt{n}|R_{c,n}| = O_p(\Delta_{\varphi,n}^{1/2} + \Delta_{b,n}^{1/2}).$$

□

Full remainder bound used by the proof. The compact statement in the main text omits the componentwise rate bookkeeping. The exact decomposition used in the proof is

$$\begin{aligned}\sqrt{n}R_n = O_p\Big[\varepsilon_{w,n} + \{\delta_{h,n}^{\mathrm{ref}}\}^{1/2} + \delta_{\Pi,n}^{\dagger} + a_{h,\dagger,n}\Big] \\ + O_p\big[\sqrt{n}\{a_{\beta,n} + \chi_{w,n}r_{\gamma,n} + r_{\gamma,n}^2\}\big].\end{aligned} \tag{D.15}$$

D.7 Proof of Theorem 3

Proof of Theorem 3. The nuisance and targeted-calibration rates invoked in the theorem are, explicitly,

$$\varepsilon_{w,n} + \{\delta_{h,n}^{\mathrm{ref}}\}^{1/2} + \delta_{\Pi,n}^{\dagger} + a_{h,\dagger,n} = o_p(1), \qquad \sqrt{n}\{a_{\beta,n} + \chi_{w,n}r_{\gamma,n} + r_{\gamma,n}^2\} = o_p(1).$$

The exact decomposition gives $\widehat{\mu} - \mu = \Gamma_{n,\dagger} + R_{c,n} + R_{w,n}$. Proposition D.3 and Lemma D.2 imply

$$\sqrt{n}R_{c,n} = O_p\Big[\varepsilon_{w,n} + \{\delta_{h,n}^{\mathrm{ref}}\}^{1/2} + \delta_{\Pi,n}^{\dagger} + a_{h,\dagger,n}\Big].$$

Proposition D.2 yields

$$|R_{w,n}| \le \beta_{w,n} \le a_{\beta,n} + \chi_{w,n}r_{\gamma,n} + C_w M_h r_{\gamma,n}^2.$$

Multiplying the last display by $\sqrt{n}$ and combining both remainders proves (D.15). □

D.8 Approximate n-indexed Fixed-center Expansion

Theorem 3A (Approximate n-indexed fixed-center expansion)**.** *Consider an n-indexed version of the preceding construction. Let $(\Pi_n^{\dagger}, h_n^{\dagger}, w_n^{\dagger}, \varphi_n^{\dagger}, b_n^{\dagger})$ denote the row-n fixed-center objects, let P_n be the row-n real law, let Q_n be the corresponding marginal law of a Q_{eval} generated text, and let $\mu_n = P_n Y$. Define*

$$\Gamma_{n,\dagger}^{(A)} = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\left[b_n^{\dagger}(X_i, Y_i) - P_n b_n^{\dagger} + \frac{1}{m}\sum_{c=1}^{m}\{\varphi_n^{\dagger}(\widetilde{X}_{i,c}^{\mathrm{eval}}) - Q_n\varphi_n^{\dagger}\}\right].$$

Assume that Assumptions A–C, Theorem 2, and Propositions D.1, D.2, and D.3 hold uniformly after replacing the fixed center by the row-n fixed center. If

$$\varepsilon_{w,n} + \{\delta_{h,n}^{\mathrm{ref}}\}^{1/2} + \delta_{\Pi,n}^{\dagger} + a_{h,\dagger,n} \to 0$$

and

$$\sqrt{n}\{a_{\beta,n} + \chi_{w,n}r_{\gamma,n} + r_{\gamma,n}^2\} = o_p(1),$$

then

$$\widehat{\mu} - \mu_n = \Gamma_{n,\dagger}^{(A)} + o_p(n^{-1/2}).$$

When the target population law is fixed, $\mu_n = \mu$.

Proof. The exact decomposition in Lemma D.1 is algebraic, so the same argument applied row by row gives

$$\widehat{\mu} - \mu_n = \Gamma_{n,\dagger}^{(A)} + R_{c,n}^{(A)} + R_{w,n}^{(A)}.$$

The foldwise proof of Lemma D.2 uses only held-out complete-cluster independence conditional on $\mathcal{T}_k$. Under the uniform n-indexed conditions it yields

$$\sqrt{n}R_{c,n}^{(A)} = O_p\Big[\varepsilon_{w,n} + \{\delta_{h,n}^{\mathrm{ref}}\}^{1/2} + \delta_{\Pi,n}^{\dagger} + a_{h,\dagger,n}\Big],$$

where the center-replacement rates come from Proposition D.3. The prediction-targeted calibration part satisfies

$$|R_{w,n}^{(A)}| \le \beta_{w,n} \le a_{\beta,n} + \chi_{w,n} r_{\gamma,n} + C_w M_h r_{\gamma,n}^2$$

by Proposition D.2. The two displayed rate conditions therefore imply $\sqrt{n}\{R_{c,n}^{(A)} + R_{w,n}^{(A)}\} = o_p(1)$, proving the expansion. □

E Prompt-group Central Limit Theorem, Variance Consistency, and Wald Coverage

E.1 Fixed-center Group Notation

Under the fixed-group design with m generated texts per real unit, define

$$B_i^{\dagger} = b^{\dagger}(X_i, Y_i) - Pb^{\dagger}, \qquad A_{Q,i}^{\dagger} = \frac{1}{m}\sum_{c=1}^{m}\{\varphi^{\dagger}(\widetilde{X}_{i,c}^{\mathrm{eval}}) - Q\varphi^{\dagger}\}.$$

Writing $\xi_{i,c}^{\dagger} = \varphi^{\dagger}(\widetilde{X}_{i,c}^{\mathrm{eval}}) - Q\varphi^{\dagger}$, the fixed-$m$ group variance expands as

$$\sigma_Q^2 = \frac{1}{m^2}\left\{\sum_{c=1}^{m}\mathrm{Var}(\xi_{i,c}^{\dagger}) + 2\sum_{1\le c<c'\le m}\mathrm{Cov}(\xi_{i,c}^{\dagger}, \xi_{i,c'}^{\dagger})\right\}.$$

Thus the m generated siblings are averaged within a group but their covariance is not discarded. Then

$$\Gamma_{n,\dagger} = \frac{1}{n}\sum_{i=1}^{n}(B_i^{\dagger} + A_{Q,i}^{\dagger}).$$

For $i \in I_k$, let

$$\widehat{A}_{k,i} = \frac{1}{m}\sum_{c=1}^{m}\widehat{\varphi}_k(\widetilde{X}_{i,c}^{\mathrm{eval}}), \qquad \widehat{B}_{k,i} = \widehat{b}_k(X_i, Y_i).$$

The reported additive estimator is

$$\widehat{\sigma}_P^2 = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}(\widehat{B}_{k,i} - \bar{B}_k)^2,$$

$$\widehat{\sigma}_Q^2 = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}(\widehat{A}_{k,i} - \bar{A}_k)^2, \qquad \widehat{\sigma}_{\mathrm{tot}}^2 = \widehat{\sigma}_P^2 + \widehat{\sigma}_Q^2.$$

E.2 Full and Additive Score Variance

Let $H^{\dagger} = h^{\dagger}(Z^{\dagger})$, and define $V_Y = \mathrm{Var}(Y)$, $V_H = \mathrm{Var}(H^{\dagger})$, $V_Q = \mathrm{Var}(A_Q^{\dagger})$, and $C_{BQ} = \mathrm{Cov}(B^{\dagger}, A_Q^{\dagger})$. Because $H^{\dagger} = \mathbb{E}(Y \mid Z^{\dagger})$, the residual $B^{\dagger} = Y - H^{\dagger}$ is orthogonal to $H^{\dagger}$ and $\mathrm{Var}(B^{\dagger}) = V_Y - V_H$. The exact fixed-center score variance, before any covariance restriction, is therefore

$$\mathrm{Var}(B^{\dagger} + A_Q^{\dagger}) = V_Y - V_H + V_Q + 2C_{BQ}. \tag{E.1}$$

Moreover,

$$V_Q = \mathrm{Var}\{\mathbb{E}(A_Q^{\dagger} \mid X)\} + \mathbb{E}\{\mathrm{Var}(A_Q^{\dagger} \mid X)\},$$

so the group-statistic variance contains both between-source variation in the transported score and within-source generation noise. Under regime (E) below, $C_{BQ} = 0$; under regime (A), the row-average cross term is $o(1)$. Only under one of these covariance-control routes does (E.1) have the additive limit $\sigma_P^2 + \sigma_Q^2$ estimated by the reported procedure.

E.3 Moment Condition for Additive Inference

Throughout Appendix E, the standing complete-cluster design applies. For the exact fixed-law regime, the only additional moment requirement is $\mathbb{E}(Y^2) < \infty$. The indexed triangular-array regime instead uses the uniform $(4+\delta_0)$-moment condition stated in Proposition E.2. Because $w^\dagger$ is bounded by Assumption C and $h^\dagger = \mathbb{E}(Y \mid Z^\dagger)$ is square integrable when $\mathbb{E}(Y^2) < \infty$, the fixed-law synthetic group mean has finite second moment. Define $\sigma_P^2 := \mathrm{Var}(B^\dagger)$, $\sigma_Q^2 := \mathrm{Var}(A_Q^\dagger)$, and $0 < \sigma_P^2 + \sigma_Q^2 < \infty$. These conditions are used only by the additive central limit theorem and variance-consistency result below.

E.4 Derived Residual–synthetic Orthogonality

Proposition E.1 (Derived residual–synthetic covariance bound)**.** *Let $r_Y^\dagger(X) = \mathbb{E}(Y \mid X) - \mathbb{E}(Y \mid Z^\dagger)$, $r_Q^\dagger(X) = \mathbb{E}(A_Q^\dagger \mid X) - \mathbb{E}(A_Q^\dagger \mid Z^\dagger)$, and $\delta_{Q,n}^\dagger = \|r_Q^\dagger\|_{L_2(P)}$. Under the standing label-free generation design, $(\widetilde{X}_{i,1}^{\mathrm{eval}}, \ldots, \widetilde{X}_{i,m}^{\mathrm{eval}}) \perp Y_i \mid X_i$,*

$$\mathrm{Cov}(B^\dagger, A_Q^\dagger) = \mathbb{E}\{r_Y^\dagger(X) r_Q^\dagger(X)\}. \tag{E.2}$$

Consequently,

$$|\mathrm{Cov}(B^\dagger, A_Q^\dagger)| \le \|r_Y^\dagger\|_{L_2(P)} \delta_{Q,n}^\dagger. \tag{E.3}$$

The quantity $\delta_{Q,n}^\dagger$ is defined by the population law rather than imposed as an assumption. Under Assumption C and $\mathbb{E}(Y^2) < \infty$, it is automatically finite; indeed,

$$(\delta_{Q,n}^\dagger)^2 \le \|A_Q^\dagger\|_{L_2}^2 \le 4 M_w \mathbb{E}(Y^2). \tag{E.4}$$

On the joint event of Theorem 2, if the conditions of Proposition C.5 hold, then

$$|\mathrm{Cov}(B^\dagger, A_Q^\dagger)| \le \delta_{Q,n}^\dagger \left[a_{\psi,n} + \left\{ C_{Y,\ell} C_{\mathrm{leak}} \left(c_{2,n} + \delta_{\mathrm{feas},n} + L_{J,2} \delta_{\Pi,n}^\dagger \right) + C_{Y,\ell} \eta_{\mathrm{leak},n} \right\}^{1/2} \right] =: \mathfrak{r}_{\perp,n}. \tag{E.5}$$

For an indexed row-wise sequence, define $\mathfrak{r}_{\perp,n,i}$ by the same right-hand side with the row-specific quantities $\delta_{Q,n,i}^\dagger, a_{\psi,n,i}, c_{2,n,i}, \delta_{\mathrm{feas},n,i}, \delta_{\Pi,n,i}^\dagger$, and $\eta_{\mathrm{leak},n,i}$, namely

$$\mathfrak{r}_{\perp,n,i} := \delta_{Q,n,i}^\dagger \left[a_{\psi,n,i} + \left\{ C_{Y,\ell} C_{\mathrm{leak}} \left(c_{2,n,i} + \delta_{\mathrm{feas},n,i} + L_{J,2} \delta_{\Pi,n,i}^\dagger \right) + C_{Y,\ell} \eta_{\mathrm{leak},n,i} \right\}^{1/2} \right]. \tag{E.6}$$

Here the indexed quantities are conditional row-wise versions of the corresponding population or training-event bounds, with the constants in Proposition C.5 kept uniform in i and n. For the exact structural regime, Proposition E.1 yields additive orthogonality whenever either $r_Y^\dagger(X) = 0$ a.s. or $r_Q^\dagger(X) = 0$ a.s. More generally, for an approximate n-indexed sequence of working representations, prediction rules, calibration rules, or generation mechanisms, the product bound in (E.5) *is sufficient when its row-wise version converges to zero on average.*

Proof. Because $A_Q^\dagger$ is centered,

$$\mathrm{Cov}(B^\dagger, A_Q^\dagger) = \mathbb{E}[(Y - h^\dagger(Z^\dagger)) A_Q^\dagger].$$

The generation kernel depends on X and an external randomization variable but not on Y, so

$$\mathbb{E}(A_Q^\dagger \mid X, Y) = \mathbb{E}(A_Q^\dagger \mid X).$$

Writing $m_Q(X) = \mathbb{E}(A_Q^\dagger \mid X)$ and conditioning first on (X, Y) gives

$$\mathrm{Cov}(B^\dagger, A_Q^\dagger) = \mathbb{E}\{r_Y^\dagger(X) m_Q(X)\}.$$

Since $h^\dagger(Z^\dagger) = \mathbb{E}(Y \mid Z^\dagger)$,

$$\mathbb{E}\{r_Y^\dagger(X) \mid Z^\dagger\} = \mathbb{E}\{\mathbb{E}(Y \mid X) - \mathbb{E}(Y \mid Z^\dagger) \mid Z^\dagger\} = 0.$$

Therefore

$$\mathbb{E}\left[r_Y^\dagger(X)\mathbb{E}\{m_Q(X) \mid Z^\dagger\}\right] = 0,$$

and hence

$$\begin{aligned}\mathrm{Cov}(B^\dagger, A_Q^\dagger) &= \mathbb{E}\left[r_Y^\dagger(X)\{m_Q(X) - \mathbb{E}(m_Q(X) \mid Z^\dagger)\}\right] \\ &= \mathbb{E}\{r_Y^\dagger(X) r_Q^\dagger(X)\},\end{aligned}$$

which proves (E.2). Cauchy–Schwarz gives (E.3).

It remains to verify that the derived score residual has a finite second moment. Conditional expectation is an L_2 contraction, so

$$\delta_{Q,n}^\dagger \le \|A_Q^\dagger\|_2.$$

By Jensen's inequality within the prompt group,

$$\|A_Q^\dagger\|_2^2 \le \frac{1}{m}\sum_{c=1}^{m}\mathbb{E}\left[\{\varphi^\dagger(\widetilde{X}_c^{\mathrm{eval}}) - Q\varphi^\dagger\}^2\right] \le 4Q\{(\varphi^\dagger)^2\}.$$

Since $\varphi^\dagger = w^\dagger h^\dagger$ and $0 \le w^\dagger \le M_w$,

$$Q\{(\varphi^\dagger)^2\} = Q\{(w^\dagger)^2(h^\dagger)^2\} \le M_w Q\{w^\dagger(h^\dagger)^2\} = M_w P\{(h^\dagger)^2\} \le M_w \mathbb{E}(Y^2),$$

where the last inequality follows from the L_2 contraction property of conditional expectation. This proves (E.4).

Finally, Proposition C.5 supplies the displayed upper bound for $\|r_Y^\dagger\|_2$. Substituting it into (E.3) proves (E.5). This proves the product bound. In the exact structural regime, the displayed conditions make one factor in (E.2) vanish. In the approximate n-indexed regime, the same product bound is used row-wise and averaged. □

E.5 Additive Fixed-center Limit and Variance Consistency

Proposition E.2 (Additive fixed-center limit and componentwise variance consistency)**.** *Consider either of the following two regimes.*

(E) Exact structural regime. *The complete real-unit prompt groups are i.i.d. from a fixed joint law and* $\mathbb{E}(Y^2) < \infty$. *Proposition E.1 yields* $\mathrm{Cov}(B^\dagger, A_Q^\dagger) = \mathbb{E}\{r_Y^\dagger(X) r_Q^\dagger(X)\}$. *Assume that either* $r_Y^\dagger(X) = 0$ *a.s. or* $r_Q^\dagger(X) = 0$ *a.s. Then* $\mathrm{Cov}(B^\dagger, A_Q^\dagger) = 0$.

(A) Approximate n-indexed regime. *The target-population law may remain fixed, but the working fixed-center contributions are allowed to depend on* n*; alternatively, the complete-cluster law itself may vary with* n. *For row* n, *let* $\{(B_{n,i}^\dagger, A_{Q,n,i}^\dagger) : 1 \le i \le n\}$ *be independent complete clusters and define* $\sigma_{P,n}^2 = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}(B_{n,i}^\dagger)$ *and* $\sigma_{Q,n}^2 = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}(A_{Q,n,i}^\dagger)$. *Assume* $\sigma_{P,n}^2 \to \sigma_P^2$ *and* $\sigma_{Q,n}^2 \to \sigma_Q^2$,

$$\frac{1}{n}\sum_{i=1}^{n}\mathfrak{r}_{\perp,n,i} \longrightarrow 0, \tag{E.7}$$

where $\mathfrak{r}_{\perp,n,i}$ *denotes the row-wise version of the upper bound in Proposition E.1. Equivalently,*

$$|\mathrm{Cov}(B_{n,i}^\dagger, A_{Q,n,i}^\dagger)| \le \mathfrak{r}_{\perp,n,i}$$

for each row element. Assume that the centered component variables satisfy $\mathbb{E}B_{n,i}^\dagger = 0$ *and* $\mathbb{E}A_{Q,n,i}^\dagger = 0$, *and, for some* $\delta_0 > 0$,

$$\sup_{n\ge 1}\max_{1\le i\le n}\left\{\mathbb{E}|B_{n,i}^\dagger|^{4+\delta_0} + \mathbb{E}|A_{Q,n,i}^\dagger|^{4+\delta_0}\right\} < \infty. \tag{E.8}$$

Also assume that the deterministic fold proportions satisfy $\min_k |I_{k,n}|/n \geq \pi_{\min} > 0$. *Condition* (E.8) *is stronger than is needed for the CLT alone, but it simultaneously supplies the row-wise Lyapunov condition and the componentwise laws of large numbers required by the reported additive variance estimator.*

Under (E),

$$\sqrt{n}\,\Gamma_{n,\dagger} \xrightarrow{d} N(0, \sigma_P^2 + \sigma_Q^2).$$

Under (A), the same conclusion holds for $\sqrt{n}\,\Gamma_{n,\dagger}^{(A)}$. *If, in addition,* $\Delta_{b,n} = o_p(1)$ *and* $\Delta_{\varphi,n} = o_p(1)$, *then*

$$\widehat{\sigma}_P^2 \xrightarrow{p} \sigma_P^2, \qquad \widehat{\sigma}_Q^2 \xrightarrow{p} \sigma_Q^2.$$

Proof. Under (E), let $\xi_i^\dagger = B_i^\dagger + A_{Q,i}^\dagger$. Exact orthogonality gives

$$\mathrm{Var}(\xi_i^\dagger) = \sigma_P^2 + \sigma_Q^2,$$

and $\mathbb{E}|\xi_i^\dagger|^2 < \infty$. The ordinary i.i.d. central limit theorem gives

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i^\dagger \xrightarrow{d} N(0, \sigma_P^2 + \sigma_Q^2).$$

Under (A), write $\xi_{n,i}^\dagger = B_{n,i}^\dagger + A_{Q,n,i}^\dagger$. Its row-average variance is

$$\frac{1}{n} \sum_{i=1}^n \mathrm{Var}(\xi_{n,i}^\dagger) = \sigma_{P,n}^2 + \sigma_{Q,n}^2 + \frac{2}{n} \sum_{i=1}^n \mathrm{Cov}(B_{n,i}^\dagger, A_{Q,n,i}^\dagger).$$

Proposition E.1 applied row-wise and (E.7) give

$$\left| \frac{2}{n} \sum_{i=1}^n \mathrm{Cov}(B_{n,i}^\dagger, A_{Q,n,i}^\dagger) \right| \leq \frac{2}{n} \sum_{i=1}^n \mathfrak{r}_{\perp,n,i} \to 0.$$

Hence the row-average variance converges to $\sigma_P^2 + \sigma_Q^2$. By $|x+y|^{4+\delta_0} \leq 2^{3+\delta_0}(|x|^{4+\delta_0} + |y|^{4+\delta_0})$, (E.8) gives

$$\sup_{n,i} \mathbb{E}|\xi_{n,i}^\dagger|^{4+\delta_0} < \infty.$$

Taking the Lyapunov exponent to be $4 + \delta_0$,

$$\frac{\sum_{i=1}^n \mathbb{E}|\xi_{n,i}^\dagger|^{4+\delta_0}}{\{n(\sigma_P^2 + \sigma_Q^2 + o(1))\}^{2+\delta_0/2}} = O(n^{-1-\delta_0/2}) \to 0.$$

The Lyapunov central limit theorem for triangular arrays therefore gives the same normal limit.

It remains to prove componentwise variance consistency.

For the real-side variance, write

$$\widehat{B}_{k,i} = b^\dagger(X_i, Y_i) + d_{k,i}, \qquad d_{k,i} = \widehat{b}_k(X_i, Y_i) - b^\dagger(X_i, Y_i).$$

For $S_k^2(a) = |I_k|^{-1} \sum_{i \in I_k} (a_i - \bar{a}_k)^2$,

$$S_k^2(b^\dagger + d) - S_k^2(b^\dagger) = 2C_{P,k,n} + D_{P,k,n},$$

where

$$D_{P,k,n} = S_k^2(d) \leq \frac{1}{|I_k|} \sum_{i \in I_k} d_{k,i}^2,$$

$$|C_{P,k,n}| \leq \{S_k^2(b^\dagger) D_{P,k,n}\}^{1/2}.$$

Conditional on $\mathcal{T}_k$, the held-out observations are independent of $\widehat{b}_k$, so

$$\mathbb{E}\left[\frac{1}{|I_k|} \sum_{i \in I_k} d_{k,i}^2 \,\middle|\, \mathcal{T}_k \right] = \|\widehat{b}_k - b^\dagger\|_{L_2(P)}^2.$$

For every $\epsilon, \eta > 0$, conditional Markov gives

$$\begin{aligned}\Pr(D_{P,k,n} > \epsilon) &\le \Pr\{\|\widehat{b}_k - b^\dagger\|_{L_2(P)}^2 > \eta\} \\ &\quad + \mathbb{E}\left[\mathbf{1}\{\|\widehat{b}_k - b^\dagger\|_2^2 \le \eta\} \Pr(D_{P,k,n} > \epsilon \mid \mathcal{T}_k)\right] \\ &\le \Pr\{\|\widehat{b}_k - b^\dagger\|_2^2 > \eta\} + \frac{\eta}{\epsilon}.\end{aligned}$$

Because K is fixed and $\Delta_{b,n} = o_p(1)$, first letting $n \to \infty$ and then $\eta \downarrow 0$ gives the fold-weighted $D_{P,k,n} = o_p(1)$. Consequently $C_{P,k,n} = o_p(1)$ by Cauchy–Schwarz and the fixed-center second-moment bound. Finally,

$$\begin{aligned}\sum_k \pi_k S_k^2(b^\dagger) &= \sum_k \pi_k \left[\mathbb{P}_{k,n}(b^\dagger)^2 - \{\mathbb{P}_{k,n} b^\dagger\}^2\right] \\ &\xrightarrow{p} P(b^\dagger)^2 - (Pb^\dagger)^2 = \sigma_P^2,\end{aligned}$$

by the complete-cluster law of large numbers. Hence $\widehat{\sigma}_P^2 \to_p \sigma_P^2$. Under (A), (E.8) yields the required fold-wise laws of large numbers explicitly. For every fold,

$$\begin{aligned}\mathrm{Var}\left\{\frac{1}{|I_{k,n}|}\sum_{i\in I_{k,n}} (B_{n,i}^\dagger)^2\right\} &= \frac{1}{|I_{k,n}|^2}\sum_{i\in I_{k,n}} \mathrm{Var}\{(B_{n,i}^\dagger)^2\} \\ &\le \frac{C}{|I_{k,n}|} \longrightarrow 0,\end{aligned}$$

and

$$\mathrm{Var}\left(\frac{1}{|I_{k,n}|}\sum_{i\in I_{k,n}} B_{n,i}^\dagger\right) \le \frac{C}{|I_{k,n}|} \longrightarrow 0.$$

Hence

$$S_{k,n}^2(B_n^\dagger) - \frac{1}{|I_{k,n}|}\sum_{i\in I_{k,n}} \mathbb{E}(B_{n,i}^\dagger)^2 \xrightarrow{p} 0.$$

Weighting by $\pi_{k,n} = |I_{k,n}|/n$ and summing over the fixed number of folds gives

$$\sum_k \pi_{k,n} S_{k,n}^2(B_n^\dagger) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}(B_{n,i}^\dagger)^2 \xrightarrow{p} 0.$$

Because the component variables are centered, the second term is $\sigma_{P,n}^2 \to \sigma_P^2$. Combining this fixed-center result with the same conditional-Markov plug-in perturbation above proves $\widehat{\sigma}_P^2 \to_p \sigma_P^2$ under (A).

For the synthetic side, set

$$A_i^{\dagger,\mathrm{unc}} = \frac{1}{m}\sum_{c=1}^m \varphi^\dagger(\widetilde{X}_{i,c}^{\mathrm{eval}}), \qquad D_{k,i} = \widehat{A}_{k,i} - A_i^{\dagger,\mathrm{unc}}.$$

Let $S_k^2(a) = |I_k|^{-1}\sum_{i\in I_k}(a_i - \bar{a}_k)^2$. Then

$$S_k^2(A^\dagger + D) - S_k^2(A^\dagger) = 2C_{Q,k,n} + D_{Q,k,n}, \tag{E.9}$$

where

$$C_{Q,k,n} = \frac{1}{|I_k|}\sum_{i\in I_k}(A_i^{\dagger,\mathrm{unc}} - \bar{A}_k^\dagger)(D_{k,i} - \bar{D}_k),$$

$$D_{Q,k,n} = \frac{1}{|I_k|}\sum_{i\in I_k}(D_{k,i} - \bar{D}_k)^2 \le \frac{1}{|I_k|}\sum_{i\in I_k} D_{k,i}^2.$$

Conditional on $\mathcal{T}_k$, Jensen's inequality within a prompt group gives

$$\begin{aligned}\mathbb{E}(D_{k,i}^2 \mid \mathcal{T}_k) &= \mathbb{E}\left[\left\{\frac{1}{m}\sum_{c=1}^{m}(\widehat{\varphi}_k - \varphi^\dagger)(\widetilde{X}_{i,c}^{\mathrm{eval}})\right\}^2 \middle| \mathcal{T}_k\right] \\ &\le \frac{1}{m}\sum_{c=1}^{m}\mathbb{E}\left[(\widehat{\varphi}_k - \varphi^\dagger)^2(\widetilde{X}_{i,c}^{\mathrm{eval}}) \mid \mathcal{T}_k\right] \\ &= \|\widehat{\varphi}_k - \varphi^\dagger\|_{L_2(Q)}^2.\end{aligned}$$

For any $\epsilon, \eta > 0$, conditional Markov gives

$$\Pr(D_{Q,k,n} > \epsilon) \le \Pr\{\|\widehat{\varphi}_k - \varphi^\dagger\|_2^2 > \eta\} + \frac{\eta}{\epsilon}.$$

Because $\Delta_{\varphi,n} = o_p(1)$ and K is fixed, $D_{Q,k,n} = o_p(1)$ fold-weightedly. By Cauchy–Schwarz,

$$|C_{Q,k,n}| \le \{S_k^2(A^\dagger)D_{Q,k,n}\}^{1/2} = o_p(1),$$

since the fixed-center group means have finite second moment. Hence

$$\widehat{\sigma}_Q^2 = \sum_k \pi_k S_k^2(A^\dagger) + o_p(1).$$

Under (E), the complete clusters are i.i.d., so the ordinary fold-centered law of large numbers gives

$$\sum_k \pi_k S_k^2(A^\dagger) \xrightarrow{p} \mathrm{Var}(A_Q^\dagger) = \sigma_Q^2.$$

Under (A), the same calculation used for the real component gives, for each fold,

$$\mathrm{Var}\left\{\frac{1}{|I_{k,n}|}\sum_{i\in I_{k,n}}(A_{Q,n,i}^\dagger)^2\right\} \le \frac{C}{|I_{k,n}|} \longrightarrow 0,$$

$$\mathrm{Var}\left(\frac{1}{|I_{k,n}|}\sum_{i\in I_{k,n}}A_{Q,n,i}^\dagger\right) \le \frac{C}{|I_{k,n}|} \longrightarrow 0.$$

Consequently,

$$\sum_k \pi_{k,n} S_{k,n}^2(A_{Q,n}^\dagger) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(A_{Q,n,i}^\dagger)^2 \xrightarrow{p} 0.$$

The second term is $\sigma_{Q,n}^2 \to \sigma_Q^2$. Together with the plug-in perturbation bound above, this proves $\widehat{\sigma}_Q^2 \to_p \sigma_Q^2$ under (A). Therefore the stated componentwise variance consistency holds under either regime. □

E.6 Proof of Theorem 4

Proof of Theorem 4. Under regime (E), Theorem 3 gives

$$\widehat{\mu} - \mu = \Gamma_{n,\dagger} + o_p(n^{-1/2}).$$

Under regime (A), Theorem 3A gives

$$\widehat{\mu} - \mu_n = \Gamma_{n,\dagger}^{(A)} + o_p(n^{-1/2}).$$

Proposition E.2 gives the corresponding additive fixed-center CLT and componentwise variance consistency for the leading term in either regime. Since $\sigma_P^2 + \sigma_Q^2 > 0$, the continuous mapping theorem and Slutsky's theorem yield the standardized normal limit and the corresponding Wald coverage statement. □

Corollary E.1 (Coverage robustness to residual–synthetic covariance)**.** *Assume the fixed-center expansion, the complete-cluster CLT and moment conditions, the fixed-K, positive-fold-proportion design, and the componentwise variance consistency in Proposition E.2. For the row-n fixed-center contributions, define $V_{\mathrm{add},n} = \sigma_{P,n}^2 + \sigma_{Q,n}^2$, $C_{\perp,n} = n^{-1}\sum_{i=1}^{n} \mathrm{Cov}(B_{n,i}^{\dagger}, A_{Q,n,i}^{\dagger})$, and $\rho_{\perp,n} = 2C_{\perp,n}/V_{\mathrm{add},n}$, and suppose $V_{\mathrm{add},n} \to V_{\mathrm{add}} > 0$. If the relative residual–synthetic covariance contribution $\rho_{\perp,n} \to \rho > -1$, the componentwise consistency gives $\widehat{\sigma}_{\mathrm{tot}}^2/V_{\mathrm{add},n} \xrightarrow{p} 1$, and the additive interval has limiting coverage*

$$2\Phi\left\{\frac{z_{1-\alpha/2}}{\sqrt{1+\rho}}\right\} - 1,$$

and hence recovers nominal coverage when $\rho = 0$.

E.7 Proof of Corollary E.1

Proof of Corollary E.1. Write $V_{\mathrm{full},n} = V_{\mathrm{add},n} + 2C_{\perp,n}$ and

$$T_n = \frac{\sqrt{n}(\widehat{\mu} - \mu_n)}{\sqrt{V_{\mathrm{full},n}}}.$$

Then

$$\begin{aligned}\frac{\sqrt{n}(\widehat{\mu} - \mu_n)}{\widehat{\sigma}_{\mathrm{tot}}} &= T_n \left(\frac{V_{\mathrm{full},n}}{V_{\mathrm{add},n}}\right)^{1/2} \left(\frac{V_{\mathrm{add},n}}{\widehat{\sigma}_{\mathrm{tot}}^2}\right)^{1/2} \\ &= T_n\sqrt{1+\rho_{\perp,n}} \left(\frac{V_{\mathrm{add},n}}{\widehat{\sigma}_{\mathrm{tot}}^2}\right)^{1/2}.\end{aligned}$$

By the assumed central limit theorem, variance consistency, and Slutsky's theorem,

$$\frac{\sqrt{n}(\widehat{\mu} - \mu_n)}{\widehat{\sigma}_{\mathrm{tot}}} \xrightarrow{d} \sqrt{1+\rho}\, Z, \qquad Z \sim N(0,1).$$

Therefore,

$$\begin{aligned}\Pr\{\mu_n \in \widehat{C}_{n,\alpha}\} &= \Pr\left\{\left|\frac{\sqrt{n}(\widehat{\mu} - \mu_n)}{\widehat{\sigma}_{\mathrm{tot}}}\right| \le z_{1-\alpha/2}\right\} \\ &\longrightarrow \Pr\left\{|Z| \le \frac{z_{1-\alpha/2}}{\sqrt{1+\rho}}\right\} \\ &= 2\Phi\left\{\frac{z_{1-\alpha/2}}{\sqrt{1+\rho}}\right\} - 1.\end{aligned}$$

The nominal-coverage conclusion follows by setting $\rho = 0$. □

Corollary E.2 (Asymptotic interval-width ratio)**.** *Under the fixed-law exact covariance-control regime of Theorem 4, suppose $\mathrm{Var}(Y) \in (0,\infty)$, and let $\widehat{C}_{n,\alpha}^{\mathrm{label}}$ be the classical label-only Wald interval with a consistent variance estimator. Writing $\mathrm{len}(C)$ for interval width, for fixed $0 < \alpha < 1$,*

$$\frac{\mathrm{len}(\widehat{C}_{n,\alpha})}{\mathrm{len}(\widehat{C}_{n,\alpha}^{\mathrm{label}})} \xrightarrow{p} \left\{\frac{\sigma_P^2 + \sigma_Q^2}{\mathrm{Var}(Y)}\right\}^{1/2}.$$

Hence MaSI is asymptotically shorter exactly when $\mathrm{Var}(Y) - \sigma_P^2 > \sigma_Q^2$: the prediction component explains more outcome variance than the auxiliary group statistic contributes.

E.8 Proof of Corollary E.2

Proof. Because $h^{\dagger}(Z^{\dagger}) = \mathbb{E}(Y \mid Z^{\dagger})$,

$$\mathrm{Var}(Y) = \mathbb{E}\{\mathrm{Var}(Y \mid Z^{\dagger})\} + \mathrm{Var}\{h^{\dagger}(Z^{\dagger})\}.$$

Moreover,

$$\sigma_P^2 = \mathrm{Var}\{Y - h^{\dagger}(Z^{\dagger})\} = \mathbb{E}\{\mathrm{Var}(Y \mid Z^{\dagger})\}.$$

Therefore

$$\mathrm{Var}(Y) - (\sigma_P^2 + \sigma_Q^2) = \mathrm{Var}\{h^{\dagger}(Z^{\dagger})\} - \sigma_Q^2 = \mathcal{G}_{\mathrm{MA}}.$$

The interval-width ratio follows by dividing the asymptotic standard deviations. □

F Supplementary Experimental Designs and Results

This appendix provides the designs, diagnostics, and complete results summarized in Section 5. The estimator, variance estimator, and theoretical derivations remain in the main text and Appendices A–E. For results-first reading, Section F.5 collects the complete simulation, controlled-text, and MHS comparisons; Sections F.2–F.4 provide representation, optimization, transport, covariance, and theory-facing diagnostics.

F.1 Experimental Designs and Comparators

All experiments follow the same statistical convention. The target is defined by the real labels of the real observational units, while generated texts enter only as source-conditioned, model-derived auxiliary information. Q_{style} is used for content-space learning and Q_{eval} for final inference. Texts attached to the same real unit form one prompt group, so generated records do not increase the number of independent units. All methods share the target, real resamples, and outer-fold splits within a setting. Generated inputs are matched only at the level explicitly stated: representation comparisons reuse the same available Q_{eval} texts, and generation-count-matched arms fix the total number of generations but may change their prompt source. Among the generation-count-matched comparator and channel-role arms, only the MHS same-pool shuffled-role ablation reuses the exact same $5 + 5$ text pool.

F.1.1 Common Experimental Setup and Reproducibility

In the primary simulations and empirical studies, all representation learning, outcome prediction, density-ratio estimation, and tuning decisions are performed within the corresponding outer-training folds. The separately labeled style-dominant representation stress test is the sole exception: to reproduce the reported diagnostic, its constraint thresholds are fixed by a simulation-only DGP pilot, while its representation, outcome learner, density ratio, and final estimator remain fitted from the observed training folds. Held-out folds are used only for final estimation and evaluation. Density ratios are estimated by prior-corrected logistic classification of real versus evaluation texts; source-group-preserving inner folds choose finite-sample stabilization by balancing content-coordinate and fitted-score moments. The simulation study draws a fresh i.i.d. sample from the specified data-generating process in each Monte Carlo replication. In the controlled text and MHS experiments, repeated samples of real observational units are drawn with replacement from the corresponding empirical evaluation population, so no finite-population correction is applied.

The main simulation uses $K = 5$ outer folds, $m = 10$ generated covariates per channel and real observational unit, $n = 200$, and $R = 300$ Monte Carlo replications; the sample-size sensitivity experiments use $R = 200$ replications. The controlled text and MHS experiments use $K = 5$ outer folds, five Q_{style} texts and five Q_{eval} texts per real unit, $n \in \{100, 200, 400, 600, 800\}$, and $R = 200$ replications. Texts are generated through the SiliconFlow DeepSeek-V3 endpoint (`deepseek-ai/DeepSeek-V3`) (DeepSeek-AI et al., 2024) and embedded by `Qwen/Qwen3-Embedding-8B` (Zhang et al., 2025) into 1024-dimensional vectors. The controlled renderer uses temperature 0.6, top-p 0.9, and a 220-token cap; the MHS style and evaluation channels use temperatures 0.35 and 0.55 and top-p values 0.85 and 0.9, respectively, with five prespecified variants per channel. The anonymous release provides the synthetic simulation implementation, fixed paper-scale configurations, automated tests, reproduction commands, expected aggregate summaries, and the audited controlled-rendering and MHS channel prompt templates. This appendix records the model identifiers, request settings, generation roles, evaluation populations, and complete reported results for the controlled-rendering and MHS analyses; the remaining releasable analysis code will be made public upon acceptance, subject to dataset licenses and model-service terms.

Table F.1: Comparator definitions and matching status. For the primary experiments, learned components and tuning choices are fitted within each outer-training fold. Unless a row is explicitly labeled as a negative control, uncertainty is estimated across real prompt groups. Generation-count matching fixes the number of generated texts; exact reuse of the same text pool is stated separately.

Method	Definition and role
Classical	Uses the mean and sample variance of the real labels; it is the real-data reference.
Oracle MaSI	Replaces the learned projector with the latent content projector and otherwise applies Equation (2) and the Wald interval in Section 3.2; it is an unavailable oracle-representation benchmark.
MaSI	Learns a constrained content map from real texts and Q_{style} rewrites, freezes it before evaluating Q_{eval}, forms the density-ratio-weighted auxiliary term, and applies the real-label residual correction in Equation (2). The primary MHS design uses five texts per channel and Equation (3).
Raw PPI	Applies unweighted PPI residual correction to Q_{eval}, using identity coordinates in simulation and standardized raw embeddings in the text experiments.
PPI (PCA-20)	Applies the same prompt-group PPI correction after fold-wise PCA to 20 dimensions; it is the compact-representation comparator.
PPI (PCA-5)	Applies prompt-group PPI after fold-wise PCA to five dimensions; it matches the rank used by MaSI.
PPI++	This generated-text adaptation applies training-fold correction-multiplier selection to source-averaged Q_{eval} predictions. It evaluates correction-strength tuning under source-conditioned generation rather than the original target-unlabeled sampling regime. The generation-count-matched MHS arm uses ten Q_{eval} texts.
IPW-PPI	Adds fold-local logistic real-to-synthetic density-ratio weights to the PPI correction. The generation-count-matched MHS arm uses ten Q_{eval} texts.
Single-channel evaluation split	Splits ten Q_{eval} texts into five for representation learning and five for inference, then applies the same downstream MaSI correction without Q_{style}.
Same-pool shuffled-role assignment	Randomly reassigns five texts from each channel between representation learning and inference, then applies the same downstream MaSI correction. It is the only channel ablation that reuses the exact primary $5 + 5$ text pool.
MaSI (5 style + 10 evaluation texts)	Keeps five Q_{style} texts and increases Q_{eval} from five to ten. This is a 15-generation sensitivity analysis, not a matched comparison.
FPPI	Retains the three Q_{eval} variants with the smallest absolute training-fold prediction drift before applying the PPI correction.
PPI (independent-text variance; negative control)	Uses the same point estimator as the corresponding raw or PCA-20 PPI method, but incorrectly treats same-prompt generations as independent in the variance calculation.
Synthetic-label controls	Replace generated covariates with LLM-generated outcomes and use either direct averaging or a real-label block correction. They are negative controls for the role of real-label anchoring.

F.1.2 Simulation Data-generating Mechanisms

The simulation experiments generate real observational units, Q_{style} paired rewrites, and Q_{eval} synthetic covariates in a fixed representation space. Let $O_\star = [V_\star, B_\star] \in \mathbb{R}^{20\times 20}$ be a fixed orthogonal basis generated once, with $V_\star \in \mathbb{R}^{20\times 5}$ spanning the content space and $B_\star \in \mathbb{R}^{20\times 15}$ spanning its orthogonal complement. For each real unit, $C_i \sim N(\mu_c, \Sigma_c)$, $\xi_i \sim N(0, \Sigma_s)$, and

$S_i = HC_i + \xi_i$. We then set $U_i = V_\star C_i + B_\star S_i + e_i$, where $e_i \sim N(0, 0.05^2 I_{20})$. We set $\mu_c = (0.5, -0.3, 0.2, 0.1, -0.2)^\top$, $\Sigma_c = \mathrm{diag}(1, 0.8, 0.6, 0.5, 0.4)$, and $\beta = (1, 0.8, 0.6, 0.4, 0.2)^\top$. The continuous linear outcome is $Y_i = 1 + \beta^\top C_i + \alpha_s q_\star^\top S_i + \varepsilon_i$, with $\varepsilon_i \sim N(0, 0.4^2)$. The continuous nonlinear outcome is $Y_i = 1 + \beta^\top C_i + \alpha_s q_\star^\top S_i + g(C_i) + \varepsilon_i$, using the same noise law and the centered component

$$g(C_i) = 0.5C_{i1}^2 + 0.4\sin(C_{i2}) + 0.3C_{i1}C_{i3} - \mathbb{E}\{0.5C_{i1}^2 + 0.4\sin(C_{i2}) + 0.3C_{i1}C_{i3}\}.$$

For the binary and ordinal outcomes, let $L_i = \beta^\top(C_i - \mu_c) + \varepsilon_i$. The binary outcome is $Y_i = \mathbb{1}\{L_i > 0\}$, and the ordinal outcome takes values $0, \ldots, 5$ according to thresholds $\tau_j = s_L \Phi^{-1}(j/6)$, $j = 1, \ldots, 5$, where $s_L = (\beta^\top \Sigma_c \beta + 0.4^2)^{1/2} = 1.409$. Thus the numerical thresholds are approximately $(-1.363, -0.607, 0, 0.607, 1.363)$. For both continuous responses, the target is $\mu = \mathbb{E}(Y_i)$, with Gaussian regression noise. The binary and ordinal responses are bounded, but the Gaussian embedding and style draws make ΔU unbounded in every simulation. Accordingly, the bounded-envelope concentration in Theorem 1 is not invoked literally for these Gaussian-coordinate runs; they are finite-sample robustness checks for the same mechanism. Boundedness is one sufficient route to the uniform concentration step, not a claim that a realized finite sample by itself satisfies a bounded envelope.

Q_{style} keeps $\widetilde{C}_{i,c} = C_i$ and changes style coordinates by prompt-specific shifts and Gaussian style noise. Q_{eval} varies the outcome-relevant content direction in whitened content coordinates. Write $W_i = \Sigma_c^{-1/2}(C_i - \mu_c)$ and $v_y = \Sigma_c^{1/2}\beta / \|\Sigma_c^{1/2}\beta\|_2$. For the correctly directed setting, $v_{\mathrm{eval}} = v_y$; for the wrong-direction negative control, v_{eval} is a unit vector orthogonal to v_y. In the phase-map sensitivity, the alignment parameter $\kappa \in [0, 1]$ is defined through $v_{\mathrm{eval}} = \kappa v_y + \sqrt{1 - \kappa^2}\, v_\perp$, where $v_\perp$ is a fixed unit direction orthogonal to v_y; hence $\kappa = 1$ is aligned and $\kappa = 0$ is orthogonal. With $P_{\mathrm{eval}} = v_{\mathrm{eval}} v_{\mathrm{eval}}^\top$, the evaluation content draw is

$$\widetilde{W}_{i,c} = (I - P_{\mathrm{eval}})W_i + \rho P_{\mathrm{eval}} W_i + \sqrt{1 - \rho^2}\, P_{\mathrm{eval}} \zeta_{i,c}, \qquad \zeta_{i,c} \sim N(0, I_5),$$

and $\widetilde{C}_{i,c} = \mu_c + \Sigma_c^{1/2}\widetilde{W}_{i,c}$. Smaller ρ corresponds to stronger outcome-relevant variation. The sensitivity analyses compare correctly directed variation, wrongly directed variation, style-only rewriting, different values of ρ, different generation counts, and different sample sizes.

Table F.2: Numerical parameters of the simulation data-generating mechanisms.

Quantity	Value or construction
Representation and content dimensions	$d = 20, r = 5$
Orthogonal basis	One random orthogonal matrix $O_\star = [V_\star, B_\star]$, drawn independently of the Monte Carlo samples and then held fixed across replications
Content mean and covariance	$\mu_c = (0.5, -0.3, 0.2, 0.1, -0.2)^\top$, $\Sigma_c = \mathrm{diag}(1, 0.8, 0.6, 0.5, 0.4)$
Style loading and covariance	$H = 0_{15\times 5}$, $\Sigma_s = \mathrm{diag}(1, 1, 0.8, 0.8, 0.6, \ldots, 0.6)$
Outcome and style parameters	$\beta = (1, 0.8, 0.6, 0.4, 0.2)^\top$, $q_\star = e_1 \in \mathbb{R}^{15}$, $\alpha_s = 0$ in the reported main mechanisms
Real embedding noise	$e_i \sim N(0, 0.05^2 I_{20})$
Outcome noise	$\varepsilon_i \sim N(0, 0.4^2)$
Q_{style} content and style law	$\widetilde{C}_{i,c} = C_i$, $\widetilde{S}_{i,c} = a_c + 0.3S_i + \eta_{i,c}$, $\eta_{i,c} \sim N(0, 0.6^2 I_{15})$; the shifts a_c are centered, normalized prompt-shift vectors with norm 0.8
Q_{eval} content law	$\widetilde{W}_{i,c} = (I - P_{\mathrm{eval}})W_i + \rho P_{\mathrm{eval}} W_i + \sqrt{1-\rho^2} P_{\mathrm{eval}} \zeta_{i,c}$, followed by $\widetilde{C}_{i,c} = \mu_c + \Sigma_c^{1/2}\widetilde{W}_{i,c}$
Q_{eval} style law	Same style-shift and style-noise law as Q_{style}
Main retention parameters	$\rho \in \{0.5, 0.7\}$

Continued on the next page

Table F.2 continued

Quantity	Value or construction
Sensitivity retention parameters	$\rho \in \{1.0, 0.9, 0.7, 0.5, 0.3\}$
Generation counts	The main simulation uses $m = 10$ per channel; generation-count sensitivity uses $m \in \{1, 3, 5, 10, 20\}$ per channel
Outcome learners	Ridge, second-order ridge, logistic ridge, and multinomial logistic regression for linear, nonlinear, binary, and ordinal outcomes, respectively

The controlled LLM text experiment lies between pure simulation and real data. It first constructs a latent population of product reviews, and then lets an LLM render text from product-experience facts. These facts are treated as S_i^{src} in Section 2; the baseline review is X_i^{text}, and the theoretical covariate for the generation law is $X_i = (X_i^{\mathrm{text}}, S_i^{\mathrm{src}})$. In this experiment S_i^{src} is available to the renderer as generation-side context but is hidden from the encoder and from all representation, prediction, and ratio fits; those fits use the rendered text X_i^{text} only. The outcome Y_{soft} is generated by the latent model as a probability and is evaluated on $[0, 1]$, so no outcome clipping is needed; fitted predictions are kept in the same interval to keep the prediction and loss classes bounded. The LLM does not access labels, the target mean, or evaluation metrics. The final evaluation population contains 1000 LLM-rendered baseline text units, each with one baseline review, five Q_{style} style rewrites, and five Q_{eval} evaluation texts. The generation log contains 11000 parsed texts with no missing or failed outputs. The PPI-like baselines use the same real samples and generated texts, while MaSI additionally uses Q_{style} paired rewrites to learn the content space.

The Measuring Hate Speech experiment treats a prespecified set of $N_{\mathrm{emp}} = 2{,}000$ comments from the cleaned corpus as a fixed empirical evaluation population (Sachdeva et al., 2022). The comments are selected before text generation, model fitting, and inspection of inferential results using a prespecified stratified eligibility design that samples records across severity and ambiguity strata among comments with nonmissing text and outcome values. Let $P_{\mathrm{emp}} = N_{\mathrm{emp}}^{-1} \sum_{j=1}^{N_{\mathrm{emp}}} \delta_{(X_j, Y_j)}$. The outcome is the continuous hate-speech score, and the target is its empirical-population mean $\mu_{\mathrm{emp}} = \mathbb{E}_{P_{\mathrm{emp}}}(Y) = -0.97038$. Since P_{emp} has finite support, $Y \in [a_{\mathrm{emp}}, b_{\mathrm{emp}}]$ P_{emp}-a.s., where $a_{\mathrm{emp}} = \min_{1 \le j \le N_{\mathrm{emp}}} Y_j$ and $b_{\mathrm{emp}} = \max_{1 \le j \le N_{\mathrm{emp}}} Y_j$; no outcome clipping is used. For each Monte Carlo replication and labeled sample size n, real units are sampled independently with replacement from P_{emp}. Thus, the reported coverage, bias, and RMSE evaluate repeated-sampling performance relative to this fixed empirical target, and no finite-population correction is required. The Q_{style} prompt instructs the renderer to preserve the target, stance, hostility level, identity reference, and ambiguity of a comment while changing surface expression; the diagnostics assess the resulting empirical stability. Q_{eval} generates evaluation-oriented comments around the severity of the original comment and is used for the synthetic correction term. The prompt templates use only the input text and do not send labels, annotation-derived variables, or numeric scores to the LLM. The generated-text PPI++ adaptation tunes the prediction-correction strength under source-conditioned generation (Angelopoulos et al., 2023b); it is not an evaluation of PPI++ in its original target-unlabeled sampling regime. IPW-PPI estimates the real–synthetic density ratio and applies importance weighting (Shimodaira, 2000; Huang et al., 2007; Sugiyama et al., 2008; Cortes et al., 2010); following FPPI's selective-use principle, our FPPI baseline retains synthetic predictions deemed reliable from training-side information (Xu & Sun, 2026). Negative controls include the Q_{style} single-channel design, ordinary zero-shot/few-shot generation, independent-text variance aggregation, and direct synthetic labels. They probe whether apparent shortening survives calibration, real-label correction, and within-prompt dependence accounting.

F.2 Simulation: Applicability and Mechanism Diagnostics

All quantities below use MaSI as specified in Sections 3–4. The main audit has $n = 200$, $K = 5$, $m = 10$ per channel, four outcome types, two values of ρ, and 300 replications per setting. Prompt groups are complete by construction, and every prespecified generation is retained without quality-score, outcome, or inferential filtering. Completeness is therefore a design fact rather than a performance metric.

For replication b, let $\widehat{v}_{Y,b}$ be the sample variance of its real labels and let $\widehat{\sigma}^2_{\mathrm{tot},b} = \widehat{\sigma}^2_{P,b} + \widehat{\sigma}^2_{Q,b}$. The reported finite-sample gain fraction is $\widehat{G}_b = (\widehat{v}_{Y,b} - \widehat{\sigma}^2_{\mathrm{tot},b})/\widehat{v}_{Y,b}$. Its mean is reported as a diagnostic of Corollary E.2; it is not used to choose a method, discard a generation, or modify the primary interval.

For the simulation oracle basis $V_\star$ and its unit outcome direction v_y, define fold-wise outcome-direction retention as $c_{Y,k} = \|\widehat{V}_k^\top V_\star v_y\|_2^2$. Table F.3 reports its mean over the fitted folds and replications.

Table F.3: Aggregate applicability diagnostics in the main simulation. Setting-level quantities are averaged equally over the eight outcome–ρ combinations; replication-level quantities use all 2,400 MaSI fits. The fitted additive-variance reduction is the empirical analogue of the normalized gain $\mathcal{G}_{\mathrm{MA}}/\operatorname{Var}(Y)$. Here $\widehat{\rho}_{\perp,n} = 2\widehat{C}_\perp/(\widehat{\sigma}^2_P + \widehat{\sigma}^2_Q)$, where $\widehat{C}_\perp$ is the fold-aggregated empirical residual–synthetic covariance; this diagnostic does not alter the additive variance estimator.

Diagnostic	Result
Mean coverage [range across settings]	0.9125 [0.9000, 0.9233]
Mean width ratio	0.7355
Fitted additive-variance reduction, mean	45.20%
Replications with positive fitted reduction	100.0%
Density-ratio relative ESS, mean [minimum]	0.99986 [0.99184]
Maximum clipping fraction	0
$\widehat{\rho}_{\perp,n}$, mean	-0.0390
$\lvert\widehat{\rho}_{\perp,n}\rvert$, 95th percentile	0.1031
Outcome-direction retention, mean	0.9816

The spectral audit repeats the main mechanism over 1,500 fold fits. The empirical eigengap has mean 0.5179 and 5th, median, and 95th percentiles 0.4881, 0.5185, and 0.5448. The median upper-to-lower boundary eigenvalue ratio is 171.75. The initializer's projector distance to the simulation oracle $\Pi^\star$ has mean 0.0399 and 95th percentile 0.0457. This establishes an empirically localized start relative to the oracle benchmark in this design; Figure F.1 separately audits localization relative to the pseudo-true center $\Pi^\dagger$.

Figure 2 reports 30 prespecified confirmation runs from a separately locked centered-linear audit. The displayed path minimizes robust-scaled distance to the multivariate median of five run summaries: projector-distance reduction, population-risk reduction, initializer-to-$\Pi^\dagger$ distance, the fraction of distance-decreasing updates, and the largest positive constraint margin. Bands show interquartile ranges across all runs, and trajectory appearance is not used for selection.

Table F.4: Thirty-run centered-linear representation-learning audit underlying Figure 2. Projector distance is measured to the unavailable simulation oracle $\Pi^\star$; excess risk is population risk above the oracle risk. The last column reports the across-run interquartile interval $[Q_1, Q_3]$.

Quantity	Mean	Median	Interquartile interval
Initial projector distance	0.1235	0.1236	[0.1145, 0.1290]
Final projector distance	0.0664	0.0660	[0.0601, 0.0710]
Projector-distance reduction	0.4588	0.4607	[0.4095, 0.5077]
Initial excess risk	0.01586	0.01543	[0.01347, 0.01724]
Final excess risk	0.00300	0.00277	[0.00215, 0.00357]
Excess-risk reduction	0.8056	0.8211	[0.7450, 0.8521]
Paths with $\max_q [g_q]_+ \le 0.03$		30/30	

Table F.5: Fold-level constraint audit under the main optimizer. Each outcome uses 30 replications and five folds at $\rho = 0.5$. The table reports the rewrite-stability margin g_0 alongside the maximum positive constraint violation; the last column gives the fraction with $\max_q [g_q]_+ \leq 0.03$.

Outcome	Median g_0	95th percentile $\max_q [g_q]_+$	Fraction within 0.03
Linear	0.0044	0.0101	1.000
Nonlinear	0.0226	0.0333	0.867
Binary	0.0259	0.0398	0.687
Ordinal	0.0416	0.0541	0.060

Table F.6: Generation-count sensitivity, averaged over linear and ordinal outcomes and $\rho \in \{0.5, 0.7\}$, with 200 replications per setting. Coverage ranges show the four individual settings; interval reduction is relative to Classical.

m	Mean coverage	Coverage range	Mean interval reduction
1	0.921	[0.915, 0.930]	-2.6%
3	0.905	[0.880, 0.925]	19.6%
5	0.896	[0.885, 0.910]	24.9%
10	0.911	[0.890, 0.925]	28.6%
20	0.900	[0.890, 0.905]	30.8%

Table F.7: Sample-size diagnostics averaged over $\rho \in \{0.5, 0.7\}$, with 200 replications per outcome–ρ setting. The absolute relative covariance is $|\widehat{\rho}_{\perp,n}|$; feasibility uses the fixed 0.03 tolerance. For each individual outcome–ρ setting, the approximate 95% Monte Carlo half-width near 0.90 coverage is 0.042; the displayed ρ-averaged coverages have smaller Monte Carlo error.

	Linear			Ordinal		
n	Coverage	Mean $\lvert\widehat{\rho}_{\perp,n}\rvert$	Feasibility rate	Coverage	Mean $\lvert\widehat{\rho}_{\perp,n}\rvert$	Feasibility rate
100	0.905	0.0534	1.000	0.913	0.0652	0.000
200	0.923	0.0257	1.000	0.900	0.0398	0.015
400	0.883	0.0144	1.000	0.885	0.0203	0.995
800	0.883	0.0097	1.000	0.915	0.0122	1.000
1600	0.903	0.0079	1.000	0.913	0.0081	1.000
3200	0.905	0.0061	1.000	0.910	0.0062	1.000

Tables F.5 and F.7 distinguish a finite-sample optimization boundary from inferential failure. At $n = 200$, the ordinal learner often exceeds the fixed rewrite-stability tolerance even though coverage is close to nominal. Feasibility rises to 0.995 by $n = 400$ and 1.000 thereafter, while the absolute covariance contribution decreases for both outcomes. We therefore use the centered-linear audit for the strongest mechanism claim and the outcome-wise results to define when that interpretation is empirically supported.

F.3 Simulation: Scope and Theory-facing Sensitivities

The following figures isolate two questions that the aggregate comparisons cannot answer: whether the fixed-center theoretical chain appears in an exact structural regime, and whether interval shortening disappears when generated content is weakly aligned with the outcome.

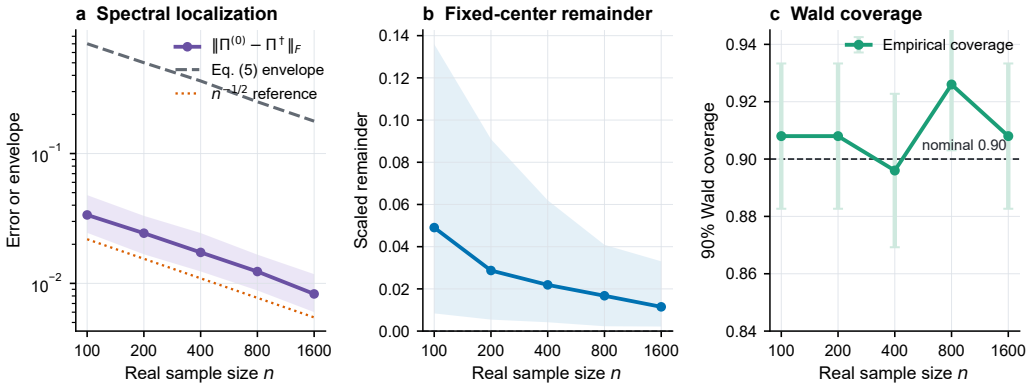

Figure F.1: Theory-aligned finite-sample diagnostics in the exact structural regime. Panel (a) tracks spectral localization relative to $\Pi^\dagger$ and its $n^{-1/2}$ reference on the $n \in \{100, 200, 400, 800, 1600\}$ grid; panel (b) tracks the square-root-n-scaled fixed-center expansion remainder in Theorem 3 (the detailed bound is Eq. (D.15) in Appendix D); panel (c) reports empirical coverage of the 90% Wald interval against the nominal 0.90 level. The diagnostic uses $m = 3$, five folds, and 500 Monte Carlo replications per sample size; panel (a) uses a log–log scale. Shading in panels (a) and (b) denotes the 10th–90th percentiles, and coverage bars in panel (c) are 95% Monte Carlo intervals. This benchmark is a special case of regime (E), with $w^\dagger \equiv 1$ and $\mathrm{Cov}(B^\dagger, A_Q^\dagger) = 0$; it does not audit the ALM or learned-ratio/tangent-rate conditions.

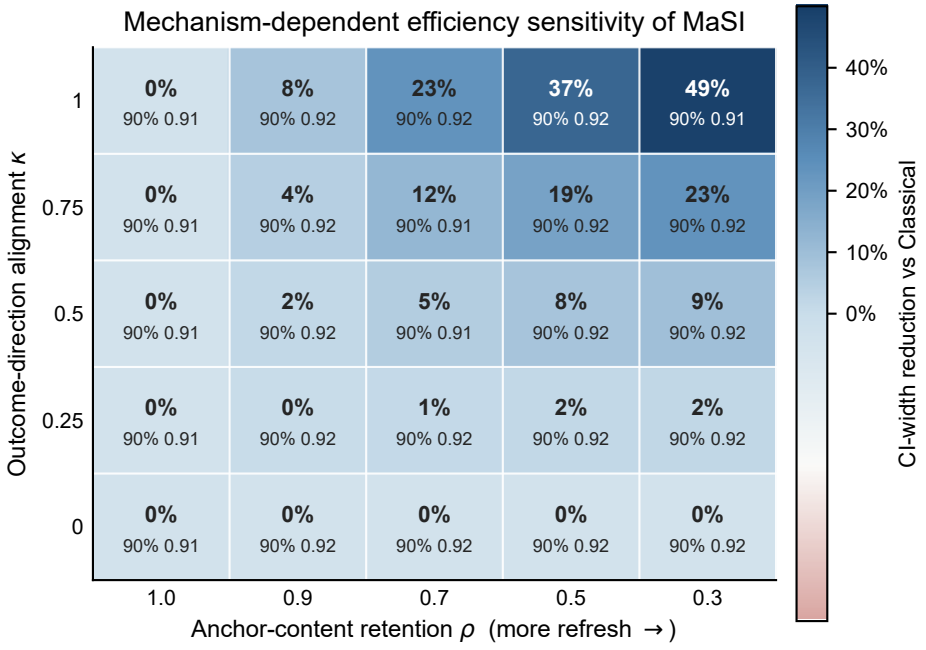

Figure F.2: Supplementary interval-width sensitivity map for the linear simulation. The map varies the anchor-content retention parameter ρ and the outcome-direction alignment parameter κ, defined in the simulation protocol by $v_{\mathrm{eval}} = \kappa v_y + \sqrt{1-\kappa^2}\, v_\perp$. Each cell reports the interval-width reduction relative to Classical and the empirical 90% coverage for the $n = 200$, five-fold, $m = 10$ setting with 300 Monte Carlo replications. This is a sensitivity analysis of the variance-gain condition, not an additional coverage guarantee.

F.4 MHS: Inferential and Channel Diagnostics

All diagnostics use the same outer-fold splits and held-out real units as the primary MHS analysis; where a diagnostic changes the generation mechanism or fitting target, it uses a dedicated training-fold refit. Generated texts attached to the same real unit are first averaged within their Q_{eval} prompt group; no generated text is treated as an independent observation. The analysis uses the frozen rendered texts and embeddings and contains 200 Monte Carlo replications at each labeled sample size.

For $i \in I_k$, recall from Appendix E that

$$\widehat{B}_{k,i} = \widehat{b}_k(X_i, Y_i), \qquad \widehat{A}_{k,i} = \frac{1}{m}\sum_{c=1}^{m} \widehat{\varphi}_k(\widetilde{X}_{i,c}^{\mathrm{eval}}).$$

Let

$$\bar{B}_k = \frac{1}{|I_k|}\sum_{i\in I_k} \widehat{B}_{k,i}, \qquad \bar{A}_k = \frac{1}{|I_k|}\sum_{i\in I_k} \widehat{A}_{k,i}.$$

The empirical residual–synthetic cross-covariance is reported only as an approximate-orthogonality diagnostic:

$$\widehat{C}_{\perp,k} = \frac{1}{|I_k|}\sum_{i\in I_k} (\widehat{B}_{k,i} - \bar{B}_k)(\widehat{A}_{k,i} - \bar{A}_k), \qquad \widehat{C}_{\perp} = \sum_{k=1}^{K} \pi_k \widehat{C}_{\perp,k}.$$

We report its relative contribution to the additive variance,

$$\widehat{\rho}_{\perp,n} = \frac{2\widehat{C}_{\perp}}{\widehat{\sigma}_P^2 + \widehat{\sigma}_Q^2}.$$

This diagnostic is not added to $\widehat{\sigma}^2_{\mathrm{tot}}$, and the reported primary interval remains the Wald interval in Section 3.2. The separately tabulated source-block interval is a covariance-inclusive sensitivity and is not used for data-dependent interval selection.

To assess the population transport term in equation (D.3), define the real-text prediction $\widehat{H}_{k,i} = \widehat{h}_k\{\widehat{Z}_k(X_i)\}$ and the held-out fold discrepancy

$$\widehat{\Delta}_{w,k,n} = \frac{1}{|I_k|}\sum_{i\in I_k} \{\widehat{A}_{k,i} - \widehat{H}_{k,i}\}.$$

Its fold-weighted absolute empirical counterpart is

$$\widehat{\beta}^{\mathrm{diag}}_{w,n} = \sum_{k=1}^{K} \pi_k |\widehat{\Delta}_{w,k,n}|,$$

and we report the standardized scale

$$\widehat{R}_{\beta,n} = \frac{\sqrt{n}\,\widehat{\beta}^{\mathrm{diag}}_{w,n}}{\widehat{\sigma}_{\mathrm{tot}}}.$$

Even under exact population transport, this absolute held-out discrepancy contains empirical-process noise on the displayed root-n scale. It is therefore a comparative finite-sample diagnostic, not a consistent test that the population remainder in (D.3) is $o(n^{-1/2})$. We also report the relative density-ratio effective sample size,

$$\widehat{\mathrm{ESS}}_k^{\mathrm{rel}} = \frac{\left(\sum_{i,c} \widehat{w}_{k,i,c}\right)^2}{m|I_k| \sum_{i,c} \widehat{w}_{k,i,c}^2},$$

and the fraction of fitted weights at the clipping boundary.

To audit the density-ratio tangent term in Theorem 3, let $w_{k,\gamma}$ denote the fold-k density-ratio model. On the held-out prompt groups, define

$$\widehat{\mathbf{g}}_{w,k} = \frac{1}{|I_k|m}\sum_{i\in I_k}\sum_{c=1}^{m} \nabla_\gamma w_{k,\widehat{\gamma}_k}\{\widehat{Z}_k(\widetilde{X}_{i,c}^{\mathrm{eval}})\}\widehat{h}_k\{\widehat{Z}_k(\widetilde{X}_{i,c}^{\mathrm{eval}})\}.$$

We estimate the ratio-parameter covariance by cluster bootstrap refits over complete real-unit ratio-training clusters. With $B = 20$ bootstrap refits,

$$\overline{\gamma}_k^* = \frac{1}{B}\sum_{b=1}^{B}\widehat{\gamma}_k^{*(b)}, \qquad \widehat{\Sigma}_{\gamma,k} = \frac{n_k^{\mathrm{tr}}}{B-1}\sum_{b=1}^{B}(\widehat{\gamma}_k^{*(b)} - \overline{\gamma}_k^*)(\widehat{\gamma}_k^{*(b)} - \overline{\gamma}_k^*)^\top,$$

where n_k^{tr} is the number of complete real-unit clusters used to fit the fold-k ratio learner. The fold-level standardized tangent impact is

$$\widehat{R}_{\chi,k,n} = \sqrt{\frac{n}{n_k^{\mathrm{tr}}}}\frac{\left(\widehat{\mathbf{g}}_{w,k}^\top \widehat{\Sigma}_{\gamma,k}\widehat{\mathbf{g}}_{w,k}\right)^{1/2}}{\widehat{\sigma}_{\mathrm{tot}}},$$

and the aggregate diagnostic is

$$\widehat{R}_{\chi,n} = \left\{\sum_{k=1}^{K}\pi_k \widehat{R}_{\chi,k,n}^2\right\}^{1/2}.$$

This is a finite-sample diagnostic for the standard-deviation scale of the linearized ratio-parameter contribution relative to the leading MaSI standard deviation. It is not used as a proof that the tangent-rate condition holds automatically.

For MaSI, the empirical relative covariance diagnostic $\widehat{\rho}_{\perp,n}$ decreases from 0.113 at $n = 100$ to 0.055 at $n = 800$. It is reported to audit the approximate-orthogonality condition and is not added to the primary variance estimator or used for data-dependent interval selection. If these values were treated as fixed population contributions, the formula in Corollary E.1 would map a nominal 90% additive interval to limiting coverages 0.881 and 0.891, respectively. This sensitivity calculation is not an alternative interval and the observed decline does not establish a vanishing population covariance; direct repeated-sampling coverage remains the relevant finite-sample MHS evidence.

The prediction-targeted diagnostic displays a different pattern across generation mechanisms. For MaSI, $\widehat{R}_{\beta,n}$ remains between 0.735 and 0.803, the relative ESS remains between 0.993 and 0.999, and no fitted weight reaches the clipping boundary. The ordinary zero-shot and few-shot generation controls exhibit substantially larger standardized discrepancies, ranging from 1.67 to 3.30, together with lower relative ESS and positive clipping fractions. These finite-sample comparisons are consistent with the mechanism-designed Q_{eval} channel being better aligned with the prediction direction entering the MaSI estimator; they do not verify the population rate required by Theorem 3.

The ratio-tangent diagnostic separately assesses ratio-parameter estimation. For MaSI, $\widehat{R}_{\chi,n}$ takes values 0.058, 0.100, 0.121, 0.129, and 0.134 for $n = 100, 200, 400, 600, 800$, respectively, with average 0.108 and maximum 0.134. The corresponding relative ESS is above 0.993 and the clipping fraction is zero. Ordinary zero-shot and few-shot generation have average tangent-impact ratios 0.389 and 0.385, with maxima 0.524 and 0.527, respectively, together with lower ESS and nonzero clipping. Thus, the implemented density-ratio learner has smaller finite-sample tangent coupling under the mechanism-designed Q_{eval} channel than under the ordinary single-channel controls. This relative diagnostic does not establish that the tangent term vanishes asymptotically.

We additionally examine whether the two generation channels display their intended empirical roles. For Q_{style}, source-classification AUC is evaluated both in the raw encoder representation and in the learned content representation. Style-variant classification is evaluated in the remaining expression representation. Prediction stability is summarized by

$$\widehat{\Delta}_{\mathrm{style}}^h = \frac{1}{Nm}\sum_{i=1}^{N}\sum_{c=1}^{m}\left[\widehat{h}\{\widehat{Z}(\widetilde{X}_{i,c}^{\mathrm{style}})\} - \widehat{h}\{\widehat{Z}(X_i)\}\right].$$

For Q_{eval}, a complete synthetic prompt group is either retained with its original real observational unit or reassigned to another unit. This permutation preserves the synthetic-text pool but breaks the unit-specific correspondence. Let

$$\Delta R_{\mathrm{aug}}^2 = R_{\mathrm{aug,same}}^2 - R_{\mathrm{aug,perm}}^2, \qquad \Delta R_{\mathrm{res}}^2 = R_{\mathrm{res,same}}^2 - R_{\mathrm{res,perm}}^2.$$

Positive values indicate that the same-unit prompt group retains source-conditioned predictive association that is lost when the unit correspondence is broken.

In the MHS experiment, real and Q_{style} texts are distinguishable in the raw embedding space ($\text{AUC}_{\text{raw}} = 0.643$), but source discrimination falls to chance level in the learned content representation ($\text{AUC}_c = 0.495$–0.498). The residual expression representation retains the five style variants with accuracy 0.603, relative to chance 0.200, while the absolute prediction drift is at most 0.022. The same residual representation has a held-out outcome-prediction R^2 of 0.0749. Thus, outcome information in the residual coordinates is limited rather than completely eliminated. Combined with the substantially above-chance style-variant accuracy, this pattern is consistent with a nondegenerate residual representation that retains expression variation while restricting outcome-related signal. It also matches the approximate, rather than exact, residual–synthetic orthogonality formulation used in the theory. For Q_{eval}, both ΔR^2_{aug} and ΔR^2_{res} are positive for every sample-size–learner combination. The former ranges from 0.041 to 0.121, and the latter from 0.048 to 0.218. These results are consistent with Q_{style} supplying content-stable expression variation and Q_{eval} supplying real-unit-specific predictive variation.

Table F.8: Theory-facing diagnostics for MaSI in the Measuring Hate Speech experiment. The reported interval uses $\widehat{\sigma}^2_{\text{tot}} = \widehat{\sigma}^2_P + \widehat{\sigma}^2_Q$; $\widehat{\rho}_{\perp,n}$ is shown only as an approximate-orthogonality diagnostic. $\widehat{R}_{\beta,n}$ is the standardized held-out prediction-targeted discrepancy, $\widehat{R}_{\chi,n}$ is the standardized ratio-tangent impact, and ESS is the relative density-ratio effective sample size. These are comparative finite-sample diagnostics, not tests of the asymptotic rate conditions.

Panel A: interval and covariance diagnostics

n	$\widehat{\rho}_{\perp,n}$	Additive coverage	Additive width ratio
100	0.1132	0.905	0.9245
200	0.0815	0.890	0.9279
400	0.0685	0.905	0.9306
600	0.0582	0.880	0.9342
800	0.0549	0.905	0.9348

Panel B: transport and ratio diagnostics

n	$\widehat{\beta}^{\text{diag}}_{w,n}$	$\widehat{R}_{\beta,n}$	$\widehat{R}_{\chi,n}$	Relative ESS
100	0.1615	0.7354	0.0577	0.9986
200	0.1260	0.8033	0.0998	0.9964
400	0.0855	0.7664	0.1214	0.9950
600	0.0676	0.7396	0.1289	0.9938
800	0.0591	0.7474	0.1336	0.9931

Table F.9: Average MHS theory-facing diagnostics over the five labeled sample sizes. The width ratio is the arithmetic mean of the five sample-size-specific ratios. These diagnostics use dedicated refits; primary performance is reported in Figure 3 and Appendix Table F.17. The relative cross-covariance is shown only as an approximate-orthogonality diagnostic.

Panel A: interval and covariance diagnostics

Generation mechanism	Additive coverage	Additive width ratio	$\overline{\widehat{\rho}}_{\perp}$
MaSI Q_{eval}	0.897	0.9304	0.0753
Q_{style} single-channel	0.899	0.9730	0.0826
Ordinary zero-shot	0.883	0.9392	0.0788
Ordinary few-shot	0.862	0.9439	0.0793

Panel B: transport and overlap diagnostics

Generation mechanism	$\overline{\widehat{R}}_{\beta}$	Relative ESS	Clipping fraction
MaSI Q_{eval}	0.7584	0.9954	0.0000
Q_{style} single-channel	0.2968	0.9991	0.0000
Ordinary zero-shot	2.5877	0.7670	0.0450
Ordinary few-shot	2.4547	0.7985	0.0235

Table F.10: Ratio-tangent diagnostic for the MHS generation mechanisms. $\widehat{R}_{\chi,n}$ measures the standardized finite-sample tangent impact of ratio-parameter estimation relative to the leading MaSI standard deviation. The average and maximum are taken across the five labeled sample sizes.

Generation mechanism	Average $\widehat{R}_{\chi,n}$	Maximum $\widehat{R}_{\chi,n}$
MaSI Q_{eval}	0.108	0.134
Q_{style} single-channel	0.026	0.047
Ordinary zero-shot	0.389	0.524
Ordinary few-shot	0.385	0.527

Table F.11: Dual-channel mechanism diagnostics in the MHS experiment. Panel A assesses content stability under Q_{style}; Panel B, continued below, compares same-unit and permuted-unit Q_{eval} prompt groups.

Panel A: Q_{style} content stability

Learner	$\mathrm{AUC}_{\mathrm{raw}}$	AUC_c	Acc_s	Chance for Acc_s	$\lvert\widehat{\Delta}^h_{\mathrm{style}}\rvert$	Held-out $R^2(Y \mid Z_s)$
Ridge	0.6432	0.4981	0.6033	0.2000	0.0141	0.0749
Extra Trees	0.6432	0.4945	0.6033	0.2000	0.0222	0.0749

Table F.11 continued: Panel B, Q_{eval} same-unit versus permuted-unit groups

n	Learner	ΔR^2_{aug}	ΔR^2_{res}
100	Extra Trees	0.0579	0.0479
100	Ridge	0.1211	0.1790
200	Extra Trees	0.0445	0.0520
200	Ridge	0.0999	0.2181
400	Extra Trees	0.0472	0.0620
400	Ridge	0.0515	0.1241
600	Extra Trees	0.0533	0.0797
600	Ridge	0.0430	0.1090
800	Extra Trees	0.0607	0.0938
800	Ridge	0.0407	0.1049

F.5 Primary Performance and Mechanism Comparisons

The tables below retain all sample sizes and comparator arms. Main-text claims use setting-wise width ratios; ratio-of-mean-width summaries, where shown, are secondary because they depend on the outcome scale.

F.5.1 Simulation: Performance and Channel Mechanisms

Table F.12: MaSI results in the sample-size scaling experiment, grouped by outcome type and ρ. Coverage is the empirical coverage of the 90% confidence interval; $\sqrt{n}$-scaled width checks whether interval width follows $n^{-1/2}$ scaling.

Outcome type	ρ	n	Coverage	Mean width	$\sqrt{n}\times$ mean width
Continuous linear	0.5	100	0.900	0.298	2.982
Continuous linear	0.5	200	0.925	0.206	2.918
Continuous linear	0.5	400	0.875	0.145	2.891
Continuous linear	0.5	800	0.880	0.102	2.877
Continuous linear	0.5	1600	0.905	0.072	2.871
Continuous linear	0.5	3200	0.905	0.051	2.862
Continuous linear	0.7	100	0.910	0.363	3.626
Continuous linear	0.7	200	0.920	0.252	3.569
Continuous linear	0.7	400	0.890	0.178	3.551
Continuous linear	0.7	800	0.885	0.125	3.541
Continuous linear	0.7	1600	0.900	0.088	3.537
Continuous linear	0.7	3200	0.905	0.062	3.527
Ordinal rating	0.5	100	0.910	0.396	3.959
Ordinal rating	0.5	200	0.890	0.264	3.729
Ordinal rating	0.5	400	0.885	0.181	3.612
Ordinal rating	0.5	800	0.905	0.126	3.565
Ordinal rating	0.5	1600	0.905	0.089	3.543
Ordinal rating	0.5	3200	0.910	0.062	3.529
Ordinal rating	0.7	100	0.915	0.463	4.629
Ordinal rating	0.7	200	0.910	0.314	4.444
Ordinal rating	0.7	400	0.885	0.217	4.348
Ordinal rating	0.7	800	0.925	0.152	4.306
Ordinal rating	0.7	1600	0.920	0.107	4.287
Ordinal rating	0.7	3200	0.910	0.076	4.272

Table F.13: PPI-like method comparison and style-dominant representation stress test. The reported width ratio is the unweighted mean of the eight outcome–ρ ratios used in the main text. The stress test evaluates relative robustness under strong expression interference. Its constraint thresholds are fixed by a simulation-only DGP pilot; the learned MaSI projection and outcome learner remain estimated from training data, and this row is not used for the primary coverage claim. Oracle MaSI uses the unobserved true content direction. The independent-text row is an invalid variance negative control; in the main simulation it shares the Raw PPI point estimator and differs only in variance aggregation.

Setting	Method	Coverage	Mean setting-wise width ratio
Main simulation average	Classical	0.903	1.000
Main simulation average	Oracle MaSI	0.918	0.697
Main simulation average	MaSI	0.912	0.735
Main simulation average	Raw PPI	0.916	0.784
Main simulation average	PPI (PCA-5)	0.913	0.761
Main simulation average	PPI (independent-text variance)	0.782	0.585
Style-dominant	Classical	0.877	1.000
Style-dominant	Oracle MaSI	0.890	0.749
Style-dominant	MaSI	0.880	0.764
Style-dominant	Raw PPI	0.870	0.978
Style-dominant	PPI (PCA-5)	0.860	0.991
Style-dominant	PPI (independent-text variance)	0.870	0.976

Table F.14: Setting-wise comparison supporting the main-text outcome analysis. Each entry reports empirical coverage and width relative to Classical for the same outcome and ρ, using 300 paired replications at $n = 200$.

		MaSI		PPI PCA-5		PPI raw	
Outcome	ρ	Coverage	Width ratio	Coverage	Width ratio	Coverage	Width ratio
Linear	0.5	0.920	0.630	0.937	0.685	0.940	0.630
Linear	0.7	0.923	0.771	0.920	0.785	0.927	0.771
Nonlinear	0.5	0.903	0.624	0.930	0.664	0.913	0.732
Nonlinear	0.7	0.900	0.767	0.900	0.771	0.907	0.833
Binary	0.5	0.913	0.776	0.887	0.790	0.910	0.825
Binary	0.7	0.923	0.860	0.890	0.861	0.917	0.903
Ordinal	0.5	0.903	0.662	0.920	0.718	0.900	0.734
Ordinal	0.7	0.913	0.790	0.920	0.814	0.913	0.847

Table F.15: Setting-wise comparison between the dual-channel and single-channel generation designs. The dual-channel design uses Q_{style} for representation learning and Q_{eval} for inference; the single-channel design uses one synthetic mechanism for both roles. Width ratio is relative to Classical. This is a mechanism-sensitivity comparison, not a same-text-pool or generation-count-matched ablation.

Outcome type	ρ	Generation design	Coverage	Width ratio
Binary	0.5	Dual-channel design	0.913	0.776
Binary	0.5	Single-generation design	0.923	1.034
Binary	0.7	Dual-channel design	0.923	0.860
Binary	0.7	Single-generation design	0.907	0.871
Continuous linear	0.5	Dual-channel design	0.920	0.630
Continuous linear	0.5	Single-generation design	0.923	0.728
Continuous linear	0.7	Dual-channel design	0.923	0.771
Continuous linear	0.7	Single-generation design	0.923	0.771
Continuous nonlinear	0.5	Dual-channel design	0.903	0.624
Continuous nonlinear	0.5	Single-generation design	0.933	0.710
Continuous nonlinear	0.7	Dual-channel design	0.900	0.767
Continuous nonlinear	0.7	Single-generation design	0.910	0.773
Ordinal rating	0.5	Dual-channel design	0.903	0.662
Ordinal rating	0.5	Single-generation design	0.943	1.015
Ordinal rating	0.7	Dual-channel design	0.913	0.790
Ordinal rating	0.7	Single-generation design	0.910	0.796

The full results explain the aggregate gain. At $\rho = 0.7$, dual and single channels are nearly indistinguishable; at $\rho = 0.5$, separation shortens intervals for all four outcomes while dual-channel coverage remains 0.903–0.920. The style-dominant stress test supplies a different challenge: when expression interference is strong, MaSI retains a much shorter interval than raw or PCA-5 PPI. Thus neither nonlinearity nor channel separation alone guarantees improvement; the decisive feature is how the generation mechanism interacts with outcome-relevant structure.

F.5.2 CONTROLLED TEXT: COMPLETE RESULTS

Table F.16: Complete method comparison by sample size in the controlled text experiment. Coverage is for the 90% interval and width ratio is relative to Classical. The independent-text variance negative control shares the PCA-20 PPI point estimator and differs only by treating same-prompt generations as independent when estimating variance.

n	Method	Coverage	Mean width	Width ratio
100	Classical	0.920	0.129	1.000
100	MaSI	0.905	0.119	0.922
100	PPI (PCA-20)	0.910	0.124	0.959
100	PPI (independent-text variance)	0.850	0.109	0.845
100	PPI (PCA-5)	0.915	0.115	0.890
100	Raw PPI	0.870	0.122	0.941
200	Classical	0.910	0.092	1.000
200	MaSI	0.895	0.081	0.887
200	PPI (PCA-20)	0.900	0.082	0.889
200	PPI (independent-text variance)	0.855	0.071	0.773
200	PPI (PCA-5)	0.885	0.080	0.874
200	Raw PPI	0.875	0.086	0.939
400	Classical	0.955	0.065	1.000
400	MaSI	0.890	0.057	0.874
400	PPI (PCA-20)	0.915	0.056	0.866
400	PPI (independent-text variance)	0.850	0.048	0.748
400	PPI (PCA-5)	0.910	0.056	0.866
400	Raw PPI	0.855	0.061	0.938
600	Classical	0.910	0.053	1.000
600	MaSI	0.885	0.046	0.870
600	PPI (PCA-20)	0.885	0.045	0.856
600	PPI (independent-text variance)	0.825	0.039	0.736
600	PPI (PCA-5)	0.880	0.046	0.861
600	Raw PPI	0.845	0.049	0.926
800	Classical	0.875	0.046	1.000
800	MaSI	0.890	0.040	0.867
800	PPI (PCA-20)	0.885	0.039	0.852
800	PPI (independent-text variance)	0.810	0.034	0.732
800	PPI (PCA-5)	0.865	0.039	0.859
800	Raw PPI	0.830	0.042	0.924

Because the target is known while the covariates are LLM-rendered texts, this experiment separates textual realism from target uncertainty. Compact PCA representations already yield near-nominal coverage and competitive widths, so MaSI and PCA-based PPI have similar coverage–width performance. The corresponding independent-text variance negative control is narrower at every n but under-covers, showing that the benign representation regime does not remove within-prompt dependence.

F.5.3 MHS: PRIMARY AND MECHANISM COMPARISONS

Table F.17: Complete method comparison by sample size in the Measuring Hate Speech empirical-population experiment. Coverage is for the 90% interval and width ratio is relative to Classical. The independent-text variance negative control shares the PCA-20 PPI point estimator and differs only in variance aggregation; it is not a valid efficiency baseline.

n	Method	Coverage	Mean width	Width ratio
100	Classical	0.935	0.785	1.000
100	MaSI	0.905	0.731	0.931
100	PPI (PCA-20)	0.780	0.738	0.940
100	PPI (independent-text variance)	0.675	0.589	0.750
100	PPI (PCA-5)	0.775	0.719	0.916
100	Raw PPI	0.785	0.772	0.984
200	Classical	0.900	0.557	1.000
200	MaSI	0.890	0.518	0.930
200	PPI (PCA-20)	0.650	0.531	0.953
200	PPI (independent-text variance)	0.505	0.410	0.736
200	PPI (PCA-5)	0.600	0.520	0.934
200	Raw PPI	0.580	0.587	1.054
400	Classical	0.925	0.395	1.000
400	MaSI	0.905	0.368	0.932
400	PPI (PCA-20)	0.485	0.379	0.960
400	PPI (independent-text variance)	0.315	0.286	0.725
400	PPI (PCA-5)	0.335	0.373	0.946
400	Raw PPI	0.305	0.449	1.139
600	Classical	0.915	0.322	1.000
600	MaSI	0.880	0.301	0.935
600	PPI (PCA-20)	0.325	0.310	0.963
600	PPI (independent-text variance)	0.175	0.232	0.720
600	PPI (PCA-5)	0.185	0.306	0.951
600	Raw PPI	0.125	0.387	1.201
800	Classical	0.910	0.279	1.000
800	MaSI	0.905	0.261	0.935
800	PPI (PCA-20)	0.235	0.269	0.965
800	PPI (independent-text variance)	0.120	0.200	0.717
800	PPI (PCA-5)	0.100	0.266	0.954
800	Raw PPI	0.080	0.357	1.280

Unlike the controlled renderer, the MHS results are consistent with persistent representation and transport error: raw and PCA PPI intervals contract while coverage deteriorates with n. MaSI retains coverage near the nominal level over the same grid with a stable 6.5–7.0% width reduction, motivating the mechanism and matched comparisons below.

Representation-learning ablation on MHS. For the three learned-representation variants, we keep the downstream calibrated estimator fixed and vary only representation learning. Spectral initialization only uses $V_k^{(0)}$ without subsequent representation optimization; outcome-guided learning without constraints starts from $V_k^{(0)}$ and minimizes $\widehat{\mathcal{R}}_{y,k}^{\mathrm{tr}}$ without $g_0, \ldots, g_3$; MaSI uses the constrained representation fit in Equation (1). Classical is the real-label reference. Because the constraint set is removed jointly, the comparison does not attribute the coverage difference to any single g_q or to C_{BQ}.

Table F.18: Replication-matched representation-learning ablation on Measuring Hate Speech. Each row summarizes 200 replications. Width ratios are relative to Classical at the same n, bias is signed, and RMSE is computed against the fixed empirical-population target. Average rows are unweighted averages over the five sample sizes.

n	Method	Coverage	Width ratio	Bias	RMSE
100	Classical	0.935	1.000	+0.0029	0.2200
100	Spectral initialization only	0.925	0.996	+0.0025	0.2219
100	Outcome-guided learning without constraints	0.870	0.931	+0.0030	0.2344
100	MaSI	0.905	0.931	-0.0122	0.2220
200	Classical	0.900	1.000	+0.0050	0.1609
200	Spectral initialization only	0.900	0.993	+0.0052	0.1605
200	Outcome-guided learning without constraints	0.885	0.931	-0.0106	0.1719
200	MaSI	0.890	0.930	-0.0111	0.1606
400	Classical	0.925	1.000	+0.0083	0.1139
400	Spectral initialization only	0.920	0.994	+0.0075	0.1139
400	Outcome-guided learning without constraints	0.880	0.928	-0.0092	0.1191
400	MaSI	0.905	0.932	-0.0030	0.1132
600	Classical	0.915	1.000	+0.0045	0.0946
600	Spectral initialization only	0.915	0.993	+0.0033	0.0948
600	Outcome-guided learning without constraints	0.835	0.929	-0.0149	0.1027
600	MaSI	0.880	0.935	-0.0031	0.0946
800	Classical	0.910	1.000	+0.0034	0.0790
800	Spectral initialization only	0.915	0.994	+0.0025	0.0787
800	Outcome-guided learning without constraints	0.870	0.928	-0.0097	0.0864
800	MaSI	0.905	0.935	-0.0035	0.0784
Average	Classical	0.917	1.000	+0.0048	0.1337
Average	Spectral initialization only	0.915	0.994	+0.0042	0.1339
Average	Outcome-guided learning without constraints	0.868	0.929	-0.0083	0.1429
Average	MaSI	0.897	0.933	-0.0066	0.1338

The ablation separates spectral initialization from subsequent representation learning without changing the downstream estimator. Spectral initialization only behaves like Classical; outcome-guided learning without constraints supplies most of the shortening but has lower coverage; and MaSI retains comparable shortening with a more favorable coverage–width balance. Because all constraints change together, this supports the joint roles of initialization, outcome-guided learning, and simultaneous constraint control rather than any single margin.

Table F.19: Comparison using five evaluation texts per real unit among residual-corrected PPI methods, the Q_{style} single-channel control, and primary MaSI (five texts per channel) on Measuring Hate Speech. Classical uses the real-unit sample variance; model-assisted methods use their stated prompt-group variance estimators. Width ratios are relative to Classical. This table does not match the total generation count across methods.

n	Method	Coverage	Mean width	Width ratio	Bias
100	Classical	0.935	0.785	1.000	0.0029
100	MaSI	0.905	0.731	0.931	-0.0122
100	Q_{style} single-channel	0.915	0.752	0.958	0.0009
100	PPI (PCA-20)	0.780	0.738	0.940	-0.1879
100	PPI++	0.790	0.728	0.927	-0.1545
100	IPW-PPI	0.910	0.736	0.937	0.0006
100	FPPI	0.890	0.746	0.950	-0.0581
200	Classical	0.900	0.557	1.000	0.0050
200	MaSI	0.890	0.518	0.930	-0.0111
200	Q_{style} single-channel	0.875	0.541	0.971	0.0040
200	PPI (PCA-20)	0.650	0.531	0.953	-0.1968
200	PPI++	0.700	0.515	0.925	-0.1649
200	IPW-PPI	0.885	0.540	0.969	0.0204
200	FPPI	0.880	0.538	0.965	-0.0564
400	Classical	0.925	0.395	1.000	0.0083
400	MaSI	0.905	0.368	0.932	-0.0030
400	Q_{style} single-channel	0.905	0.387	0.980	0.0089
400	PPI (PCA-20)	0.485	0.379	0.960	-0.1892
400	PPI++	0.570	0.361	0.915	-0.1559
400	IPW-PPI	0.900	0.393	0.997	0.0224
400	FPPI	0.885	0.384	0.973	-0.0457
600	Classical	0.915	0.322	1.000	0.0045
600	MaSI	0.880	0.301	0.935	-0.0031
600	Q_{style} single-channel	0.895	0.316	0.982	0.0052
600	PPI (PCA-20)	0.325	0.310	0.963	-0.1896
600	PPI++	0.435	0.293	0.911	-0.1597
600	IPW-PPI	0.920	0.326	1.013	0.0109
600	FPPI	0.880	0.314	0.976	-0.0466
800	Classical	0.910	0.279	1.000	0.0034
800	MaSI	0.905	0.261	0.935	-0.0035
800	Q_{style} single-channel	0.905	0.274	0.985	0.0037
800	PPI (PCA-20)	0.235	0.269	0.965	-0.1916
800	PPI++	0.310	0.253	0.907	-0.1601
800	IPW-PPI	0.940	0.284	1.018	0.0049
800	FPPI	0.890	0.272	0.978	-0.0484

The five-evaluation-text comparison shows that the main challenge is not merely representation dimension. Ordinary PCA-20 PPI retains bias near -0.19 across n, while IPW-PPI largely removes that bias but sacrifices most of the width reduction. MaSI combines small bias with a stable width gain, whereas the generated-text PPI++ adaptation and FPPI leave residual bias that increasingly dominates their contracting intervals.

Generation-count-matched MHS controls. The primary MHS analysis uses five Q_{style} and five Q_{eval} texts per real unit. The matched controls also use ten generated texts: PPI++ and IPW-PPI use ten Q_{eval} texts, the single-evaluation-channel control splits ten Q_{eval} texts between representation and inference, and the same-pool shuffled-role control reassigns the fixed five-plus-five channel pool to break role specialization. Only the shuffled-role control is a same-text-pool ablation; the other methods match generation count while changing the channel or prompt allocation. All rows use the same target, sample-size grid, outer-fold protocol, and 200 Monte Carlo replications per n. The additive interval is prespecified; the block interval estimates the prompt-group variance of the combined real-residual and synthetic contributions and is reported only as a covariance-inclusive sensitivity.

Table F.20: Comparison at a matched budget of ten generated texts per real unit on Measuring Hate Speech. Coverage and width ratios are reported for the prespecified additive interval and the covariance-inclusive block sensitivity; width ratios are relative to Classical. No data-dependent selection is made between the two variance estimators.

n	Method	Additive coverage	Additive width ratio	Block coverage	Block width ratio	Bias
100	MaSI	0.905	0.931	0.910	0.977	-0.012
100	PPI++	0.790	0.928	0.830	0.988	-0.161
100	IPW-PPI	0.915	0.938	0.920	0.982	+0.004
100	Single-channel split	0.910	0.935	0.930	0.982	-0.014
100	Shuffled roles	0.905	0.944	0.925	0.995	-0.003
200	MaSI	0.890	0.930	0.905	0.965	-0.011
200	PPI++	0.685	0.930	0.730	0.987	-0.170
200	IPW-PPI	0.880	0.969	0.885	1.003	+0.024
200	Single-channel split	0.875	0.940	0.880	0.981	-0.010
200	Shuffled roles	0.885	0.953	0.905	0.997	-0.001
400	MaSI	0.905	0.932	0.910	0.962	-0.003
400	PPI++	0.545	0.926	0.595	0.986	-0.164
400	IPW-PPI	0.895	0.995	0.910	1.026	+0.024
400	Single-channel split	0.890	0.953	0.905	0.984	+0.002
400	Shuffled roles	0.900	0.964	0.915	1.001	+0.005
600	MaSI	0.880	0.935	0.885	0.961	-0.003
600	PPI++	0.415	0.923	0.445	0.986	-0.166
600	IPW-PPI	0.915	1.009	0.925	1.039	+0.013
600	Single-channel split	0.895	0.957	0.905	0.983	+0.000
600	Shuffled roles	0.910	0.969	0.915	1.001	+0.002
800	MaSI	0.905	0.935	0.910	0.960	-0.004
800	PPI++	0.280	0.921	0.300	0.986	-0.168
800	IPW-PPI	0.930	1.013	0.935	1.043	+0.007
800	Single-channel split	0.890	0.960	0.895	0.981	+0.000
800	Shuffled roles	0.910	0.972	0.915	1.002	+0.002

Count matching does not erase the distinction. IPW-PPI attains near-nominal average coverage but a 0.985 width ratio, while the generated-text PPI++ adaptation remains under-covered at a width similar to MaSI. The same-pool shuffled-role control has average coverage 0.902 and width ratio 0.960, versus 0.897/0.933 for MaSI; prespecified channel roles yield a modest width advantage at similar coverage. Because the other arms alter the prompt source as well as allocation, they establish procedural performance at equal generation count rather than isolating a causal channel effect.

Evaluation-draw sensitivity. Table F.21 holds the five style rewrites fixed and increases only the number of evaluation texts from five to ten, using the same frozen Q_{eval} prompt configuration and downstream protocol. The latter arm therefore uses 15, rather than 10, generated texts per real unit and is not part of Table F.20. It is an empirical unequal-allocation sensitivity; no theoretical conclusion relies on extending the equal-m presentation to this arm.

Table F.21: Sensitivity to the number of MHS evaluation texts. The primary MaSI design uses five texts per channel; the sensitivity keeps five style rewrites and uses ten evaluation texts. Add. and Block follow the definitions in Table F.20.

n	Eval. texts per unit	Additive coverage	Additive width ratio	Block coverage	Block width ratio	Bias
100	5	0.905	0.931	0.910	0.977	-0.012
100	10	0.910	0.930	0.920	0.975	-0.014
200	5	0.890	0.930	0.905	0.965	-0.011
200	10	0.890	0.930	0.900	0.965	-0.012
400	5	0.905	0.932	0.910	0.962	-0.003
400	10	0.905	0.931	0.910	0.961	-0.003
600	5	0.880	0.935	0.885	0.961	-0.003
600	10	0.895	0.934	0.905	0.961	-0.003
800	5	0.905	0.935	0.910	0.960	-0.004
800	10	0.905	0.935	0.910	0.960	-0.003
Average	5	0.897	0.933	0.904	0.965	-0.0066
Average	10	0.901	0.932	0.909	0.964	-0.0070

Increasing only Q_{eval} from five to ten texts changes the average MaSI width ratio from 0.933 to 0.932. The small width change is consistent with saturation of within-source averaging in this fixed text pool; it neither separates predictive from Monte Carlo contributions nor identifies an optimal m.

F.5.4 MHS: Ordinary-generation and Synthetic-label Negative Controls

The final controls ask whether prompt specialization and real-label anchoring can be replaced by generic generation or synthetic outcomes. They are deliberately kept outside the primary ranking because they change the information supplied to the estimator.

Table F.22: Sample-size negative control for ordinary zero-shot single-channel generation with PPI-type correction methods on Measuring Hate Speech. Primary dual-channel MaSI is shown as the reference. The other methods use ordinary zero-shot single-channel LLM-generated texts and then apply ordinary PPI, PPI++, IPW-PPI, or FPPI. Width ratio is the mean interval width of the method divided by the mean interval width of Classical.

n	Method	Coverage	Width ratio	Bias
100	MaSI	0.905	0.931	-0.012
100	Ordinary PPI	0.040	0.938	-0.899
100	PPI++	0.240	0.908	-0.651
100	IPW-PPI	0.865	1.082	-0.129
100	FPPI	0.050	0.943	-0.890
200	MaSI	0.890	0.930	-0.011
200	Ordinary PPI	0.000	0.929	-1.054
200	PPI++	0.010	0.885	-0.862
200	IPW-PPI	0.590	1.180	-0.290
200	FPPI	0.000	0.936	-1.045
400	MaSI	0.905	0.932	-0.003
400	Ordinary PPI	0.000	0.922	-1.159
400	PPI++	0.000	0.881	-1.007
400	IPW-PPI	0.130	1.268	-0.412
400	FPPI	0.000	0.930	-1.144
600	MaSI	0.880	0.935	-0.003
600	Ordinary PPI	0.000	0.920	-1.159
600	PPI++	0.000	0.882	-1.030
600	IPW-PPI	0.010	1.317	-0.468
600	FPPI	0.000	0.927	-1.142

Continued on the next page

Table F.22 continued

n	Method	Coverage	Width ratio	Bias
800	MaSI	0.905	0.935	-0.004
800	Ordinary PPI	0.000	0.918	-1.185
800	PPI++	0.000	0.882	-1.054
800	IPW-PPI	0.000	1.329	-0.487
800	FPPI	0.000	0.926	-1.168

Table F.23: Sample-size negative control for ordinary few-shot single-channel generation with PPI-type correction methods on Measuring Hate Speech. Primary dual-channel MaSI is shown as the reference. The other methods use ordinary few-shot single-channel LLM-generated texts and then apply ordinary PPI, PPI++, IPW-PPI, or FPPI. Width ratio is the mean interval width of the method divided by the mean interval width of Classical.

n	Method	Coverage	Width ratio	Bias
100	MaSI	0.905	0.931	-0.012
100	Ordinary PPI	0.120	0.941	-0.697
100	PPI++	0.390	0.912	-0.461
100	IPW-PPI	0.845	1.047	-0.119
100	FPPI	0.130	0.944	-0.690
200	MaSI	0.890	0.930	-0.011
200	Ordinary PPI	0.005	0.933	-0.830
200	PPI++	0.095	0.885	-0.646
200	IPW-PPI	0.630	1.162	-0.247
200	FPPI	0.005	0.937	-0.821
400	MaSI	0.905	0.932	-0.003
400	Ordinary PPI	0.000	0.927	-0.917
400	PPI++	0.005	0.872	-0.755
400	IPW-PPI	0.240	1.243	-0.337
400	FPPI	0.000	0.931	-0.902
600	MaSI	0.880	0.935	-0.003
600	Ordinary PPI	0.000	0.927	-0.919
600	PPI++	0.000	0.871	-0.771
600	IPW-PPI	0.110	1.287	-0.371
600	FPPI	0.000	0.931	-0.902
800	MaSI	0.905	0.935	-0.004
800	Ordinary PPI	0.000	0.926	-0.940
800	PPI++	0.000	0.868	-0.788
800	IPW-PPI	0.030	1.301	-0.391
800	FPPI	0.000	0.930	-0.921

Table F.24: Direct synthetic-label negative control by labeled sample size. Few-shot and zero-shot generation produce synthetic outcome labels conditionally for each real observational unit; global generation produces labeled samples at the population level. Direct pooling uses generated labels without real-label residual correction, whereas real-label correction applies a source-group correction. Width ratio is the mean interval width divided by the mean interval width of Classical.

n	Generation mode	Method	Coverage	Width ratio	Bias
100	Few-shot	Direct pooling	0.575	0.410	0.047
100	Few-shot	Real-label correction	0.925	1.002	0.029
200	Few-shot	Direct pooling	0.500	0.410	0.052
200	Few-shot	Real-label correction	0.890	1.001	0.032

Continued on the next page

Table F.24 continued

n	Generation mode	Method	Coverage	Width ratio	Bias
400	Few-shot	Direct pooling	0.515	0.410	0.053
400	Few-shot	Real-label correction	0.870	1.001	0.035
600	Few-shot	Direct pooling	0.440	0.411	0.050
600	Few-shot	Real-label correction	0.895	1.001	0.031
800	Few-shot	Direct pooling	0.480	0.411	0.049
800	Few-shot	Real-label correction	0.895	1.001	0.030
100	Zero-shot	Direct pooling	0.410	0.443	0.196
100	Zero-shot	Real-label correction	0.885	1.032	0.115
200	Zero-shot	Direct pooling	0.310	0.443	0.198
200	Zero-shot	Real-label correction	0.830	1.031	0.115
400	Zero-shot	Direct pooling	0.140	0.444	0.208
400	Zero-shot	Real-label correction	0.730	1.030	0.120
600	Zero-shot	Direct pooling	0.065	0.444	0.202
600	Zero-shot	Real-label correction	0.710	1.029	0.115
800	Zero-shot	Direct pooling	0.030	0.444	0.200
800	Zero-shot	Real-label correction	0.640	1.029	0.113
100	Global	Direct pooling	0.000	0.592	0.630
100	Global	Real-label correction	0.725	0.576	0.067
200	Global	Direct pooling	0.000	0.590	0.631
200	Global	Real-label correction	0.695	0.569	0.031
400	Global	Direct pooling	0.000	0.589	0.631
400	Global	Real-label correction	0.650	0.564	0.017
600	Global	Direct pooling	0.000	0.590	0.631
600	Global	Real-label correction	0.710	0.564	0.009
800	Global	Direct pooling	0.000	0.590	0.631
800	Global	Real-label correction	0.690	0.563	0.007

Across these controls, generic zero- and few-shot covariates retain substantial transport bias, and direct synthetic-label mixing can produce very short but invalid intervals. Real-label correction removes part of the distortion but generally restores little or none of the precision gain. Together with the MHS diagnostics above, these results support using generated texts as calibrated auxiliary covariates while keeping both the target and uncertainty anchored to real observational units.