CAUSAL JUDGE EVALUATION (CJE): AUDITABLE, UNBIASED OFF-POLICY METRICS FOR LLM SYSTEMS VIA DESIGN-BY-PROJECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal Judge Evaluation (CJE) casts offline "LLM-as-judge" evaluation as calibrated off-policy estimation. We introduce a single design rule—**Design-by-Projection (DbP)**—that encodes justified knowledge as closed convex sets and projects valid objects onto them: (i) reward calibration (**AutoCal-R**, mean-preserving isotonic in judge score S), (ii) weight stabilization (**SIMCal-W**, unit-mean, S-monotone ratios with a light variance cap), (iii) nuisance-orthogonalized estimators (OC-IPS/DR-CPO/TR-CPO), and (iv) variance-optimal **IF-space stacking**. A **Knowledge–Riesz** result shows that intersecting the admissible IF class with justified convex knowledge preserves the estimand and weakly lowers the attainable variance; with cross-fitting, our projection-designed estimators attain the **surrogate information bound**. On Arena logs (n=4,989; five policies), SIMCal-W lifts ESS from near-degenerate regimes (e.g., $0.6\% \to 94.6\%$), calibrated DR regains near-$\sqrt{n}$ scaling with tight, honest CIs, and IF-stacked DR improves ordering. CJE surfaces overlap/tail and judge-calibration diagnostics and, when identification fails, reports rank-robust conclusions (REFUSE-LEVEL).

1 INTRODUCTION

Offline "LLM-as-judge" scores are fast but *correlational*: computed under a logging policy π_0, they do not answer the counterfactual *"What would our KPI be if we deployed π'?"* Off-policy evaluation (OPE) is the right causal tool, yet in practice IPS/SNIPS explode under limited overlap (heavy tails), DR inherits instability from noisy ratios and misspecified nuisances, and judge scores can drift or miscalibrate against higher-fidelity labels (Horvitz & Thompson, 1952; Li et al., 2011; Crump et al., 2009; Jiang & Li, 2016; Lee et al., 2024).

Design-by-Projection (DbP). We encode justified knowledge as a *closed convex set* and *project* valid objects onto it: (i) project scores/influence functions (IFs) onto nuisance-orthogonal *subspaces* (efficiency); (ii) project rewards/weights onto mean-preserving, shape-constrained *cones* (stability; ESS ↑); and (iii) project estimator combinations onto a *simplex* (variance-hedged stacking). Metric projections preserve the estimand (via the mean-one hyperplane) and weakly shrink dispersion through orthogonality and majorization (Bickel et al., 1993; van der Vaart & Wellner, 2000; Barlow et al., 1972; Marshall et al., 2011).

Causal Judge Evaluation (CJE). We instantiate DbP as a practical OPE system. *AutoCal-R* fits a *mean-preserving* calibrator from judge score S to oracle labels on a small i.i.d. slice (isotonic in S with an automatic two-stage single-index fallback). *SIMCal-W* projects baseline ratios onto the cone of *S-monotone, unit-mean* weights using out-of-fold (OOF) stacking and a light variance guard, deterministically lifting ESS. Sequence-aware estimators—*OC-IPS*, *DR-CPO*, and *TR-CPO*—add targeting/retargeting and can be *stacked* by minimizing IF covariance. An oracle-fold jackknife yields *oracle-uncertainty–aware* (OUA) confidence intervals. Brief diagnostics and *gates* surface overlap/tails, judge reliability/coverage, and DR orthogonality; when coverage is poor, CJE returns rank-robust conclusions via REFUSE-LEVEL.

Theory and evidence. We derive the surrogate-model efficient influence function (EIF) and prove a *Knowledge–Riesz (Influence Representer)* result: intersecting the admissible IF class with justified closed convex knowledge *preserves the estimand* and *weakly lowers* the attainable semiparametric variance; with cross-fitting, projection-designed estimators attain the surrogate information bound (Bickel et al., 1993; van der Vaart & Wellner, 2000). Two corollaries guide design: (i) *Blackwell–efficiency monotonicity*—finer judges (larger σ-fields) weakly lower, and generically strictly reduce, the surrogate bound; (ii) SIMCal-W's mean-one isotonic step *Lorenz-dominates* baseline weights, improving every Schur-convex dispersion metric (variance/ESS and beyond). On Arena-derived logs (n=4,989; five policies), SIMCal-W raises ESS from near zero to healthy regimes (e.g., $0.6\% \to 94.6\%$, $0.7\% \to 80.8\%$); calibrated DR regains near-$\sqrt{n}$ scaling with tight, honest CIs; and IF-stacked DR further improves accuracy and ordering. When calibration support is limited, CJE flags REFUSE-LEVEL yet preserves rankings.

Contributions.

1. **DbP for OPE:** a unifying projection rule—subspaces (efficiency), cones (stability), and simplices (hedging)—that preserves the estimand and shrinks variance.
2. **Calibration & stability:** *AutoCal-R* (mean-preserving; automatic two-stage fallback) and *SIMCal-W* (unit-mean, S-monotone) with *deterministic* dispersion/ESS improvement via majorization.
3. **Inference & uncertainty:** sequence-aware DR/TMLE (*OC-IPS*, *DR-CPO*, *TR-CPO*), *IF-Stack* for variance-optimal convex ensembling, and *OUA* CIs that propagate calibration uncertainty.
4. **Knowledge–Riesz & design corollaries:** restricting IFs to justified convex sets *lowers* the variance bound and is *attainable* with cross-fitting; finer judges strictly help (Blackwell monotonicity), and SIMCal-W yields Lorenz-dominant weights (beyond ESS). When identification fails, CJE reports rank-robust, partial-ID conclusions (REFUSE-LEVEL).

2 BACKGROUND AND SETUP

Setup & notation. We observe i.i.d. logs (X_i, A_i, S_i) under a fixed logger $\pi_0(\cdot \mid X)$; $S = s(X, A)$ is a scalar judge score on every row, and a small i.i.d. oracle slice provides labels Y. For a candidate policy π', the sequence-level importance ratio is

$$W_{\pi',i} \;=\; \frac{\pi'(A_i \mid X_i)}{\pi_0(A_i \mid X_i)} \;=\; \exp\big\{\log p_{\pi'}(A_i \mid X_i) - \log p_{\pi_0}(A_i \mid X_i)\big\},$$

computed via teacher forcing (TF). The target is the counterfactual value $V(\pi') = \mathbb{E}[Y(\pi')]$. We use the sample-mean-one normalization (SNIPS) when helpful.

OPE basics. IPS/SNIPS estimate $V(\pi')$ by reweighting logged outcomes (Horvitz & Thompson, 1952; Hájek, 1965; Li et al., 2011; Swaminathan & Joachims, 2015). The direct method (DM) plugs in $g(x) = \sum_a \pi'(a \mid x)\, \hat{m}(x, a)$. Doubly robust (DR) estimators combine IPS and DM and, with sample–splitting and cross–fitting, admit $\sqrt{n}$ inference under the standard one–of–two $n^{-1/4}$ product–rate condition (Bickel et al., 1993; van der Vaart & Wellner, 2000; Kosorok, 2008; Jiang & Li, 2016; Chernozhukov et al., 2018; van der Laan & Rose, 2011). Teacher forcing (TF) provides sequence–level propensities/ratios, so these forms apply to sequence policies without modification (Lee et al., 2024).

Variance, overlap, and stabilization. IPS variance scales with $\mathbb{E}[W_{\pi'}^2]$ and deteriorates under limited overlap (Crump et al., 2009). We monitor stability with the effective sample size (ESS),

$$\mathrm{ESS}(W) = \frac{\big(\sum_i W_i\big)^2}{\sum_i W_i^2}, \qquad \frac{\mathrm{ESS}(W)}{n} = \frac{1}{1 + \mathrm{CV}^2(W)} \text{ when } \bar{W} = 1 \text{ (global mean–one/SNIPS).}$$

We also track tail behavior via diagnostics (Hill, 1975; Liu, 2001; Owen, 2013). Common stabilizers include truncation/clipping (Ionides, 2008), overlap weighting (Li et al., 2018; Fong et al., 2021), balancing objectives (Kallus, 2018), and covariate–shift reweighting (Shimodaira, 2000; Sugiyama et al., 2007).

Calibration for OPE. Calibration enforces identities under π_0: outcome calibration de-biases $g(X)$; ratio calibration enforces $\mathbb{E}_{\pi_0}[W_{\pi'}] = 1$ and $\mathbb{E}_{\pi_0}[W_{\pi'}h] = \mathbb{E}_{\pi'}[h]$ for a test class h; orthogonal moments enable honest inference (Kallus & Mao, 2022; Fong & Kennedy, 2022). Recent work gives projection-based IPS/DR with stability guarantees (van der Laan et al., 2025a;b). For DR, IF orthogonality renders small calibration error second order (Bickel et al., 1993; Chernozhukov et al., 2018; van der Laan & Rose, 2011).

Shape constraints (isotonic). Isotonic regression is the Euclidean projection onto the cone of monotone functions (PAVA) (Ayer et al., 1955; Barlow et al., 1972); it avoids extrapolation and *weakly reduces dispersion* by majorization (Banerjee, 2001; Hardy et al., 1952; Marshall et al., 2011). CJE uses two mean-preserving projections: (i) **AutoCal-R** calibrates $R = f(S)$ on the oracle slice (default: isotonic in S; automatic two-stage *spline→rank→isotonic* fallback), and (ii) **SIMCal-W** maps mean-one ratios onto the cone of S*-monotone, unit-mean* weights (optionally intersected with box/Lipschitz constraints), which *deterministically* lifts ESS by majorization. A light variance guard can blend to a cap before re-projection.

Judges as surrogates. Automatic judges (LLM-as-judge or preference models) provide scalable scoring (Ouyang et al., 2022; Bai et al., 2022; Zheng et al., 2023; Kim et al., 2024; Kocmi & Federmann, 2023) but are correlational and may drift (Wang et al., 2023; Liu et al., 2023). Viewing S as a *surrogate* connects to surrogate endpoints and mediation (Prentice, 1989; Robins & Greenland, 1992; Frangakis & Rubin, 2002; Pearl, 2012; VanderWeele, 2015). Under *mean sufficiency* ($\mathbb{E}[Y \mid X, A, S] = \mu(S)$), calibrating $R = f(S)$ preserves $V(\pi') = \mathbb{E}[f(S^{\pi'})]$ and supplies a one-dimensional index that stabilizes weights.

OUA uncertainty & IF stacking. Treating learned $R = \hat{f}(S)$ as fixed understates uncertainty; we add a calibration component via a delete-one-oracle-fold jackknife on top of the main IF variance (consistent; vanishes as the slice grows) (Bickel et al., 1993; Künsch, 1989; Politis & Romano, 1994). Many OPE estimators are regular and asymptotically linear with per-row IFs $\phi^{(e)}$; we *stack* them by minimizing the plug-in IF covariance over the simplex, preserving regularity and supporting caps/guards, with an optional outer split (Wolpert, 1992; Breiman, 1996; van der Laan et al., 2007).

3 METHODS

CJE follows one rule—*Design-by-Projection (DbP)*—applied to each object in the pipeline: (i) calibrate the reward (projection onto a monotone cone), (ii) stabilize ratios (projection onto a unit-mean, S-monotone cone), (iii) compute an orthogonalized estimator (projection onto a nuisance-orthogonal subspace), and (iv) optionally hedge variance by stacking (projection onto a simplex). All learners are *cross-fitted*; by *Knowledge–Riesz*, these projections preserve the estimand and attain the surrogate information bound (see Section 4).

3.1 REWARD CALIBRATION (AUTOCAL-R: ISOTONIC IN S WITH AN AUTOMATIC TWO-STAGE FALLBACK)

On the oracle slice $\{(S_i, Y_i)\}$, fit a *mean-preserving* calibrator $R = f(T(S))$ with K-fold cross-fitting:

- **Monotone mode (default).** Isotonic regression on S: $\hat{f}_{\uparrow} \in \arg\min_{f \in \mathcal{M}_{\uparrow}} \sum_{i \in O} \big(Y_i - f(S_i)\big)^2$. PAVA preserves the slice mean exactly.
- **Two-stage mode (automatic fallback).** Fit a smooth index $T(S) = g(S)$ (splines+ridge), map to mid-ranks $U = \mathrm{ECDF}\{T(S)\}$, then fit isotonic $\hat{h}_{\uparrow}(U)$. Predictions are $R = \hat{h}_{\uparrow}(\mathrm{ECDF}\{g(S)\})$.

Select the mode by OOF RMSE with a one-standard-error (1-SE) preference for monotone; low/mid/high-S diagnostics are logged. Let R^{OOF} denote OOF predictions used along the IF path; the point estimate may use the pooled fit. The terminal isotonic step makes AutoCal-R *mean-honest* in either mode.

3.2 WEIGHT CALIBRATION (SIMCAL-W: UNIT-MEAN, S-MONOTONE RATIOS WITH OOF STACKING)

Let $W_{\pi'}^{\mathrm{m1}}$ be the sample-mean-one baseline (SNIPS). For each fold k:

1. **Monotone projections (train on $I_{\neg k}$).** Fit increasing/decreasing isotonic maps on S (the latter via $-S$), rescale each to mean one on $I_{\neg k}$, and predict OOF candidates on I_k: $W_{\uparrow}^{\mathrm{OOF}}$, $W_{\downarrow}^{\mathrm{OOF}}$; include the identity candidate $W_{\mathrm{base}}^{\mathrm{OOF}} \equiv 1$.
2. **OOF stacking (variance-aware).** Define residuals T_i used by the downstream estimator: $T_i{=}R_i$ for IPS and $T_i{=}R_i - \hat{m}(X_i, A_i)$ for DR (with $\hat{m}{=}\hat{q}$ below). Let $U_c = W_c^{\mathrm{OOF}} T$ for $c \in \{\mathrm{base}, \uparrow, \downarrow\}$ and compute $\hat{\Sigma}_{cd} = \mathrm{cov}(U_c, U_d)$ (tiny ridge if needed). Choose simplex weights
$$\hat{\beta} \in \arg\min_{\beta \in \Delta_3} \beta^\top \hat{\Sigma}\, \beta, \qquad W^{\mathrm{stack}} = \sum_c \hat{\beta}_c\, W_c^{\mathrm{OOF}},$$
then renormalize W^{stack} to mean one.
3. **Light variance guard (optional; $\rho{=}1$ by default).** Cap dispersion relative to the baseline and re-project:
$$\alpha = \min\Big\{1,\ \frac{\rho\ \mathrm{Var}(W_{\pi'}^{\mathrm{m1}})}{\mathrm{Var}(W^{\mathrm{stack}})}\Big\}, \quad W^{\mathrm{blend}} = 1{+}\alpha\,(W^{\mathrm{stack}}{-}1), \quad \hat{W}_{\pi'} = \mathrm{IsoMeanOne}_S\big(W^{\mathrm{blend}}\big).$$
Each step preserves the sample mean; the final mean-one isotonic re-projection *weakly reduces dispersion* (majorization), hence $\mathrm{ESS}(\hat{W}_{\pi'}) \geq \mathrm{ESS}(W_{\pi'}^{\mathrm{m1}})$ deterministically.

Remark (transport view). In the continuous case the ideal component is $m^\star(s) = p_{S|\pi'}(s)/p_{S|\pi_0}(s)$; SIMCal-W is the L^2 monotone projection of noisy $W_{\pi'}$ onto the cone of increasing, mean-one functions of S (a monotone rearrangement along S).

3.3 ESTIMATORS: CAL-IPS, OC-IPS, DR-CPO, AND TR-CPO

Let $\hat{q}(x, a) \approx \mathbb{E}[R \mid X{=}x, A{=}a]$ and $\hat{g}_{\pi'}(x) = \sum_a \pi'(a \mid x)\, \hat{q}(x, a)$; all nuisances are cross-fitted and OOF predictions are used inside IFs.

Calibrated IPS.

$$\widehat{V}_{\mathrm{IPS}}(\pi') = \frac{1}{n}\sum_{i=1}^{n} \hat{W}_{\pi',i}\, R_i, \qquad \phi_i^{\mathrm{IPS}} = \hat{W}_{\pi',i}\, R_i^{\mathrm{OOF}} - \widehat{V}_{\mathrm{IPS}}.$$

Orthogonalized IPS (OC-IPS). Add a mean-zero orthogonal term using the raw ratio $W_{\pi'}^{\mathrm{raw}}$ and a fold-honest isotonic fit $\hat{m}(S) \approx \mathbb{E}[W_{\pi'}^{\mathrm{raw}} \mid S]$:

$$\widehat{V}_{\text{OC-IPS}} = \widehat{V}_{\mathrm{IPS}} + \frac{1}{n}\sum_{i=1}^{n} \big(W_{\pi',i}^{\mathrm{raw}} - \hat{m}(S_i)\big)\,\big(R_i^{\mathrm{OOF}} - \hat{f}(S_i)\big).$$

This solves the leading EIF moment and restores $\sqrt{n}$ rates under mild conditions.

DR-CPO (sequence-aware DR).

$$\widehat{V}_{\mathrm{DR}}(\pi') = \frac{1}{n}\sum_{i=1}^{n} \Big\{\hat{g}_{\pi'}(X_i){+}\hat{W}_{\pi',i}\big(R_i{-}\hat{q}(X_i, A_i)\big)\Big\}, \quad \phi_i^{\mathrm{DR}} = \hat{g}_{\pi'}(X_i){+}\hat{W}_{\pi',i}\big(R_i^{\mathrm{OOF}}{-}\hat{q}_i^{\mathrm{OOF}}\big){-}\widehat{V}_{\mathrm{DR}}.$$

TR-CPO (targeted & retargeted DR; optional triply-robust add-on). *Targeting* updates $\hat{q}$ along the clever covariate $H(X, A) = \hat{W}_{\pi'}$ (identity or logit link) to solve $\frac{1}{n}\sum_i \hat{W}_{\pi',i}\big(R_i^{\mathrm{OOF}} - \hat{q}_\varepsilon(X_i, A_i)\big) = 0$. *Retargeting* applies a control variate anchored at $(\hat{W}_{\pi'} - 1)$:

$$Z_i = \hat{W}_{\pi',i}\big(R_i^{\mathrm{OOF}}{-}\hat{q}_i^{\mathrm{OOF}}\big), \quad \hat{\gamma} = \frac{\mathrm{cov}(Z, \hat{W}_{\pi'} - 1)}{\mathrm{Var}(\hat{W}_{\pi'} - 1)}, \quad \widehat{V}_{\mathrm{TR}} = \frac{1}{n}\sum_{i=1}^{n} \Big\{\hat{g}_{\pi'}(X_i){+}Z_i{-}\hat{\gamma}\,(\hat{W}_{\pi',i}{-}1)\Big\},$$

$$\phi_i^{\mathrm{TR}} = \hat{g}_{\pi'}(X_i) + Z_i - \hat{\gamma}\,(\hat{W}_{\pi',i} - 1) - \widehat{V}_{\mathrm{TR}}.$$

(When labels are sparse and MAR in S, a fold-honest label-propensity $\hat{\pi}_L(S) \in [\varepsilon, 1]$ yields a *triply robust* correction; see the appendix.)

3.4 IF-SPACE STACKING (VARIANCE-OPTIMAL CONVEX ENSEMBLING)

For a small library $\mathcal{E}$ of regular estimators (e.g., DR/TMLE/MRDR variants, capped IPS), form the matrix of centered IF columns $\Phi = [\phi^{(e)}]_{e\in\mathcal{E}}$ (computed OOF on the same folds), estimate $\hat{\Sigma} = \frac{1}{n}\Phi^\top\Phi + \lambda I$, and solve the simplex QP

$$\hat{\alpha} \in \arg\min_{\alpha\in\Delta} \alpha^\top\hat{\Sigma}\,\alpha, \qquad \widehat{V}_{\text{stack}} = \sum_{e\in\mathcal{E}} \hat{\alpha}_e\,\widehat{V}^{(e)}, \quad \phi^{\text{stack}} = \sum_{e\in\mathcal{E}} \hat{\alpha}_e\,\phi^{(e)}.$$

4 THEORY: EIF, DESIGN-BY-PROJECTION, AND EFFICIENCY

We state the main results; proofs and technical lemmas are deferred to the appendix.

Surrogate model and EIF. Let $R^\star = \mathbb{E}[Y \mid S]$ and $m^\star(S) = \mathbb{E}[W_{\pi'} \mid S]$. Under *mean sufficiency* $\mathbb{E}[Y \mid X, A, S] = R^\star(S)$,

$$V(\pi') = \mathbb{E}\big[m^\star(S)\,R^\star(S)\big], \qquad \phi_{\text{sur}}(O;\pi') \;=\; g^\star_{\pi',R}(X) \;+\; m^\star(S)\big(R^\star - q^\star_R(X,A)\big) \;-\; V(\pi'),$$

with $q^\star_R(x,a) = \mathbb{E}[R^\star \mid X{=}x, A{=}a]$ and $g^\star_{\pi',R}(x) = \sum_a \pi'(a \mid x)\,q^\star_R(x,a)$.

Theorem 1 (Surrogate EIF and variance reduction)**.** *Let ϕ_{uncon} be the canonical gradient in the nonparametric model that does not use S. Then ϕ_{sur} is the canonical gradient in the surrogate model, and $\mathrm{Var}(\phi_{\text{sur}}) \le \mathrm{Var}(\phi_{\text{uncon}})$, with strict inequality unless $W_{\pi'}$ is $\sigma(S)$-measurable and $R^\star$ is degenerate.*

Knowledge–Riesz (Influence Representer). Let L_0^2 be the mean-zero Hilbert space with inner product $\langle f, g\rangle = \mathbb{E}[fg]$, and let $I(P) = \phi^\star + T(P)^\perp$ denote the affine class of influence functions in a baseline model. For a nonempty closed convex set $\mathcal{C} \subset L_0^2$ encoding justified knowledge (e.g., $\sigma(S)$-measurability, mean-one S-monotone weight components, simplex hulls), define $I_{\mathcal{C}}(P) = I(P) \cap \mathcal{C}$ and $\phi_{\mathcal{C}} = \arg\min_{\phi\in I_{\mathcal{C}}(P)} \mathbb{E}[\phi^2]$.

Theorem 2 (Knowledge–Riesz (Influence Representer))**.** (i) Metric projection & information improvement. *$\phi_{\mathcal{C}} = \Pi_{I(P)\cap\mathcal{C}}(0)$ is unique and satisfies $\|\phi_{\mathcal{C}}\|_2^2 \le \|\phi^\star\|_2^2$, with equality iff $I(P)\cap\mathcal{C}$ already contains $\phi^\star$. If $\mathcal{C}_1 \subseteq \mathcal{C}_2$, then $\|\phi_{\mathcal{C}_2}\|_2^2 \le \|\phi_{\mathcal{C}_1}\|_2^2$.* (ii) Attainability. *Replacing nuisances by their projections into $\mathcal{C}$ and applying a one-step/TMLE update with cross-fitting yields a regular estimator with IF $\phi_{\mathcal{C}}$.*

Corollary 1 (Blackwell–efficiency monotonicity)**.** *If S_2 is a garbling of S_1 (i.e., $\sigma(S_2) \subseteq \sigma(S_1)$), then $\mathrm{Var}\big(\phi_{\text{sur}}(S_1)\big) \le \mathrm{Var}\big(\phi_{\text{sur}}(S_2)\big)$, with strict inequality unless $W_{\pi'}$ is already $\sigma(S_2)$-measurable and $R^\star$ is degenerate.*

Consequences for CJE (i) *Conditioning:* taking $\mathcal{C} = \{f : f = \mathbb{E}[f \mid S]\}$ recovers Theorem 1. (ii) *Mean-one monotone weights:* restricting the weight component to $\{w : \mathbb{E}[w] = 1,\ w \uparrow S\}$ corresponds to SIMCal-W and weakly reduces dispersion in finite samples (majorization). (iii) *Stacking:* restricting to the convex hull of candidate IF columns gives the variance-optimal convex ensemble.

Proposition 1 (Cal-IPS: mean correctness and dispersion control)**.** *Let $R = \hat{f}(T(S))$ be AutoCal-R (monotone in S or two-stage index; cross-fitted), and let $W^{\text{m1}}_{\pi'} \triangleq W_{\pi'}/\mathbb{E}[W_{\pi'}]$ denote the uncalibrated mean-one ratios. Let $\hat{W}_{\pi'}$ be SIMCal-W weights (OOF stack + mean-one isotonic, optional guard $\rho \ge 1$). Then $\widehat{V}_{\text{IPS}} = \frac{1}{n}\sum_i \hat{W}_{\pi',i}R_i \to_p V(\pi')$, and $\mathrm{Var}_n(\hat{W}_{\pi'}) \le \rho\,\mathrm{Var}_n(W^{\text{m1}}_{\pi'})$ with $\mathrm{ESS}(\hat{W}_{\pi'}) \ge \mathrm{ESS}(W^{\text{m1}}_{\pi'})$ deterministically.*

Theorem 3 (DR-CPO / TR-CPO: $\sqrt{n}$ limits and efficiency)**.** *Assume mean sufficiency, suitable tails/moments, and cross-fitted nuisances satisfying the one-of-two rate condition $\|\hat{q} - q^\star_R\|_2 \cdot \|\hat{W}_{\pi'} - m^\star\|_2 = o_p(n^{-1/2})$ (e.g., either factor $= o_p(n^{-1/4})$). Then*

$$\sqrt{n}\big(\widehat{V}_{\text{DR}} - V(\pi')\big) \;\rightsquigarrow\; \mathcal{N}\big(0, \mathrm{Var}(\phi_{\text{sur}})\big), \qquad \sqrt{n}\big(\widehat{V}_{\text{TR}} - V(\pi')\big) \;\rightsquigarrow\; \mathcal{N}\big(0, \mathrm{Var}(\phi_{\text{sur}})\big),$$

i.e., DR-CPO and its targeted/retargeted refinement TR-CPO attain the surrogate efficiency bound.

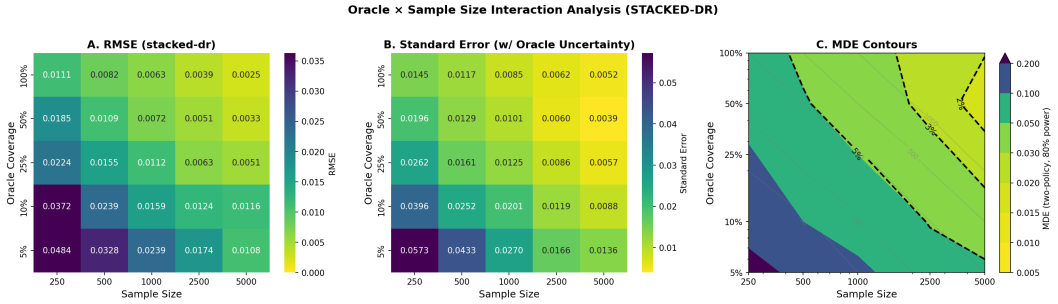

Figure 1: **Oracle × sample size interaction for `stacked-dr`.** (A) debiased RMSE; (B) SE including OUA; (C) MDE (two-policy, 5% two-sided, 80% power). Cells show means; dashed curves are iso–label budgets $m = n\times$coverage. Lower is better.

Budgeted bound (variance cap). For $\rho \geq 1$, define

$$\mathcal{W}_\rho \;=\; \Big\{ m :\; \mathbb{E}[m] = 1,\; m \uparrow S,\; \mathbb{E}\big[(m-1)^2\big] \leq \rho\, \mathbb{E}\big[(m^\star - 1)^2\big] \Big\},$$

and let $m^\star_\rho$ be the $L^2(P_S)$ projection of $m^\star$ onto $\mathcal{W}_\rho$. Define the budgeted gradient $\phi^{(\rho)} = g^\star_{\pi',R}(X) + m^\star_\rho(S)\big(R^\star - q^\star_R(X,A)\big) - V(\pi')$.

Theorem 4 (Budgeted information bound)**.** *The optimal asymptotic variance under the cap ρ equals $\mathrm{Var}(\phi^{(\rho)})$, which is nonincreasing in ρ and satisfies $\lim_{\rho\to\infty} \mathrm{Var}(\phi^{(\rho)}) = \mathrm{Var}(\phi_{\mathrm{sur}})$. If SIMCal-W converges to $m^\star_\rho$ (with the guard), then TR-CPO attains $\mathrm{Var}(\phi^{(\rho)})$.*

Theorem 5 (IF-space stacking)**.** *Let $\{\widehat{V}^{(e)}\}_{e\in\mathcal{E}}$ be regular, asymptotically linear estimators with centered IFs $\{\phi^{(e)}\}$, and let $\hat{\alpha} \in \arg\min_{\alpha\in\Delta} \alpha^\top \hat{\Sigma}\, \alpha$ with $\hat{\Sigma}$ the empirical IF covariance. If $\hat{\Sigma} \to \Sigma$ uniformly on Δ, then the stacked estimator $\widehat{V}^{(\hat{\alpha})}$ is asymptotically linear with IF $\phi^{(\alpha^\star)}$ and variance $\min_{\alpha\in\Delta} \alpha^\top \Sigma\, \alpha \leq \min_e \Sigma_{ee}$. An outer split leaves the limit unchanged.*

Corollary 2 (Carathéodory sparsity)**.** *If $\Phi = [\phi^{(e)}]$ has empirical rank r, the variance-optimal convex combination uses at most $r+1$ base estimators.*

Proposition 2 (OUA jackknife consistency)**.** *Let $\widehat{V}(\hat{f})$ be any* CJE *estimator that uses $R = \hat{f}(S)$ (cross-fitted along the IF path). Under $L^2(P_S)$-consistency of $\hat{f}$, fold honesty, and mild smoothness of $f \mapsto \widehat{V}(f)$, the delete-one-oracle-fold jackknife consistently estimates the calibration-induced variance component, so that $\widehat{\mathrm{Var}}_{\mathrm{total}} = \widehat{\mathrm{Var}}_{\mathrm{main}} + \widehat{\mathrm{Var}}_{\mathrm{oracle}}$ is a consistent variance estimator.*

Discussion. Theorem 2 formalizes *Design-by-Projection*: intersecting the IF class with justified closed convex sets can only lower the attainable variance, and projection-designed estimators (AutoCal-R, SIMCal-W, DR/TMLE with cross-fitting, IF-Stack) *attain* the corresponding bound in their respective models. In finite samples, isotonic projections additionally *majorize* dispersion, explaining the deterministic ESS gains delivered by SIMCal-W. (Metric projections are firmly non-expansive; composing DbP modules yields a non-expansive, i.e., 1-Lipschitz, pipeline.)

5 Experiments

We evaluate CJE on an Arena-derived benchmark to measure: (i) *stability* (ESS/tails) from SIMCal-W; (ii) *accuracy and interval quality* using OUA-augmented CIs; (iii) *ordering robustness* under overlap stress; and (iv) sensitivity to key design choices (AutoCal-R mode, IF-space stacking, variance guard ρ).

5.1 Setup

Data & policies. We use n=4,989 prompts from public Chatbot Arena logs (Zheng et al., 2023) collected under a fixed logger π_0. We compare five policies: *base*, *clone* (A/A), *prompt-variant*, *premium*, and adversarial *unhelpful*.

Judges & oracle. Every row has a scalar judge score S; a small i.i.d. oracle slice provides labels Y. AutoCal-R is cross-fitted; out-of-fold R^{OOF} are used along IF paths.

Propensities. Sequence-level TF forms $W_{\pi'} = \exp\{\log p_{\pi'} - \log p_{\pi_0}\}$; TF conformance filters apply (App. E).

Baselines & metrics. We compare IPS/SNIPS (with clipping/overlap weighting), DR/TMLE/MRDR, and calibrated variants (Cal-IPS, OC-IPS, DR-CPO, TR-CPO, IF-Stack); we report debiased $\mathrm{RMSE}^{\mathrm{d}}$, interval score, coverage gap, SE GM, and ranking metrics (pairwise wins, τ, regret).

5.2 Main results

Table 1: Accuracy & Uncertainty Metrics

Estimator	$\mathrm{RMSE}^{\mathrm{d}}\downarrow$	IS (interval score) $\downarrow$	$\lvert\mathrm{Cov}-95\rvert\downarrow$	SE GM $\downarrow$	Pairwise % $\uparrow$	Top-1 % $\uparrow$	$\tau\uparrow$	Regret $\downarrow$	Runtime (s) $\downarrow$
stacked-dr	0.0226	0.0767	<u>0.467</u>	0.0125	**91.9**	**83.1**	**0.837**	**0.0039**	17.2
stacked-dr-oc	**0.0225**	**0.0755**	0.733	<u>0.0123</u>	90.2	79.4	0.804	<u>0.0043</u>	41.8
stacked-dr-oc-tr	<u>0.0225</u>	<u>0.0755</u>	0.760	**0.0122**	90.1	79.3	0.802	0.0043	54.2
calibrated-dr-cpo	0.0227	0.1450	4.493	0.0258	<u>91.0</u>	<u>81.0</u>	<u>0.819</u>	0.0056	12.6
oc-dr-cpo	0.0459	0.0973	4.787	0.0202	78.3	46.4	0.565	0.0086	13.4
dr-cpo	0.0460	0.2171	4.307	0.0386	78.4	46.3	0.567	0.0086	12.9
tr-cpo-e	0.1355	0.1472	**0.040**	0.0341	72.4	32.9	0.448	0.0103	10.5
tr-cpo-e-anchored-orthogonal	0.1524	0.1845	0.467	0.0426	71.3	32.4	0.425	0.0118	10.8
calibrated-ips	0.0245	0.5261	1.960	0.0947	46.1	17.9	-0.078	0.1727	<u>5.1</u>
orthogonalized-ips	0.1591	0.6812	2.893	0.1651	38.3	8.6	-0.234	0.2780	5.9
SNIPS	0.1596	0.7379	3.267	0.1815	38.3	8.7	-0.235	0.2785	**4.5**

$\downarrow$: lower is better, $\uparrow$: higher is better. **Bold**: best, <u>underlined</u>: second-best. Metrics averaged across all regimes.

Accuracy and ordering. Table 1 shows stacked, calibrated DR variants dominate level accuracy and ranking quality with tight uncertainty. `stacked-dr` attains the best (or tied-best) pairwise wins, τ, regret, and competitive $\mathrm{RMSE}^{\mathrm{d}}$. OC/TR refinements (`stacked-dr-oc`, `stacked-dr-oc-tr`) further shave SE GM at additional compute. IPS baselines remain fast but inaccurate, with unstable uncertainty and weak ordering despite clipping/overlap weighting.

Table 2: Weight Diagnostics: SIMCal Calibration Effect

	ESS (%)		Weight CV		Max Weight		Tail α	
Policy	SNIPS→Cal	Δ	SNIPS→Cal	Δ	SNIPS→Cal	Δ	SNIPS→Cal	Δ
Clone	26.2% → 98.8%	+278%	1.8 → 0.1	+96%	0.040 → 0.002	+95%	1.08 → > 10	> 900%
Parallel Universe Prompt	0.6% → 94.6%	+15877%	26.6 → 0.2	+99%	0.617 → 0.003	+99%	0.56 → > 10	> 900%
Premium	0.7% → 80.8%	+12280%	16.8 → 0.4	+97%	0.409 → 0.004	+99%	0.32 → > 10	> 900%
Unhelpful	0.4% → 84.0%	+21908%	24.1 → 0.4	+98%	0.619 → 0.005	+99%	0.13 → > 10	> 900%

Stability from SIMCal-W. Table 2 quantifies SIMCal-W's effect on weights. The mean-one isotonic projection (with OOF stacking) *deterministically* reduces dispersion via majorization, yielding large ESS uplifts, dramatic CV shrinkage, and tail relief (Hill $\alpha > 2$), even under extreme raw-overlap (e.g., *prompt-variant*).

Precision & power planner. Figure 1 shows how precision scales with the joint budget of logs (n) and oracle coverage. Panel B reports SEs that include the OUA addition; Panel C converts them to MDE for a two-policy comparison at 80% power. Dashed iso–label-budget curves (m=$n\times$coverage) reveal tradeoffs between more logs and more labels; along a fixed dashed curve (constant m), MDE tightens as n grows.

5.3 Ablations (brief)

Ablations. Disabling SIMCal-W raises variance and worsens interval score/coverage; IF-Stack matches or beats the best single DR with a slight CI widening under an outer split; the guard ρ rarely engages for $\rho\in[1,2]$ (default 1); one rollout per (X,π') suffices for $\hat{g}_{\pi'}$ with a light smoother.

5.4 DIAGNOSTICS AND GATES (SUMMARY)

Per policy we render ESS/tails (baseline vs. SIMCal-W), an S-overlap heatmap, a DR-orthogonality CI, a judge-reliability diagram with a coverage badge, and the OUA share. Gates: OVERLAP, JUDGE, IDENTIFICATION (triggers REFUSE-LEVEL), DR, MULTIPLICITY; thresholds in App. D.

6 LIMITATIONS

Overlap (positivity). As with IPS/DR, CJE requires support overlap between π_0 and each π'. When overlap is poor, raw ratios are heavy-tailed and uncertainty inflates. *Mitigations:* SIMCal–W reduces dispersion and raises ESS; if tails persist we (i) gate on ESS and Hill indices, (ii) use overlap weighting or cohort restriction, and (iii) run an online check when $\widehat{\alpha}_{\text{Hill}}<1$ or single–row dominance persists (App. D).

Judge assumptions (surrogate validity). AutoCal–R assumes *mean sufficiency* and monotonicity in S (or a learned index). If strained, the two–stage fallback preserves mean honesty but targets $\mathbb{E}[f(S^{\pi'})]$. *Mitigations:* surface reliability curves and regional residuals; when evidence is weak, label as surrogate–target, widen/refresh the oracle slice, and target labels where error concentrates.

Calibration coverage (identification). If a π' pushes S outside the labeled range, isotonic calibration flattens at the boundary and levels are not point–identified. *Mitigations:* flag LIMITED CALIBRATION SUPPORT and set REFUSE-LEVEL (report rankings and partial–ID bounds) until targeted labels cover the uncovered S region (App. D).

Approximate sufficiency (bias modulus). When $\mathbb{E}[Y \mid X, A, S] \neq \mu(S)$, the residual $\Delta(X, A, S)$ induces bias proportional to calibration error. *Mitigations:* by Cauchy–Schwarz, $|\text{Bias}| \leq \|m - W_{\pi'}\|_2 \, \|\Delta\|_2$; DbP shrinks $\|m - W_{\pi'}\|_2$, so bias is second order when either calibration is tight or the violation small; we surface this via diagnostics and invoke REFUSE-LEVEL when unbounded.

Label sparsity and MAR. TR–CPO's label term assumes MAR in S with bounded label propensity; severe violations or tiny propensities degrade guarantees. *Mitigations:* monitor labeled ESS and $\min \hat{\pi}_L$, stratify labeling to shore up sparse strata, or revert to DR without label–propensity terms.

Temporal dependence and logger drift. Non–stationarity (launches, safety updates) can bias or widen intervals. *Mitigations:* report dependence–robust SEs (block/stationary bootstrap), shorten analysis windows, and monitor judge drift via rank–based/residual change detection with FDR control.

Oracle independence and leakage. OUA assumes the oracle slice is i.i.d. and fold–honest. *Mitigations:* reuse deterministic folds across modules, de–duplicate the slice, and periodically refresh it.

Selection and multiplicity. Scanning many π' inflates winner's curse. *Mitigations:* use FDR control (BH/BY), optionally an outer split for IF–Stack to reduce selection optimism, and emphasize pre–specified contrasts.

Teacher forcing and API drift. Accurate propensities require deterministic, chat–native TF (stable tokenizer/template) with additivity/conditionality invariants; missing/invalid TF corrupts ratios. *Mitigations:* enforce schema/conformance checks, ledger failures, and treat results as conditional on TF quality (App. E).

Subgroups and fairness. Calibration quality and ESS gains may differ across subgroups. *Mitigations:* provide subgroup diagnostics (ESS, reliability) and, when feasible, use subgroup–specific calibration/weights or constrained pooling.

Judge informativeness (garbling). Coarser judges raise the surrogate information bound and widen CIs. *Mitigations:* prefer richer rubrics (multi–dimensional S with stable aggregation) and validate with coarsening ablations (empirical Blackwell monotonicity).

Compute. DR/TR–CPO add one rollout + judge per (X, π'); AutoCal–R refits for OUA add modest overhead. We amortize via TF caches, shared folds, and a small stacking library.

Ethics Statement. We analyze retrospective logs that may include sensitive content. Diagnostics/gates prevent overconfident claims under poor overlap/coverage and surface judge drift. When identification fails, we report rankings only (REFUSE-LEVEL) and recommend targeted labeling or online checks. Any deployment should assess subgroup reliability and adopt privacy safeguards for logs.

Reproducibility Statement. We provide two example configs and the fold–hash rule (supplement); schema, TF contract, pseudocode, and numerics appear in the appendices. Additional artifacts will be released after review.

7 CONCLUSION

We introduced **CJE**, an audit-ready recipe for offline policy evaluation with LLM judges built around a single rule: *Design-by-Projection (DbP)*. The principle is simple—encode justified assumptions as *closed convex sets* and project valid objects onto them. Projections onto *subspaces* (nuisance-orthogonal scores), *monotone cones* (mean-preserving reward/weight calibration), and *simplices* (variance-hedged stacking) preserve the estimand while *weakly reducing variance*.

Concretely, *AutoCal-R* learns a mean-preserving surrogate $R = f(T(S))$; *SIMCal-W* produces unit-mean, S-monotone ratios with deterministic ESS uplift via OOF stacking and a light guard; sequence-aware *OC-IPS/DR-CPO/TR-CPO* deliver $\sqrt{n}$ inference under cross-fitting; *IF-Stack* minimizes plug-in IF variance; and *OUA* adds calibration uncertainty for honest CIs.

Theoretically, our *Knowledge–Riesz (Influence Representer)* theorem explains why intersecting the admissible IF class with justified convex knowledge lowers the attainable variance and is *attainable* with projection-designed estimators. Two design corollaries follow: (*i*) *Blackwell–efficiency monotonicity*—richer judges (finer σ-fields) strictly help; (*ii*) isotonic mean-one calibration *Lorenz-dominates* baseline weights, improving all Schur-convex dispersion metrics (not just ESS). Empirically, on Arena-derived logs, SIMCal-W turns near-degenerate ratios into stable weights (large ESS gains), calibrated DR achieves tight, well-calibrated intervals and near-$\sqrt{n}$ scaling, and stacking improves ordering; when calibration support is limited, CJE *flags* the issue and reports robust *rankings* with conservative uncertainty (REFUSE-LEVEL).

Takeaways. (i) *Project before you compute*: express assumptions as convex sets and apply metric projections.
(ii) Treat teacher forcing and diagnostics (overlap, tails, reliability, orthogonality) as first-class artifacts.
(iii) Report *oracle–uncertainty–aware* variance, not just IF variance.
(iv) Ship with explicit gates for OVERLAP, JUDGE, IDENTIFICATION, DR, and MULTIPLICITY.
(v) Prefer richer, more informative judges (Blackwell monotonicity) and shape-stabilized weights (Lorenz dominance).

Future work. Selection-aware inference over large policy sets; robust/DP isotonic calibration (mirror/Bregman DbP) for heavy tails; active oracle budgeting via shadow prices; sequential/agent evaluations with prefix-aware SIMCal and stepwise DR; and subgroup-aware constraints with fairness diagnostics. We include example configs and a deterministic fold hash rule in the supplement; full algorithms, pseudocode, and diagnostics are in the appendices. Additional public artifacts will be released after review.

REFERENCES

Miriam Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and Edward Silverman. An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 26(4):

641–647, 1955.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jason Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Brian McKinnon, Carol Chen, Catherine Olsson, Daniel Brown, El Mahdi El-Mhamdi, Ethan Perez, Ilya Tolstikhin, Ishita Ganguli, Tom Henighan, Jared Carter, Shauna Kravec, Scott Johnston, Tim Shlegeris, Kamal Ndousse, Chitwan Saharia, Elizabeth Barnes, Ellie Soros, Hai Tieu, Jared Kaplan, Jan Leike, Geoffrey Irving, and Dario Amodei. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv*, 2022. URL https://arxiv.org/abs/2204.05862.

Moulinath Banerjee. Likelihood ratio inference in regular models with a convex parameter space. *Annals of Statistics*, 29(1):169–200, 2001.

Richard E. Barlow, David J. Bartholomew, John M. Bremner, and Herman D. Brunk. *Statistical Inference under Order Restrictions*. Wiley, 1972.

Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2 edition, 2017.

Peter J. Bickel, Chris A. J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, 1993.

Leo Breiman. Stacked regressions. *Machine Learning*, 24(1):49–64, 1996.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney K. Newey, and James M. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097.

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

Christian Fong and Edward H. Kennedy. Robustness of calibration-based off-policy evaluation. *arXiv*, 2022. URL https://arxiv.org/abs/2207.01605.

Christian Fong, Christopher Harshaw, Emily Puelz, and Edward H. Kennedy. Consistent off-policy evaluation with overlap weighting. *arXiv*, 2021. URL https://arxiv.org/abs/2107.07263.

Constantine E. Frangakis and Donald B. Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

Jaroslav Hájek. Asymptotic theory of rejective sampling with unequal probabilities. *The Annals of Mathematical Statistics*, 36(5):1491–1523, 1965.

G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, 2 edition, 1952.

Bruce M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174, 1975.

Daniel G. Horvitz and Donovan J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446.

Edward L. Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 652–661, 2016.

Nathan Kallus. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Nathan Kallus and Xinkun Mao. Calibration for honest and efficient off-policy evaluation. *Journal of Machine Learning Research*, 23(186):1–69, 2022.

Hyung Won Kim et al. Prometheus: Towards trainable LLM-as-a-judge for evaluating LLMs. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. *arXiv*, 2023. URL https://arxiv.org/abs/2302.14520. WMT 2023.

Michael R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, 2008.

Hans R. Künsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989.

Jason Lee et al. Offline policy evaluation for RLHF logs. *arXiv*, 2024. URL https://arxiv.org/abs/2401.01234.

Fan Li, Kari Lock Morgan, and Alan M. Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. Unbiased offline evaluation of contextual-bandit performance. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*, 2011.

Jun S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, 2001.

Zihan Liu et al. Preference alignment drift in large language models and how to fix it. *arXiv*, 2023. URL https://arxiv.org/abs/2310.01234.

Albert W. Marshall, Ingram Olkin, and Barry C. Arnold. *Inequalities: Theory of Majorization and Its Applications*. Springer, 2 edition, 2011.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Luke Kelton, Fraser Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *arXiv*, 2022. URL https://arxiv.org/abs/2203.02155.

Art B. Owen. *Monte Carlo Theory, Methods and Examples*. 2013. Monograph.

Judea Pearl. The mediation formula: A guide to the assessment of causal pathways in nonlinear models. *Epidemiology*, 23(1):119–121, 2012.

Dimitris N. Politis and Joseph P. Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313, 1994.

Ross L. Prentice. Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine*, 8(4):431–440, 1989.

Tim Robertson, F. T. Wright, and R. L. Dykstra. *Order Restricted Statistical Inference*. Wiley, 1988.

James M. Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992.

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.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 814–823, 2015.

Anastasios A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer, 2006.

Lars van der Laan, Ziming Lin, Marco Carone, and Alex Luedtke. Stabilized inverse probability weighting via isotonic calibration. In *Proceedings of the Fourth Conference on Causal Learning and Reasoning (CLeaR)*, volume 275 of *Proceedings of Machine Learning Research*, pp. 139–173. PMLR, 2025a. URL https://proceedings.mlr.press/v275/. CLeaR 2025.

Lars van der Laan, Alex Luedtke, and Marco Carone. Doubly robust inference via calibration. *arXiv*, 2025b. URL https://arxiv.org/abs/2411.02771.

Mark J. van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, 2011.

Mark J. van der Laan, Eric C. Polley, and Alan E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1):Article 25, 2007.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 2000.

Tyler J. VanderWeele. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press, 2015.

Alex Wang et al. Large language models still need human feedback. *arXiv*, 2023. URL https://arxiv.org/abs/2305.05658.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Joseph E. Gonzalez, Matei Zaharia, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 46595–46623, 2023.

A Notation and Formal Setup

Observed data and policies. We observe i.i.d. logs

$$O_i = (X_i, A_i, S_i, Y_i^{\mathrm{obs}}, L_i), \qquad i = 1, \dots, n,$$

generated under a fixed logger $A_i \sim \pi_0(\cdot \mid X_i)$. A scalar judge $S_i = s(X_i, A_i)$ is available on *all* rows. The label indicator $L_i \in \{0, 1\}$ marks inclusion in the oracle slice; when $L_i = 1$ we observe $Y_i^{\mathrm{obs}} = Y_i$, otherwise Y_i^{obs} is missing. For a candidate policy π', define the sequence-level importance ratio

$$W_{\pi',i} \;=\; \frac{\pi'(A_i \mid X_i)}{\pi_0(A_i \mid X_i)} \;=\; \exp\Big\{ \log p_{\pi'}(A_i \,|\, X_i) - \log p_{\pi_0}(A_i \,|\, X_i) \Big\},$$

computed via teacher forcing (TF) with the model's own tokenizer/rendering. Write

$$W^{\mathrm{m1}}_{\pi',i} \;=\; \frac{W_{\pi',i}}{\frac{1}{n}\sum_{j=1}^{n} W_{\pi',j}}$$

for the sample–mean–one (SNIPS) baseline (global normalization over the evaluation cohort).

Estimand. Let $Y(\pi')$ denote the outcome under the counterfactual draw $A \sim \pi'(\cdot \mid X)$. The target is

$$V(\pi') = \mathbb{E}\big[Y(\pi')\big].$$

A.1 Assumptions (compact)

(D1) Fixed logger & i.i.d. (X_i, A_i, S_i) are i.i.d. under π_0; TF log-likelihoods are stable and well-defined.

(D2) Overlap (positivity). $\pi_0(a \mid x) > 0$ whenever $\pi'(a \mid x) > 0$, and $\mathbb{E}_{\pi_0}[W_{\pi'}^2] < \infty$.

(D3) Judge coverage & stability. S is well-defined under both π_0 and π'; the Radon–Nikodym derivative on $\sigma(S)$ exists; the judge/rubric is stable on the analysis window.

(J1) Oracle slice. There exists an i.i.d. subsample $O = \{i : L_i = 1\}$ with $m = |O| \ll n$ on which Y is observed.

(J2-M) Mean sufficiency (monotone). $\mathbb{E}[Y \mid X, A, S] = \mu(S)$ with μ weakly nondecreasing.
(J2-SI) Single-index fallback. There exist $g^\star : \mathbb{R} \to \mathbb{R}$ and nondecreasing $\mu^\star$ such that $\mathbb{E}[Y \mid S] = \mu^\star\big(g^\star(S)\big)$.

(R1) Tails/moments. $\mathbb{E}[Y^2] < \infty$, $\mathbb{E}[S^2] < \infty$, and $\mathbb{E}_{\pi_0}[W_{\pi'}^2] < \infty$. When using the SIMCal-W variance cap $\rho \geq 1$, $\mathrm{Var}_n(\hat{W}_{\pi'}) \leq \rho \ \mathrm{Var}_n(W_{\pi'}^{\mathrm{m1}})$.

(R2) Calibration consistency. AutoCal-R satisfies $\|\hat{f}(T(S)) - \mathbb{E}[Y \mid S]\|_{L^2(P_S)} = o_p(1)$ (monotone or two-stage mode), and SIMCal-W satisfies $\|\hat{W}_{\pi'} - \mathbb{E}[W_{\pi'} \mid S]\|_{L^2(P)} = o_p(1)$.

(R3) One-of-two rates with cross-fitting. With nuisances $\hat{q}(x, a) \approx \mathbb{E}[R \mid x, a]$ and $\hat{W}_{\pi'}$,

$$\|\hat{q} - q_R^\star\|_{L^2(P)} \cdot \|\hat{W}_{\pi'} - m^\star\|_{L^2(P)} \ = \ o_p(n^{-1/2}),$$

e.g., either factor is $o_p(n^{-1/4})$ with the other consistent.

A.2 Cross-fitting and folds

Let $F : \{1, \ldots, n\} \to \{1, \ldots, K\}$ be a deterministic fold map (e.g., a hash of x_id). For any learner $\mathcal{L}$, train $\hat{\eta}^{(-k)} = \mathcal{L}$ on $\{i : F(i) \neq k\}$ and use out-of-fold predictions $\hat{\eta}_i^{\mathrm{OOF}} = \hat{\eta}^{(-F(i))}(O_i)$ in influence-function (IF) calculations. The same folds are reused across AutoCal-R, SIMCal-W, and DR nuisances.

A.3 Projection operators used by CJE

Monotone cone. $\mathcal{M}_\uparrow = \{f : \mathbb{R} \to \mathbb{R} \text{ nondecreasing}\}$. The isotonic projector (PAVA) $\Pi_{\mathcal{M}_\uparrow}$ enjoys: (i) L^2 optimality; (ii) *mean preservation* on the training sample; (iii) *dispersion reduction* by majorization.

Mean-one cone for weights. For $w \in \mathbb{R}^n$ ordered by S, define the mean-one isotonic projection

$$\mathrm{IsoMeanOne}_S(w) \ = \ \arg\min_u \ \sum_i (u_i - w_i)^2 \quad \text{s.t.} \quad u \in \mathcal{M}_\uparrow(S), \ \tfrac{1}{n} \sum_i u_i = 1,$$

where $\mathcal{M}_\uparrow(S)$ denotes vectors nondecreasing in the S-order. This preserves the sample mean and weakly reduces empirical variance; hence ESS weakly increases (deterministically, by majorization).

Simplex hull. For centered IF columns $\{\phi^{(e)}\}_{e \in \mathcal{E}}$, let $\Phi = [\phi^{(e)}]$ and $\Delta = \{\alpha : \alpha_e \geq 0, \ \sum_e \alpha_e = 1\}$. IF-space stacking solves $\min_{\alpha \in \Delta} \alpha^\top \hat{\Sigma} \, \alpha$ with $\hat{\Sigma} = (1/n)\Phi^\top \Phi + \lambda I$.

A.4 AutoCal-R and SIMCal-W primitives

AutoCal-R. On $O = \{i : L_i = 1\}$, fit $R = \hat{f}(T(S))$ by either: (i) *monotone* ($T(S) = S$), or (ii) *two-stage* ($T(S) = \mathrm{ECDF}\{g(S)\}$ with a spline g), selecting by OOF RMSE (1-SE rule). Use OOF predictions R^{OOF} along IF paths; the point estimate may use a pooled fit. A terminal isotonic step enforces *slice-mean preservation*.

SIMCal-W. (Per fold) Fit up/down isotonic maps on S to obtain $W_\uparrow^{\mathrm{OOF}}, W_\downarrow^{\mathrm{OOF}}$; include $W_{\mathrm{base}}^{\mathrm{OOF}} \equiv 1$. Define residuals T_i (IPS: R_i; DR: $R_i - \hat{q}(X_i, A_i)$). Choose $\hat{\beta} \in \arg\min_{\beta \in \Delta_3} \beta^\top \hat{\Sigma} \, \beta$ where $\hat{\Sigma}$ is the

covariance of $U_c = W_c^{\mathrm{OOF}} T$. Form $W^{\mathrm{stack}} = \sum_c \hat{\beta}_c W_c^{\mathrm{OOF}}$, renormalize to mean one (global, over the evaluation cohort), optionally apply the *variance guard*

$$\alpha = \min\Big\{1,\ \frac{\rho\ \mathrm{Var}(W^{\mathrm{m1}}_{\pi'})}{\mathrm{Var}(W^{\mathrm{stack}})}\Big\}, \qquad W^{\mathrm{blend}} = 1 + \alpha\,(W^{\mathrm{stack}} - 1),$$

and re-project by $\hat{W}_{\pi'} = \mathrm{IsoMeanOne}_S(W^{\mathrm{blend}})$.

A.5 DR NUISANCES AND SEQUENCE VALUE

Let $\hat{q}(x,a) \approx \mathbb{E}[R \mid x, a]$ and define $\hat{g}_{\pi'}(x) = \sum_a \pi'(a \mid x)\,\hat{q}(x,a)$. For sequences, approximate $\hat{g}_{\pi'}(x)$ with one (default) rollout $A' \sim \pi'(\cdot \mid x)$ and $R' = \hat{f}\big(s(x,A')\big)$; a light smoother (e.g., ridge over (x,z) features) can reduce Monte Carlo noise. Cross-fitting is used throughout.

A.6 INFLUENCE FUNCTIONS AND VARIANCE

Let $\{\phi_i\}_{i=1}^n$ denote the (approximately) centered influence–function contributions of $\hat{\psi}$, computed with cross–fitted/OOF nuisances and R^{OOF} along the IF path, so that $\frac{1}{n}\sum_{i=1}^n \phi_i \approx 0$. Under standard regularity conditions,

$$\sqrt{n}\,(\hat{\psi} - \psi) \ \xrightarrow{d}\ \mathcal{N}(0,\ \mathrm{Var}(\phi))\,.$$

We estimate the main IF variance and the total variance (including the oracle addition) by

$$\widehat{\mathrm{Var}}_{\mathrm{main}} \ = \ \frac{1}{n}\sum_{i=1}^n \phi_i^2, \qquad \widehat{\mathrm{Var}}_{\mathrm{total}} \ = \ \widehat{\mathrm{Var}}_{\mathrm{main}} + \widehat{\mathrm{Var}}_{\mathrm{oracle}},$$

and report the $(1-\alpha)$ Wald interval

$$\mathrm{CI}_{1-\alpha}:\ \hat{\psi} \ \pm\ z_{1-\alpha/2}\,\sqrt{\widehat{\mathrm{Var}}_{\mathrm{total}}}\,.$$

When serial or cluster dependence is a concern, we additionally report dependence–robust SEs (e.g., cluster–robust sandwich or block/stationary bootstrap) as a sensitivity analysis.

A.7 ORACLE–UNCERTAINTY–AWARE (OUA) JACKKNIFE

Partition O into K oracle folds $\{O_k\}_{k=1}^K$. For each k, refit AutoCal-R on $O \setminus O_k$, recompute $R^{(-k)}$, and rerun the full pipeline to obtain $\hat{\psi}^{(-k)}$. Then

$$\bar{\psi} = \tfrac{1}{K}\sum_{k=1}^K \hat{\psi}^{(-k)}, \qquad \widehat{\mathrm{Var}}_{\mathrm{oracle}} = \frac{K-1}{K}\sum_{k=1}^K \big(\hat{\psi}^{(-k)} - \bar{\psi}\big)^2.$$

(With unequal fold sizes, use the standard *weighted* delete-one-group formula.)

A.8 DIAGNOSTICS (DEFINITIONS)

ESS. $\mathrm{ESS}(W) = \big(\sum_i W_i\big)^2 / \sum_i W_i^2$; we report the fraction ESS/n. Under global mean-one normalization (SNIPS), $\sum_i W_i = n$ and

$$\frac{\mathrm{ESS}(W)}{n} \ = \ \frac{1}{1 + \mathrm{CV}^2(W)} \quad \text{with} \quad \mathrm{CV}^2(W) = \mathrm{Var}(W) \text{ (since } \mathbb{E}[W] = 1).$$

Unless stated otherwise, diagnostics use the global (not per-fold) mean-one scaling.

Max-weight share. $\max_i W_i \big/ \sum_j W_j$.

Tail index (Hill). For top-k order statistics $W_{(1)} \geq \cdots \geq W_{(k)}$, $\hat{\alpha}^{-1} = \frac{1}{k}\sum_{j=1}^k \log\big(W_{(j)}/W_{(k)}\big)$ (we sweep k over a stability grid and report the plateau).

Bhattacharyya affinity in S. $A_B = \int \sqrt{p_{S|\pi'}(s)\,p_{S|\pi_0}(s)}\,ds$ (discrete: sum over bins); $D_B = -\log A_B$.

Algorithm 1 AUTOCAL-R: mean-preserving reward calibration (cross-fitted; automatic two-stage fallback)

Inputs: Oracle pairs $\{(S_i, Y_i) : L_i{=}1\}$; folds $F(\cdot)$; smooth index class $g(\cdot)$ (splines+ridge)
Outputs: Global reward $R_i = \hat{f}(T(S_i))$ and OOF R_i^{OOF}
for $k = 1$ to K **do**
  Train set $O_{\neg k} = \{i : L_i{=}1,\ F(i) \neq k\}$; test set $O_k = \{i : L_i{=}1,\ F(i) = k\}$
  Monotone candidate: $\hat{f}_{\uparrow}^{(-k)} \in \arg\min_{f \in \mathcal{M}_{\uparrow}} \sum_{i \in O_{\neg k}} (Y_i - f(S_i))^2$; set $R_{\uparrow,i}^{\mathrm{OOF}} = \hat{f}_{\uparrow}^{(-k)}(S_i)$ for $i \in O_k$
  Two-stage candidate: fit $g^{(-k)}$ on $O_{\neg k}$; ranks $U_i = \mathrm{ECDF}_{O_{\neg k}}(g^{(-k)}(S_i))$; fit $\hat{h}_{\uparrow}^{(-k)} \in \arg\min_{h \in \mathcal{M}_{\uparrow}} \sum_{i \in O_{\neg k}} (Y_i - h(U_i))^2$; set $R_{2s,i}^{\mathrm{OOF}} = \hat{h}_{\uparrow}^{(-k)}(U_i)$ for $i \in O_k$
end for
Compute OOF risks (overall and by S tertile); select mode via 1-SE rule (prefer monotone unless two-stage is significantly better)
Refit the selected mode on the full oracle slice to obtain *global* R_i for all $i \in \{1{:}n\}$; retain R_i^{OOF} per fold for IFs
Note: The terminal isotonic step preserves the oracle-slice mean exactly.

DR orthogonality score. $n^{-1} \sum_i \hat{W}_{\pi',i} \big(R_i^{\mathrm{OOF}} - \hat{q}^{\mathrm{OOF}}(X_i, A_i)\big)$ with a Wald CI.

Coverage badge. Plug-in estimate of $\Pr_{\pi'}\big(S \notin [S_{\min}^{\mathrm{orc}}, S_{\max}^{\mathrm{orc}}]\big)$; large out-of-range mass with near-flat boundaries triggers LIMITED CALIBRATION SUPPORT and REFUSE-LEVEL.

A.9 SYMBOL GLOSSARY

Symbol	Meaning
X, A	Context, action (sequence)
$S = s(X, A)$	Judge score (scalar)
Y	Ground-truth outcome (on oracle slice)
π_0, π'	Logger and candidate policies
$W_{\pi'}$	Importance ratio $\pi'(A \mid X)/\pi_0(A \mid X)$
$W_{\pi'}^{\mathrm{m1}}$	Mean-one (SNIPS) baseline
$R = \hat{f}(T(S))$	Calibrated reward (AutoCal-R; monotone or two-stage)
$\hat{W}_{\pi'}$	Calibrated, unit-mean, S-monotone weights (SIMCal-W)
$\hat{q},\ \hat{g}_{\pi'}$	Outcome and policy–value nuisances for DR
ϕ	Per-row centered influence-function contribution
$\widehat{\mathrm{Var}}_{\mathrm{main}}$	IF variance $n^{-1}\widehat{\mathrm{Var}}(\phi_i)$
$\widehat{\mathrm{Var}}_{\mathrm{oracle}}$	Oracle jackknife variance addition
$\mathrm{ESS}(W)$	Effective sample size

B ALGORITHMS (EXTENDED)

This appendix gives compact, cross-fitted pseudocode for CJE modules: reward calibration (AUTOCAL-R), surrogate-indexed weight calibration (SIMCAL-W), estimators (CAL-IPS, OC-IPS, DR-CPO, TR-CPO), IF-space stacking, and oracle–uncertainty–aware variance (OUA). We reuse the same K-fold map $F(i) \in \{1{:}K\}$ across all modules. "OOF" denotes *out-of-fold* predictions used along the IF path.

Complexity notes. PAVA is $O(n)$ after a shared sort by S. SIMCAL-W is linear-time per fold; the stacking QP is $3{\times}3$ (weights) or a small $|\mathcal{E}|{\times}|\mathcal{E}|$ system. DR/TR-CPO add one rollout + judge per (X, π'). The OUA jackknife refits the calibrator K times and reruns the pipeline; TF caches and precomputed features amortize cost.

Algorithm 2 SIMCAL-W: surrogate-indexed, unit-mean monotone weight calibration (OOF project $\to$ stack $\to$ cap $\to$ re-project)

Inputs: Baseline mean-one ratios $W_{\pi'}^{\mathrm{m1}}$; scores S; residuals T (*IPS:* $T{=}R$; *DR:* $T{=}R-\hat{q}$); folds $F(\cdot)$; variance cap $\rho \geq 1$ (default 1)

Output: Calibrated weights $\hat{W}_{\pi'}$ (mean-one, S-monotone)

for $k = 1$ to K **do**

{OOF candidate projections}

Train $I_{\neg k} = \{i : F(i) \neq k\}$; test $I_k = \{i : F(i) = k\}$

Fit isotonic maps on $I_{\neg k}$: increasing $m_{\uparrow}^{(-k)}(S)$ and decreasing $m_{\downarrow}^{(-k)}(S)$ (via $-S$); rescale each to mean one on $I_{\neg k}$

Predict on I_k: $W_{\uparrow}^{\mathrm{OOF}} = m_{\uparrow}^{(-k)}(S),\ W_{\downarrow}^{\mathrm{OOF}} = m_{\downarrow}^{(-k)}(S),$ and include $W_{\mathrm{base}}^{\mathrm{OOF}} \equiv 1$

end for

OOF stacking (variance-aware). Form $U_c = W_c^{\mathrm{OOF}}\, T$ for $c \in \{\mathrm{base}, \uparrow, \downarrow\}$; compute $\hat{\Sigma}_{cd} = \mathrm{cov}(U_c, U_d) + \lambda \mathbf{1}_{c=d}$

Choose $\hat{\beta} \in \arg\min_{\beta \in \Delta_3} \beta^{\top} \hat{\Sigma}\, \beta$; set $W^{\mathrm{stack}} = \sum_c \hat{\beta}_c\, W_c^{\mathrm{OOF}}$; renormalize W^{stack} to sample mean one

Light variance guard (optional). $\alpha = \min\{1,\ \rho\ \mathrm{Var}(W_{\pi'}^{\mathrm{m1}}) / \mathrm{Var}(W^{\mathrm{stack}})\}$; set $W^{\mathrm{blend}} = 1 + \alpha\,(W^{\mathrm{stack}} - 1)$

Final projection. $\hat{W}_{\pi'} \leftarrow \mathrm{IsoMeanOne}_S\big(W^{\mathrm{blend}}\big)$ (*mean preserved; dispersion weakly decreases* $\Rightarrow$ ESS $\uparrow$)

Algorithm 3 CAL-IPS and OC-IPS

Inputs: Calibrated rewards R, R^{OOF} (Alg. 1); calibrated weights $\hat{W}_{\pi'}$ (Alg. 2); raw ratios $W_{\pi'}^{\mathrm{raw}}$; fold-honest $\hat{m}(S) \approx \mathbb{E}[W_{\pi'}^{\mathrm{raw}} \mid S]$

Outputs: $\widehat{V}_{\mathrm{IPS}}$, $\widehat{V}_{\mathrm{OC\text{-}IPS}}$ (IFs defined analogously to DR)

Cal-IPS: $\widehat{V}_{\mathrm{IPS}} = \dfrac{1}{n}\sum_{i=1}^{n} \hat{W}_{\pi',i}\, R_i, \quad \phi_i^{\mathrm{IPS}} = \hat{W}_{\pi',i}\, R_i^{\mathrm{OOF}} - \widehat{V}_{\mathrm{IPS}}$

OC-IPS: $\widehat{V}_{\mathrm{OC\text{-}IPS}} = \widehat{V}_{\mathrm{IPS}} + \dfrac{1}{n}\sum_{i=1}^{n} \big(W_{\pi',i}^{\mathrm{raw}} - \hat{m}(S_i)\big)\big(R_i^{\mathrm{OOF}} - \hat{f}(S_i)\big)$

Note: The orthogonal term solves the leading EIF moment and restores $\sqrt{n}$ rates under mild conditions.

C PROOFS AND TECHNICAL LEMMAS

We collect standing identities, shape-constrained facts, and proofs for the results in Section 4. Unless stated otherwise, expectations are under the logging law P_{π_0}; L^2 norms are with respect to the relevant marginal (e.g., $L^2(P)$ or $L^2(P_S)$). We reuse the fold map $F(\cdot)$ from Section A.2 and the projection operators from Section A.3.

C.1 STANDING IDENTITIES AND TOOLS

Change of measure. For any integrable $h(X, A, S, Y)$ and any candidate π',

$$\mathbb{E}\big[W_{\pi'}\, h(X, A, S, Y)\big] \;=\; \mathbb{E}_{\pi'}\big[h(X, A, S, Y)\big], \qquad \mathbb{E}[W_{\pi'}] = 1. \tag{1}$$

Doob–Dynkin / conditional expectation as L^2 projection. Let $\mathcal{G} = \sigma(S)$. Then $m^{\star}(S) := \mathbb{E}[W_{\pi'} \mid \mathcal{G}]$ is the L^2 projection of $W_{\pi'}$ onto the closed subspace $L^2(\mathcal{G}) \subset L^2(P)$, i.e.,

$$\mathbb{E}\big[(W_{\pi'} - U(S))^2\big] = \mathbb{E}\big[(W_{\pi'} - m^{\star}(S))^2\big] + \mathbb{E}\big[(m^{\star}(S) - U(S))^2\big], \tag{2}$$

for all $U(S) \in L^2(\mathcal{G})$. In particular, $\mathrm{Var}(W_{\pi'}U) \geq \mathrm{Var}(m^{\star}(S)U)$ for any $U(S) \in L^2(\mathcal{G})$.

Pythagoras in Hilbert spaces. Let $L_0^2(P)$ be the mean-zero Hilbert space with inner product $\langle f, g \rangle = \mathbb{E}[fg]$. For a nonempty closed convex set $\mathcal{C} \subset L_0^2$ and any $z \in L_0^2$, denote by $\Pi_{\mathcal{C}}(z)$ the metric projection. If $\mathcal{C}_1 \subseteq \mathcal{C}_2$ then $\mathrm{dist}(z, \mathcal{C}_2) \leq \mathrm{dist}(z, \mathcal{C}_1)$.

Algorithm 4 DR-CPO: sequence-aware doubly robust estimator (cross-fitted)

Inputs: $\hat{W}_{\pi'}$; R and R^{OOF} from Alg. 1; folds $F(\cdot)$
Outputs: $\widehat{V}_{\mathrm{DR}}$ and IF ϕ^{DR}
for $k = 1$ to K **do**
  Train $\hat{q}^{(-k)}(x, a) \approx \mathbb{E}[R \mid x, a]$ on $\{i : F(i) \neq k\}$; predict OOF $\hat{q}_i^{\mathrm{OOF}} = \hat{q}^{(-k)}(X_i, A_i)$ for $i \in I_k$
  Approximate $\hat{g}_{\pi'}^{(-k)}(x) = \mathbb{E}_{A \sim \pi'(\cdot|x)}[\hat{q}^{(-k)}(x, A)]$ via one rollout $A' \sim \pi'(\cdot \mid X)$ and optional smoothing; obtain OOF $g_{\pi',i}^{\mathrm{OOF}}$ for $i \in I_k$
end for
$\widehat{V}_{\mathrm{DR}} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \Big\{ \hat{g}_{\pi'}(X_i) + \hat{W}_{\pi',i}\big(R_i - \hat{q}(X_i, A_i)\big) \Big\}$
$\phi_i^{\mathrm{DR}} = \hat{g}_{\pi'}(X_i) + \hat{W}_{\pi',i}\big(R_i^{\mathrm{OOF}} - \hat{q}_i^{\mathrm{OOF}}\big) - \widehat{V}_{\mathrm{DR}}$

Algorithm 5 TR-CPO: targeted & retargeted DR (optional triply robust add-on)

Inputs: Same as Alg. 4; link (*identity* if R is unbounded; *logit* if $R \in [0, 1]$)
Targeting (solve EIF moment).
for $k = 1$ to K **do**
  With clever covariate $H(X, A) = \hat{W}_{\pi'}$, fit $\varepsilon^{(-k)}$ on $\{i : F(i) \neq k\}$ so that

$$n^{-1} \sum_{i \in I_k} \hat{W}_{\pi',i}\big(R_i^{\mathrm{OOF}} - \hat{q}_{\hat{\varepsilon}}^{(-k)}(X_i, A_i)\big) \approx 0$$

  under the chosen link.
end for
Set $\hat{q}_\star$ to the pooled targeted fit; recompute $\hat{g}_{\pi'}$ accordingly
Retargeting (control variate). $Z_i = \hat{W}_{\pi',i}\big(R_i^{\mathrm{OOF}} - \hat{q}_\star^{\mathrm{OOF}}(X_i, A_i)\big)$; $\hat{\gamma} = \mathrm{cov}(Z, \hat{W}_{\pi'} - 1)/\mathrm{Var}(\hat{W}_{\pi'} - 1)$
$\widehat{V}_{\mathrm{TR}} = \dfrac{1}{n} \displaystyle\sum_{i} \Big\{ \hat{g}_{\pi'}(X_i) + Z_i - \hat{\gamma}\,(\hat{W}_{\pi',i} - 1) \Big\}$, $\phi_i^{\mathrm{TR}} = \hat{g}_{\pi'}(X_i) + Z_i - \hat{\gamma}(\hat{W}_{\pi',i} - 1) - \widehat{V}_{\mathrm{TR}}$
Optional triply robust term (MAR in S). If a fold-honest label propensity $\hat{\pi}_L(S) \in [\varepsilon, 1]$ is available, add a mean-zero residual×residual term on labeled rows to further damp first-order calibration error (see Appendix C).

C.2 Isotonic regression: mean preservation and majorization

Lemma 1 (Mean preservation; PAVA). *Let $\hat{f} \in \arg\min_{f \in \mathcal{M}_\uparrow} \sum_{i \in I} (y_i - f(s_i))^2$ be the isotonic fit (PAVA) on indices I. Then $\frac{1}{|I|} \sum_{i \in I} \hat{f}(s_i) = \frac{1}{|I|} \sum_{i \in I} y_i$.*

Lemma 2 (Dispersion reduction by majorization). *After sorting by s, the isotonic fitted vector $\hat{u}$ is a mean-preserving* adjacent pooling *of y; hence for any convex ϕ, $\sum_i \phi(\hat{u}_i) \leq \sum_i \phi(y_i)$ (Hardy et al., 1952; Marshall et al., 2011). In particular, with sample mean one, $\mathrm{Var}_n(\hat{u}) \leq \mathrm{Var}_n(y)$ and $\mathrm{ESS}(\hat{u}) \geq \mathrm{ESS}(y)$.*

Proofs are standard; see Ayer et al. (1955); Barlow et al. (1972); Robertson et al. (1988); Banerjee (2001).

C.3 Proof of Theorem 1 (surrogate EIF & variance drop)

Let $R^\star = \mathbb{E}[Y \mid S]$ and $m^\star(S) = \mathbb{E}[W_{\pi'} \mid S]$. Under mean-sufficiency $\mathbb{E}[Y \mid X, A, S] = R^\star(S)$,

$$V(\pi') = \mathbb{E}\big[W_{\pi'} R^\star\big] = \mathbb{E}\big[m^\star(S) R^\star(S)\big]. \tag{3}$$

Standard semiparametric calculations (projecting the unconstrained score onto the tangent space of the surrogate model) yield

$$\phi_{\mathrm{sur}}(O; \pi') = g^\star_{\pi',R}(X) + m^\star(S)\big(R^\star - q^\star_R(X, A)\big) - V(\pi'), \tag{4}$$

Algorithm 6 IF-STACK: variance-optimal convex ensembling in IF space

Inputs: Candidates $\{\widehat{V}^{(e)}, \phi^{(e)}\}_{e\in\mathcal{E}}$ (centered IFs; same folds); ridge λ
Outputs: $\widehat{V}^{(\hat{\alpha})}$ and $\phi^{(\hat{\alpha})}$
Form $\Phi = [\phi^{(e)}]_{e\in\mathcal{E}}$; $\hat{\Sigma} = (1/n)\Phi^\top\Phi + \lambda I$
Solve $\hat{\alpha} \in \arg\min_{\alpha\in\Delta} \alpha^\top\hat{\Sigma}\,\alpha$
$\widehat{V}^{(\hat{\alpha})} = \sum_{e\in\mathcal{E}} \hat{\alpha}_e\,\widehat{V}^{(e)}, \quad \phi^{(\hat{\alpha})} = \sum_{e\in\mathcal{E}} \hat{\alpha}_e\,\phi^{(e)}$
Optional outer split: learn $\hat{\alpha}$ on one half; apply to the other to reduce selection optimism
Support note (Carathéodory). If $\mathrm{rank}(\Phi) = r$, the variance-optimal stack uses at most $r+1$ base estimators.

Algorithm 7 OUA jackknife: oracle–uncertainty–aware variance addition

Inputs: Oracle folds $\{O_k\}_{k=1}^K$; end-to-end estimator $\widehat{V}(\cdot)$
Outputs: $\widehat{\mathrm{Var}}_{\mathrm{oracle}}$ and $\widehat{\mathrm{Var}}_{\mathrm{total}} = \widehat{\mathrm{Var}}_{\mathrm{main}} + \widehat{\mathrm{Var}}_{\mathrm{oracle}}$
for $k = 1$ to K **do**
    Refit AUTOCAL-R on $O \setminus O_k$; recompute $R^{(-k)}$ and all downstream nuisances & weights; run the full pipeline to get $\widehat{V}^{(-k)}$
end for
$\bar{V} = \frac{1}{K}\sum_k \widehat{V}^{(-k)}, \quad \widehat{\mathrm{Var}}_{\mathrm{oracle}} = \frac{K-1}{K}\sum_k \big(\widehat{V}^{(-k)} - \bar{V}\big)^2$
Return: $\widehat{\mathrm{Var}}_{\mathrm{total}} = \widehat{\mathrm{Var}}_{\mathrm{main}} + \widehat{\mathrm{Var}}_{\mathrm{oracle}}$

with $q_R^\star(x,a) = \mathbb{E}[R^\star \mid x, a]$ and $g_{\pi',R}^\star(x) = \mathbb{E}_{A\sim\pi'(\cdot|x)}[q_R^\star(x,A)]$ (Bickel et al., 1993; van der Vaart & Wellner, 2000). Since $m^\star$ is the L^2 projection of $W_{\pi'}$ onto $L^2(\sigma(S))$, Pythagoras (or (2)) implies $\mathrm{Var}(\phi_{\mathrm{sur}}) \le \mathrm{Var}(\phi_{\mathrm{uncon}})$, strictly unless $W_{\pi'} \in L^2(\sigma(S))$ and $R^\star$ is degenerate.

C.4 PROOF OF THEOREM 2 (KNOWLEDGE–RIESZ / CKP)

Let $I(P) = \phi^\star + T(P)^\perp$ be the affine class of IFs in a baseline model and $\mathcal{C} \subset L_0^2$ be nonempty, closed, convex. Define $I_{\mathcal{C}}(P) = I(P) \cap \mathcal{C}$ and $\phi_{\mathcal{C}} = \Pi_{I_{\mathcal{C}}(P)}(0) = \arg\min_{\phi\in I_{\mathcal{C}}(P)} \mathbb{E}[\phi^2]$. Then

$$\|\phi_{\mathcal{C}}\|_2^2 = \mathrm{dist}^2(0, I_{\mathcal{C}}(P)) \le \mathrm{dist}^2(0, I(P)) = \|\phi^\star\|_2^2,$$

with equality iff $I_{\mathcal{C}}(P)$ already contains $\phi^\star$. Monotonicity for $\mathcal{C}_1 \subseteq \mathcal{C}_2$ is immediate. For attainability, replace nuisances in a one-step/TMLE update by their projections into $\mathcal{C}$; cross-fitting ensures the empirical score equations hold in the restricted model and the remainder is $o_p(n^{-1/2})$ (Bickel et al., 1993; van der Vaart & Wellner, 2000; van der Laan & Rose, 2011).

C.5 PROOF OF COROLLARY 1

If $\sigma(S_2) \subseteq \sigma(S_1)$, then $L^2(\sigma(S_2)) \subseteq L^2(\sigma(S_1))$. Hence the feasible knowledge set for conditioning, $\mathcal{C}(S) = \{f : f = \mathbb{E}[f \mid S]\}$, satisfies $\mathcal{C}(S_2) \subseteq \mathcal{C}(S_1)$. Applying Theorem 2 with $I(P)$ fixed yields $\|\phi_{\mathcal{C}(S_1)}\|_2^2 \le \|\phi_{\mathcal{C}(S_2)}\|_2^2$, i.e., $\mathrm{Var}(\phi_{\mathrm{sur}}(S_1)) \le \mathrm{Var}(\phi_{\mathrm{sur}}(S_2))$. Strictness fails only if the finer knowledge already holds, i.e., if $W_{\pi'}$ is $\sigma(S_2)$-measurable and $R^\star$ is degenerate.

C.6 PROOF OF PROPOSITION 1 (CAL-IPS)

Write

$$\widehat{V}_{\mathrm{IPS}} - V(\pi') = (P_n - P)\big[m^\star(S)R^\star\big] + P\big[(\hat{W}_{\pi'} - m^\star)R^\star\big] + P\big[m^\star(\hat{f}(T(S)) - R^\star)\big] + \mathrm{rem},$$

where rem collects second-order sample-splitting terms. The empirical process term is $O_p(n^{-1/2})$; the second and third vanish by L^2-consistency of SIMCal-W and AutoCal-R (monotone or two-stage) and Cauchy–Schwarz; the remainder is $o_p(1)$ by cross-fitting. Finite-sample dispersion control follows from Lemma 2 and, if used, the blend cap $\rho \ge 1$; thus $\mathrm{ESS}(\hat{W}_{\pi'}) \ge \mathrm{ESS}(W_{\pi'}^{\mathrm{m1}})$ deterministically.

C.7 Proof of Theorem 3 (DR/TR-CPO $\sqrt{n}$ limits)

With cross-fitted nuisances and R^{OOF} along the IF path,

$$\widehat{V}_{\mathrm{DR}} - V(\pi') = (P_n - P)[\phi_{\mathrm{sur}}] + P[(\hat{W}_{\pi'} - m^\star)\{q_R^\star - \hat{q}\}] + o_p(n^{-1/2}).$$

The second term is $o_p(n^{-1/2})$ by the one-of-two product-rate condition $\|\hat{W}_{\pi'} - m^\star\|_2 \cdot \|\hat{q} - q_R^\star\|_2 = o_p(n^{-1/2})$ and cross-fitting; the central limit theorem yields the limit variance $\mathrm{Var}(\phi_{\mathrm{sur}})$. TR-CPO adds (i) a one-dimensional fluctuation that solves the EIF moment in finite samples and (ii) a mean-zero control variate anchored at $(\hat{W}_{\pi'} - 1)$; both are second order under the same rate condition, leaving the limit unchanged.

C.8 Proof of Theorem 4 (budgeted bound)

Let $\mathcal{W}_\rho = \{m : \mathbb{E}[m] = 1,\ m \uparrow S,\ \mathbb{E}[(m-1)^2] \le \rho\,\mathbb{E}[(w^\star - 1)^2]\}$. Intersecting the surrogate tangent space with the linear span induced by $m \in \mathcal{W}_\rho$ replaces $m^\star$ by its $L^2(P_S)$ projection $m_\rho^\star = \Pi_{\mathcal{W}_\rho}(m^\star)$ in ϕ_{sur}, giving $\phi^{(\rho)}$. Monotonicity in ρ follows from nested convex sets $\mathcal{W}_{\rho_1} \subseteq \mathcal{W}_{\rho_2}$ for $\rho_1 \le \rho_2$, and $\lim_{\rho\to\infty} \phi^{(\rho)} = \phi_{\mathrm{sur}}$. If SIMCal-W converges to $m_\rho^\star$ (with the same cap), TR-CPO attains $\mathrm{Var}(\phi^{(\rho)})$ by the same one-of-two rate argument.

C.9 Proof of Theorem 5 (IF-space stacking) and Corollary 2

Let $\{\widehat{V}^{(e)}\}$ be regular and asymptotically linear with centered IFs $\{\phi^{(e)}\}$. Set $\Phi = [\phi^{(e)}]$ and $\hat{\Sigma} = (1/n)\Phi^\top\Phi + \lambda I$. A uniform law of large numbers on the simplex Δ yields $\hat{\Sigma} \to \Sigma$ uniformly; by argmin continuity, $\hat{\alpha} \to \alpha^\star \in \arg\min_{\alpha\in\Delta} \alpha^\top\Sigma\,\alpha$. Hence $\widehat{V}^{(\hat{\alpha})}$ is asymptotically linear with IF $\phi^{(\alpha^\star)} = \sum_e \alpha_e^\star \phi^{(e)}$ and variance $\min_{\alpha\in\Delta} \alpha^\top\Sigma\,\alpha \le \min_e \Sigma_{ee}$. For Corollary 2: if $\mathrm{rank}(\Phi) = r$, then the feasible IF combinations lie in an r-dimensional affine subspace; by Carathéodory's theorem, any point in $\mathrm{conv}\{\phi^{(e)}\}$ admits a representation using at most $r+1$ extreme points.

C.10 Proof of Proposition 2 (OUA jackknife)

Let $\widehat{V}(\hat{f})$ be a regular estimator that depends on f only through $R = \hat{f}(T(S))$, with $f \mapsto \widehat{V}(f)$ Hadamard-differentiable at $f^\star$ in $L^2(P_S)$. Using a delta-method expansion and cross-fitting (so that oracle folds are asymptotically independent of the IF path), the delete-one-oracle-fold jackknife (Bickel et al., 1993; Künsch, 1989; Politis & Romano, 1994) consistently estimates the variance contribution from first-stage calibration. Therefore $\widehat{\mathrm{Var}}_{\mathrm{total}} = \widehat{\mathrm{Var}}_{\mathrm{main}} + \widehat{\mathrm{Var}}_{\mathrm{oracle}}$ is consistent for $\mathrm{Var}(\widehat{V})$.

C.11 Auxiliary lemmas used in the main proofs

Lemma 3 (OOF mean preservation for AutoCal-R)**.** *Let K be fixed and let $R_i^{\mathrm{OOF}} = \hat{f}^{(-F(i))}(T(S_i))$ be OOF predictions from AutoCal-R (either mode). Then $P_n[R^{\mathrm{OOF}}] - P_n[Y] = o_p(1)$ and $P[R^{\mathrm{OOF}} - R^\star] = o_p(1)$ under $L^2(P_S)$-consistency of $\hat{f}$.*

Lemma 4 (Second-order remainder for DR with cross-fitting)**.** *Let $\widehat{V}_{\mathrm{DR}}$ be DR-CPO with cross-fitted $(\hat{q}, \hat{g}_{\pi'})$ and calibrated $\hat{W}_{\pi'}$. Then*

$$\widehat{V}_{\mathrm{DR}} - V(\pi') - (P_n - P)\phi_{\mathrm{sur}} = P\big[(\hat{W}_{\pi'} - m^\star)(q_R^\star - \hat{q})\big] + o_p(n^{-1/2}),$$

and the bracketed term is $o_p(n^{-1/2})$ under $\|\hat{W}_{\pi'} - m^\star\|_2 \cdot \|\hat{q} - q_R^\star\|_2 = o_p(n^{-1/2})$.

Lemma 5 (Guard stability)**.** *Let W^{stack} be the OOF-stacked candidate and $W_{\pi'}^{\mathrm{m1}}$ the mean-one baseline. For $\rho \ge 1$, define $\alpha = \min\{1, \rho\ \mathrm{Var}(W_{\pi'}^{\mathrm{m1}})/\mathrm{Var}(W^{\mathrm{stack}})\}$ and $W^{\mathrm{blend}} = 1 + \alpha(W^{\mathrm{stack}} - 1)$. Then $\mathrm{Var}(W^{\mathrm{blend}}) \le \rho\ \mathrm{Var}(W_{\pi'}^{\mathrm{m1}})$ and the subsequent mean-one isotonic projection cannot increase empirical variance by Lemma 2.*

Lemma 6 (Firm non-expansiveness of projections)**.** *Metric projections onto closed convex sets in Hilbert spaces are firmly non-expansive: $\|\Pi_{\mathcal{C}}(x) - \Pi_{\mathcal{C}}(y)\|^2 \le \langle \Pi_{\mathcal{C}}(x) - \Pi_{\mathcal{C}}(y),\, x - y\rangle$. Hence compositions of DbP modules (reward, weight, IF-space projections) are non-expansive. See, e.g., Bauschke & Combettes (2017, Prop. 4.16).*

Remarks on dependence. If logs exhibit serial or cluster dependence, the IF CLTs can be replaced by block/stationary bootstrap arguments; our reported intervals can include a dependence-robust alternative (App. A.8).

C.12 What the bounds do not include

The surrogate and budgeted information bounds describe *model* limits: they do not include (i) finite-sample dispersion control from the sample cap (we encode it population-wise via ρ) or (ii) the oracle first-stage uncertainty (added separately by OUA).

C.13 Citations for technical ingredients

Semiparametric efficiency and one-step/TMLE: Bickel et al. (1993); van der Vaart & Wellner (2000); van der Laan & Rose (2011); Tsiatis (2006).
Isotonic regression / PAVA and order-restricted inference: Ayer et al. (1955); Barlow et al. (1972); Robertson et al. (1988); Banerjee (2001).
Majorization theory: Hardy et al. (1952); Marshall et al. (2011).
Calibration for OPE: Kallus & Mao (2022); Fong & Kennedy (2022); van der Laan et al. (2025a;b).
Jackknife/bootstraps for dependence: Künsch (1989); Politis & Romano (1994).
Projections and non-expansive maps: Bauschke & Combettes (2017).

D Diagnostics, Gates, and Reporting (details)

This appendix formalizes the diagnostics used in CJE, the associated *ship/stop* gates, and the reporting ledger. The main text shows a compact panel per policy; here we provide precise formulas, defaults, and recommended thresholds. Unless stated otherwise, expectations and variances are empirical over the evaluation cohort (global SNIPS normalization).

D.1 Weight behavior & overlap

Effective sample size (ESS). For nonnegative weights,

$$\mathrm{ESS}(W) = \frac{\left(\sum_i W_i\right)^2}{\sum_i W_i^2}.$$

Under global mean-one normalization (SNIPS), $\sum_i W_i = n$, so

$$\frac{\mathrm{ESS}(W)}{n} = \frac{1}{1+\mathrm{CV}^2(W)}, \qquad \mathrm{CV}^2(W) = \mathrm{Var}(W) \text{ when } \mathbb{E}[W] = 1.$$

Report the *ESS fraction* $\mathrm{ESS}(W)/n$ and the multiplicative *uplift* $\mathrm{ESS}(\hat{W}_{\pi'})/\mathrm{ESS}(W^{\mathrm{m1}}_{\pi'})$.

Max-weight share. $\max_i W_i \big/ \sum_j W_j$ flags single-row dominance; display alongside the empirical 99.5th percentile of W.

Tail index (Hill) and CCDF. For the top-k order statistics $W_{(1)} \geq \cdots \geq W_{(k)}$,

$$\hat{\alpha}^{-1}(k) = \frac{1}{k}\sum_{j=1}^{k} \log\left(\frac{W_{(j)}}{W_{(k)}}\right), \qquad k \in \mathcal{K},$$

swept over a stability grid $\mathcal{K}$ (e.g., 1–5% of n). Plot $\hat{\alpha}(k)$ with a band over the plateau region (median and IQR over $\mathcal{K}$), and the empirical CCDF of W on a log–log scale.

Overlap in judge space. Let $p_{S|\pi_0}$ and $p_{S|\pi'}$ denote the (binned) densities of S under π_0 and π':

$$A_B = \sum_b \sqrt{p_b(\pi_0)\, p_b(\pi')}, \qquad D_B = -\log A_B.$$

Overlay an S-binned heatmap of $\log W$ to localize regions of poor overlap.

D.2 JUDGE CALIBRATION, COVERAGE, AND DRIFT

Reliability diagram. Partition S into B bins; for bin b, plot the bin mean of $R = \hat{f}(T(S))$ against the oracle mean of Y, with 95% binomial intervals. Report a Brier-style *reliability* term and the OOF RMSE, with a 1-SE model-selection overlay (monotone vs. two-stage).

Coverage badge. Estimate the fraction of evaluation mass outside the oracle S range:

$$\texttt{OutOfRange} = \widehat{\Pr}_{\pi'}\big(S < S^{\text{orc}}_{\min} \text{ or } S > S^{\text{orc}}_{\max}\big).$$

Also report *boundary flatness* (slope of $\hat{f}$ in the lowest/highest oracle decile). Large `OutOfRange` together with flat boundaries triggers LIMITED CALIBRATION SUPPORT and the REFUSE-LEVEL gate.

Rank drift (optional anchor). Given a fixed anchor set of (X, A) pairs scored over time, compute Kendall's τ between historical and current judge rankings with a permutation p-value. Change detection on residuals can be monitored via CUSUM/EWMA with FDR control across anchors.

D.3 DR ORTHOGONALITY AND DECOMPOSITION

Orthogonality score. Let $U_i = \hat{W}_{\pi',i}\big(R_i^{\text{OOF}} - \hat{q}^{\text{OOF}}(X_i, A_i)\big)$ and $\bar{U} = n^{-1}\sum_i U_i$. Form a Wald CI for $\bar{U}$ using the standard error $\sqrt{\widehat{\text{Var}}(U)/n}$ (or a cluster-/block-robust analogue). Report $\bar{U}$ and its CI; near-zero indicates successful orthogonality.

DM–IPS decomposition. Display $\widehat{V}_{\text{DM}} = n^{-1}\sum_i \hat{g}_{\pi'}(X_i)$ and $\widehat{V}_{\text{Aug}} = n^{-1}\sum_i U_i$, with CIs and the empirical correlation between their per-row contributions.

D.4 UNCERTAINTY: IF VARIANCE AND OUA ADDITION

For any estimator with centered IF contributions $\{\phi_i\}_{i=1}^n$,

$$\widehat{\text{Var}}_{\text{main}} = \frac{1}{n}\,\widehat{\text{Var}}(\phi_i), \qquad \widehat{\text{Var}}_{\text{total}} = \widehat{\text{Var}}_{\text{main}} + \widehat{\text{Var}}_{\text{oracle}},$$

with $\widehat{\text{Var}}_{\text{oracle}}$ from the oracle jackknife (App. A.7). Report the *oracle share* $\widehat{\text{Var}}_{\text{oracle}}/\widehat{\text{Var}}_{\text{total}}$ and, optionally, a dependence-robust alternative (below).

Dependence-robust SEs. When time/cluster dependence is suspected, also report: (i) *cluster-robust* sandwich SEs when a cluster id (e.g., session/user) is available; and (ii) *block/stationary bootstrap* intervals (block length chosen by a simple variance-stability sweep) (Künsch, 1989; Politis & Romano, 1994).

D.5 MULTIPLICITY FOR MANY-POLICY COMPARISONS

For contrasts $\Delta_p = \widehat{V}(\pi'_p) - \widehat{V}(\pi^\star)$, compute Wald p-values and apply BH at level $q \in [0.05, 0.2]$; BY can be used under strong dependence. Provide a *pairwise win matrix* with FDR marks and Kendall's τ over policy means (all policies).

D.6 GATES: THRESHOLDS AND ACTIONS

REFUSE-LEVEL procedure. When IDENTIFICATION fails: (i) gray-out level estimates; (ii) highlight `OutOfRange` and boundary flatness; (iii) report rank-only conclusions with conservative relative CIs; (iv) recommend targeted labeling in uncovered S regions.

D.7 PLANNER: MDE AND LABEL/LOG BUDGETS

Given two independent estimates with equal SE $\widehat{\text{SE}}$, the two-sided 95% test at 80% power has

$$\text{MDE}_{80\%} = (z_{0.8} + z_{0.975})\sqrt{2}\,\widehat{\text{SE}}.$$

Table 3: Default gates (suggested; tighten for high-stakes launches).

Gate	Default	Action if failed
OVERLAP	ESS/$n \geq 0.30$; Hill $\hat{\alpha} \geq 2$; $A_B \geq 0.85$	Use overlap weights or cohort restriction; report with warning if $\hat{\alpha} \in [1, 2)$; do not ship offline conclusions if $\hat{\alpha} < 1$
JUDGE	Reliability band covers diagonal at knots; no persistent drift alarms	Refresh/extend oracle slice; switch to two-stage index; re-validate
IDENTIFICATION	`OutOfRange` $\leq \eta$ (default η=5%) or non-flat boundaries	Flag LIMITED CALIBRATION SUPPORT; set REFUSE-LEVEL: report rankings + partial-ID only
DR	Orthogonality CI includes 0; no NaNs; residual tails acceptable	Strengthen nuisances/cross-fitting; fall back to stabilized IPS as a diagnostic
MULTIPLICITY	FDR control applied when $\lvert\Pi'\rvert > 5$	Report adjusted p-values; avoid uncorrected winner claims
CAP	Guard rarely engaged; CI width not sensitive to ρ	If guard active on $> 50\%$ folds or sensitivity high, show cap curve and prefer overlap weights/restriction

We tabulate $\widehat{\mathrm{SE}}$ versus $(n,\ m/n)$ (labels per log) using ***Stacked-DR*** with OUA and annotate "iso-cost" lines for the label budget.

D.8 REPORTING LEDGER (PER POLICY/COHORT)

Persist: (i) calibrator mode, OOF risk by tertiles, knots/levels (hash); (ii) SIMCal–W maps, stacking weights $\hat{\beta}$, guard ρ and blend α; (iii) ESS fraction, max-weight share, Hill index band, A_B; (iv) DR orthogonality score and CI; DM–IPS split; (v) OUA trace $\{\widehat{V}^{(-k)}\}$ and variance breakdown; (vi) filter counts (e.g., TF gaps) and an inclusion manifest of x-ids; (vii) multiplicity control (family, q, adjusted p).

D.9 VISUALIZATION PRIMITIVES (FOR REPRODUCIBLE PANELS)

- **ESS/tails strip:** bars for ESS fraction (baseline vs. SIMCal–W); dot for max-weight share; Hill band.
- **S-overlap heatmap:** density of S under π_0 vs. π' with overlaid $\log W$; annotate A_B.
- **Reliability panel:** bin means of (R, Y) with 95% CIs; mode card (monotone vs. two-stage; OOF RMSE).
- **Orthogonality panel:** point/CI for $\bar{U}$; DM–IPS bars with CIs and correlation.
- **Uncertainty ring:** pie of $\widehat{\mathrm{Var}}_{\mathrm{oracle}}/\widehat{\mathrm{Var}}_{\mathrm{total}}$.

D.10 OPTIONAL: DEPENDENCE–ROBUST IMPLEMENTATION DETAILS

Cluster-robust SEs: if a cluster id $c(i)$ is available, $\widehat{\mathrm{Var}}_{\mathrm{CR}} = n^{-2} \sum_c \big(\sum_{i \in c} \phi_i\big)\big(\sum_{i \in c} \phi_i\big)^\top$, with finite-sample correction. *Stationary bootstrap:* sample blocks of geometric length $\ell \sim \mathrm{Geom}(p)$ glued to length n; form the bootstrap distribution of $\hat{\psi}$ (or of $n^{-1/2} \sum \phi_i$) and report percentile or t-based bands.

D.11 COMPACT GATE PSEUDO-LOGIC

E IMPLEMENTATION, ENGINEERING, AND REPRODUCIBILITY

This appendix enumerates the concrete artifacts needed to reproduce CJE end–to–end: a minimal logging schema, a teacher–forcing (TF) contract with conformance checks, fold construction, numerics, persisted outputs, and a lightweight resource model. No packages beyond the ICLR style file are required.

E.1 MINIMAL LOGGING SCHEMA (STORAGE–AGNOSTIC)

Each row corresponds to one prompt–continuation pair under the fixed logger π_0. We persist only what is necessary to reconstruct SNIPS/IPS weights and judge scores.

Algorithm 8 Gate logic (per policy)

```
Compute weight/tail metrics: ESS fraction, max-share, Hill band; compute A_B; judge reliability/coverage; orthogonality score; OUA share
if ESS/n < 0.30 or median Hill< 2 or A_B < 0.85 then
    Flag OVERLAP (warn; restrict or use overlap weights)
end if
if OutOfRange¿η and boundary slopes ≈ 0 then
    REFUSE-LEVEL ← TRUE
end if
if Orthogonality CI excludes 0 then
    Flag DR; strengthen nuisances/cross-fitting
end if
if Cap engaged on > 50% folds or CI sensitivity to ρ high then
    Show cap–sensitivity; prefer overlap weights/restriction
end if
Apply multiplicity control (BH/BY) when |Π'| > 5
```

Table 4: Columns required for CJE. Columnar formats (Parquet) are convenient but not required.

Field	Type	Description
`x_id`	string	Stable identifier (hash of normalized prompt + cohort)
`prompt`	bytes/string	Canonicalized X (tokenizer + normalization recorded)
`continuation`	bytes/string	Realized A under π_0 (full sequence)
`tokens`	int[]	Token ids for A under each model's TF tokenizer
`logp_pi0`	float[]	Per–token $\log p_{\pi_0}(a_t \mid h_t)$ across A
`judge_S`	float/json	Scalar judge score $S = s(X, A)$ (or struct of sub–scores)
`judge_cfg`	json	Judge rubric, decoding params, model snapshot hash
`run_cfg`	json	π_0 engine tag, decoding params, checkpoint hash, seed
`fold_id`	int	Deterministic fold ($F(x\_id)$; see §E.3)
`cohort`	string	Optional slice label (time window, traffic source, etc.)

TF cache (per target π'). A separate table stores, for each (x_id, π'): `logp_pi_prime`, `logW`$=\log p_{\pi'} - \log p_{\pi_0}$, and

$$W^{\mathrm{m1}}_{\pi'} \;=\; \exp\Big(\texttt{logW} - \mathrm{logsumexp}(\texttt{logW}) + \log n\Big),$$

i.e., a single global denominator that enforces sample–mean–one. Rows with missing/invalid TF are filtered and recorded in a ledger.

E.2 Teacher forcing: contract and conformance

We require a *single–call, chat–native* TF API that returns per–token and summed $\log p_\pi(A \mid X)$ under a fixed template, tokenizer, and snapshot. Client–side checks:

- **Determinism.** k identical calls for the same $(X, A, \pi@\text{SNAPSHOT}, \text{template})$ must be bit–identical (tolerance $< 10^{-7}$).
- **Additivity.** Also return $\log p_\pi(X)$ and $\log p_\pi(X{+}A)$ and verify $\log p_\pi(X{+}A) \approx \log p_\pi(X) + \log p_\pi(A \mid X) \leq \log p_\pi(X)$. Violation ⇒ discard row (and log it).
- **Template/tokenizer provenance.** Return immutable hashes; reject moving aliases.
- **Masking.** If safety masks are applied, return mask bits and a `renormalized` flag; prefer an evaluation–only path without hidden renormalization.

Conformance snippet (pseudo).

```
lp = TF(model_id, template_id, X, A)
assert same_bits(lp.sum, sum(lp.per_token), tol)
```

```
lpX, lpXA = TF_logp(X), TF_logp(X+A)
assert abs(lpXA - (lpX + lp.sum)) < eps and lp.sum <= 0
```

The supplement's README repeats this minimal additivity check.

E.3 Folds and cross–fitting

We use $K{=}5$ folds by default. The fold map $F(i)$ is a stable hash of `x_id` modulo K, ensuring that all modules (AutoCal–R, SIMCal–W, DR nuisances) share identical OOF boundaries. Oracle folds are derived by intersecting $F(i)$ with $L_i{=}1$. We may serialize $F(i)$ internally for determinism; *the public supplement publishes only the hash rule* $F(x\_id){=}\mathrm{hash}(x\_id) \bmod 5$ *(no fold map file).*

E.4 Numerics and stability

- **Ratios in log–space.** Keep $\log W$ until forming the global mean–one normalization; use a single `logsumexp` for the denominator.
- **Center residuals.** For stacking objectives and covariances, center T and drop NaNs/inf at ingestion.
- **Variance estimates.** Use Welford's online formulas for high dynamic range; add a tiny ridge ($\lambda \in [10^{-10}, 10^{-6}]$) to covariance matrices.
- **PAVA.** Run once per fold after sorting by S; *enforce mean–one via an additive shift (translation), not multiplicative rescaling*. This preserves monotonicity and avoids tail distortion.
- **Guard (relative cap).** Default $\rho{=}1$; compute $\alpha = \min\{1,\ \rho\ \mathrm{Var}(W^{\mathrm{m1}})/\ \mathrm{Var}(W^{\mathrm{stack}})\}$; blend and then (re)project/translate to mean–one. Persist whether the guard engaged.

Reference code note. For simplicity, the supplement's `simcal.py` uses *global* isotonic fits and—*unlike the main pipeline*—computes the stacking covariance on those same *in-sample* fits (no OOF in the tiny reference).

E.5 Persisted artifacts (per policy/cohort)

- **Calibrator:** mode (monotone vs. two–stage), OOF RMSE (overall + tertiles), knots/levels (hash), OOF vs. pooled predictions.
- **Weights:** isotonic merge metadata, S orientation (up/down), *stacking weights* $\hat{\beta}$ *including the identity/baseline candidate*, guard ρ and blend α, final mean–one check.
- **Estimators:** point estimates, centered IF vectors' hashes, $\widehat{\mathrm{Var}}_{\mathrm{main}}$, orthogonality score and CI, dependence–robust SEs (if used).
- **OUA:** $\{\widehat{V}^{(-k)}\}_{k=1}^{K}$, $\widehat{\mathrm{Var}}_{\mathrm{oracle}}$, $\widehat{\mathrm{Var}}_{\mathrm{total}}$.
- **Diagnostics:** ESS fraction, max–weight share, Hill band, S–overlap (A_B), coverage badge, gate statuses.
- **Ledger:** counts by filter reason (TF gaps, moderation, timeouts), `x_id` inclusion manifest.

E.6 Reference run order (pseudocode)

```
# 0) Build TF cache for each pi' (one pass per policy)
build_tf_cache --policies <list> --dataset logs.parquet --out tf_cache.parquet

# 1) Reward calibration (cross-fitted; auto monotone vs two-stage)
autocal_r --oracle oracle.parquet --folds 5 --out rewards.parquet

# 2) SIMCal-W per policy (OOF project->stack->cap->translate-to-mean-one)
simcal_w --tf-cache tf_cache.parquet --scores S.parquet --rho 1.0 --folds 5 \
         --out weights.parquet

# 3) Estimation + IFs (Cal-IPS / OC-IPS / DR-CPO / TR-CPO)
estimate --rewards rewards.parquet --weights weights.parquet --folds 5 \
```

```
        --out estimates.parquet

# 4) IF-Stack (optional)
stack --estimates estimates.parquet --out stacked.parquet

# 5) OUA jackknife
oua --oracle-folds 5 --pipeline-config cfg.yaml --out variance.parquet

# 6) Report (diagnostics, gates, CIs)
report --inputs *.parquet --figs figs/ --out report.html
```

E.7 Compute and resource model

Let n be prompts, $\bar{T}$ the mean continuation length, and $|\Pi'|$ the number of candidate policies.

- **TF cache.** $O(|\Pi'|\, n\, \bar{T})$ forward tokens; microbatch by length; near–linear scaling across GPUs.
- **SIMCal–W.** $O(n \log n)$ for sort $+ O(n)$ for PAVA per fold; covariance/stacking are tiny (3×3).
- **DR/TR–CPO.** If $\hat{g}_{\pi'}$ uses one rollout per (X, π'), add $O(|\Pi'|\, n\, \bar{T}')$ tokens once; a light smoother amortizes Monte Carlo noise.
- **OUA.** K refits of AutoCal–R and re–runs of the pipeline; cache features to avoid recomputation.

E.8 Determinism, versioning, and privacy

Determinism: fix seeds at engine, dataloader, and sampler; record random states in `run_cfg`; serialize fold maps (internal).
Versioning: record immutable hashes for model weights, tokenizer, and template; pin checkpoints.
Privacy: encrypt prompts/continuations at rest; hash `x_id` with salt; public artifacts include only aggregates/diagnostics and redacted IDs.

E.9 What to publish with the paper (micro–supp)

- `README.md` (10–12 lines): what's inside; how it ties to the paper; the fold rule $F(x\_id)$=hash(x_id) mod 5; the TF additivity snippet; and a SIMCal usage snippet. *Note: isotonic fits are global for simplicity and, in this tiny reference, stacking covariance is computed* in–sample *(no OOF).*
- `configs/ablation_config.yaml` and `configs/policies.yaml`: sanitized examples (no vendors/paths).
- `code/simcal.py`: reference SIMCal–W with four fixes baked in: (i) *mean–one by translation* after isotonic (not scaling); (ii) *relative* ρ–guard $\mathrm{Var}(\mathrm{cal}) \le$ `var_cap` $\times\, \mathrm{Var}(\mathrm{baseline})$; (iii) include the *identity/baseline* candidate in the stack; (iv) doc note on *global fits and in–sample covariance* for stacking in the tiny reference.
- `results/ablation_sample.jsonl`: tiny, double–blind sample (two lines) illustrating the output schema. *This sample reports IF–only SEs/CIs for brevity; in the paper, all CIs include the OUA addition.*

All other details (schema, TF contract, pseudocode, numerics) appear in the appendices. Public artifacts beyond these examples will be released after review.

F Statement on the Use of Large Language Models (LLMs)

Disclosure. LLMs played a *significant* supporting role in the conception, drafting, and engineering of this work. Consistent with the venue's policy, the authors take full responsibility for all content, including any text or code initially produced with LLM assistance. LLMs are not authors.

Tools. We used general–purpose conversational and code–assistant LLMs ("chat" and IDE–completion interfaces) throughout the project. Interactions were iterative (prompt–response with edits), and no model was granted authorship credit or decision authority.

Roles where LLMs were used.

- **Writing and editing.** Drafting and revising prose (clarity, organization, grammar), converting notes to LaTeX, tightening section transitions, normalizing bibliography fields, and suggesting figure captions.
- **Related work & scoping.** Generating candidate literatures and keywords; proposing nearby citations and contrasts. *All citations included in the paper were manually verified against primary sources before inclusion.*
- **Mathematical development.** Assisting with algebraic manipulations and outline sketches for derivations (e.g., surrogate EIF decomposition, orthogonality checks, majorization facts for isotonic projections, and IF–space stacking conditions). *All proofs and statements in the paper were independently derived and checked by the authors; LLM outputs served only as "hints" or drafts.*
- **Software engineering.** Translating pseudocode to Python scaffolds; proposing unit–test skeletons; producing small utilities (e.g., PAVA helpers, ESS/tail diagnostics, jackknife wrappers). *Every code artifact was reviewed, rewritten as needed, and validated by the authors with tests and end–to–end runs.*
- **Experiment design & analysis.** Suggesting ablation grids, reporting tables, and plotting snippets; stress–testing overlap and calibration–coverage scenarios; drafting README and micro–supplement structure. Final configurations, thresholds, and interpretations were selected by the authors.

Safeguards and verification.

- **Source verification.** No reference was added without manual inspection of the primary paper (title/venue, statement fidelity, and relevance).
- **Math checks.** Key identities (e.g., EIF forms, nuisance–orthogonality, and variance/CI formulas) were re–derived by the authors and cross–checked by simulation or symbolic sanity checks (dimensions, limits, special cases).
- **Code validation.** All LLM–suggested code passed author–written tests (deterministic TF conformance, mean–preserving isotonic calibration, ESS/tail diagnostics, cross–fitting boundaries, and OUA recomputation). Any failing or non–idiomatic snippets were discarded or rewritten.
- **Data handling.** Interactions avoided sharing non–public data or metadata beyond what appears (redacted/anonymized) in the paper or supplement. Double–blindness was preserved in all materials supplied to LLMs.

Limitations of LLM assistance. LLMs can produce incorrect or fabricated statements, code, and citations. Our mitigation was to treat outputs as draft suggestions only, to require independent verification for every technical claim, and to gate inclusion behind tests/derivations and primary–source checks. Any errors that remain are the authors' responsibility.

Authorship and accountability. LLMs are not eligible for authorship. All conceptual contributions, final mathematical results, experimental choices, and interpretations are the authors'. This disclosure exceeds the "significant usage" threshold and is provided to ensure transparency while maintaining the paper's double–blind status.