Large Language Models Enhanced Covariate-adjusted Response-adaptive Randomization Design

Yanping Li
UC Berkeley
yanpingli@berkeley.edu

Xinrui Ruan
UC Berkeley
xinruiruan@berkeley.edu

Jingshen Wang*
UC Berkeley
jingshenwang@berkeley.edu

Waverly Wei*
University of Southern California
waverly@marshall.usc.edu

Abstract

Covariate-adjusted response-adaptive randomization (CARA) designs can improve statistical efficiency and participant welfare in randomized experiments by learning from accrued data and dynamically updating treatment allocation. In practice, two issues limit their impact. Many experiments are sample-size constrained because enrollment is difficult and trials are expensive. At the same time, studies increasingly collect rich pre-treatment information, including unstructured data such as text, images, and clinical notes, yet most CARA designs rely on a small set of structured covariates when computing allocation probabilities. Recent advances in large language models (LLMs) offer a new opportunity to extract useful signals from unstructured inputs and external knowledge. However, few-shot LLM predictions are not ordinary fixed fitted predictions because prompt demonstrations sampled from accrued trial data create data-dependent randomness and can induce correlation across predictions. If used naively to drive allocation updates, LLM-generated signals may weaken CARA's ability to improve power and participant welfare. We propose a CARA design that integrates few-shot LLM-predicted outcomes through resampling-based aggregation, calibration, and effective residual variances to guide both adaptive treatment allocation and downstream treatment effect estimation. Theoretical results and simulation studies show that, when calibrated few-shot LLM auxiliary signals reduce residual variation, coupling an LLM-enhanced estimator with an adaptive allocation rule can improve statistical efficiency and participant welfare.

1 Introduction

Covariate-adjusted response-adaptive randomization (CARA) designs learn from accrued trial data and update treatment assignment probabilities over stages, with the dual aims of improving statistical efficiency and promoting ethical or welfare-oriented goals [Hu and Rosenberger, 2006, Rosenberger et al., 2001, Hu et al., 2015, Rosenberger and Lachin, 2015]. These aims are particularly compelling in modern randomized experiments that are *sample-size constrained* because recruitment is difficult and trials are expensive, yet *data rich at baseline* because pre-treatment records increasingly include structured variables and unstructured inputs such as free-text surveys, clinical notes, and images. In principle, CARA can use accumulating trial information to improve both power and participant welfare. In practice, however, CARA's empirical gains are often under-explored in these regimes [Robertson et al., 2023, Offer-Westort et al., 2021].

*Corresponding authors

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Two practical issues drive this gap. First, when interim sample sizes are small, stagewise estimates are noisy, which can make allocation updates unstable [Hu and Rosenberger, 2006, Zhang et al., 2007, Hu and Hu, 2012]. Second, most CARA procedures rely on a small set of structured covariates when computing allocation probabilities. High-dimensional and unstructured baseline inputs are harder to use during enrollment because prediction models refitted during enrollment can overfit and naive plug-in predictions can inject noise or bias. As a result, outcome-relevant information in unstructured baseline records is often ignored, even when it could improve both design and analysis.

Recent advances in large-scale pretrained language models (LLMs) offer a way to extract auxiliary outcome signals from unstructured baseline inputs in sample-size constrained trials. Motivated by evidence that LLM-based predictions can increase power at the analysis stage [Ruan et al., 2025], a natural next step is to use few-shot LLM signals to guide adaptive allocation. Unlike analysis-stage LLM adjustment for fixed randomized experiments, our setting uses few-shot LLM auxiliary signals inside a CARA design to update future treatment allocation probabilities, requiring predictable randomization and inference under adaptive allocation. Few-shot LLM prediction differs from standard supervised prediction because it is not a fixed fitted function. It conditions on demonstration examples sampled from accrued trial data, which can make predictions data-dependent and correlated across subjects. This stochastic layer is delicate in CARA designs, where allocation probabilities are updated across stages and validity relies on predictable randomization and martingale arguments. LLM-generated predictions may also contain systematic errors or subgroup miscalibration, so naive use can misdirect assignments and weaken efficiency or welfare.

This manuscript studies how to integrate arm-specific few-shot LLM auxiliary signals into CARA while preserving predictable randomization and valid inference. The design treats LLM outputs as auxiliary signals rather than plug-in outcomes. It aggregates prompt-level predictions through a resampling-based procedure inspired by classical U-statistic theory, calibrates the aggregated signal within strata, and uses the resulting effective residual variances to update allocation probabilities. For efficiency maximization, this yields an LLM-enhanced Neyman-type allocation. For welfare improvement, it yields a constrained allocation rule that trades off participant welfare and statistical precision. The final estimator uses design-known propensities and cross-fitting to support inference under predictable randomization [Tsiatis, 2006, Hahn, 1998, Hirano et al., 2003, Ma and Wang, 2020, Bibaut and Kallus, 2025]. Our work is complementary to adaptive Neyman allocation, batch-adaptive experimentation, Neyman-regret, and semiparametric inference for adaptive experiments [Hahn et al., 2011, Dai et al., 2023, Zhao, 2023, Neopane et al., 2025], whereas our focus is on resampled few-shot LLM auxiliary signals that enter both CARA updates and final inference.

Our contributions are summarized as follows:

1. We introduce a CARA design procedure that uses arm-specific few-shot LLM auxiliary predictions to guide adaptive allocation. The procedure aggregates data-dependent and prompt-randomized outputs, calibrates the resulting signals within strata, converts them into effective residual variances, and updates allocation probabilities under efficiency and welfare objectives.

2. We develop an average treatment effect estimator that uses design-known propensities, cross-fitting, and calibrated LLM signals through residual-variance reduction rather than treating LLM outputs as ground-truth counterfactual outcomes. We establish allocation consistency and asymptotic normality by extending CARA martingale arguments to include the prompt-aggregation layer.

3. We characterize multi-layer efficiency gains from calibrated LLM adjustment, LLM-guided allocation, and few-shot aggregation when it improves the auxiliary signal. Theory and simulations show that these gains can improve efficiency and participant welfare when calibrated auxiliary signals reduce residual variation.

2 CARA: Problem Setup

In this section, we develop a CARA framework that uses few-shot LLM auxiliary predictions from the full baseline record to stabilize allocation updates and improve final-stage inference.

2.1 CARA enrollment

The experiment has $S < \infty$ stages. In stage s, n_s individuals are enrolled, with cumulative sample size $N_s = \sum_{r=1}^{s} n_r$, total sample size $N = N_S$, and stage fraction $r_s = n_s/N$, with S and $r_1, \ldots, r_S$ fixed as N grows. Individuals are indexed by enrollment order $i = 1, \ldots, N$, and $S_i = s$ if $N_{s-1} < i \leq N_s$, with $N_0 = 0$.

Following the Neyman-Rubin potential outcomes framework [Neyman, 1923, Rubin, 1974], let $T_i \in \{0, 1\}$ denote treatment assignment and let $Y_i(t)$ denote the potential outcome under arm $t \in \{0, 1\}$. The observed outcome is $Y_i = T_i Y_i(1) + (1 - T_i)Y_i(0)$. The average treatment effect (ATE) is $\tau = \mathbb{E}\{Y(1) - Y(0)\}$, and the stratum-specific effect is $\tau(z) = \mathbb{E}\{Y(1) - Y(0) \mid Z = z\}$. Each subject has baseline record $W_i = (Z_i, X_i, U_i)$, where $Z_i \in \{1, \ldots, J\}$ is a low-dimensional stratification variable, $X_i \in \mathbb{R}^d$ contains structured covariates, and U_i contains high-dimensional or unstructured inputs such as clinical notes, images, or missing-heavy records.

2.2 CARA allocation rules

Let $\mathcal{G}_i$ denote all information generated up to individual i, including baseline records, assignments, and outcomes, whether or not they have been observed. Outcomes may be delayed, so the history $\mathcal{H}_s \subseteq \mathcal{G}_{N_s}$ available at the end of stage s contains only the outcomes observed by then.

A CARA design specifies stratum-specific allocation functions $\boldsymbol{e} = \{e_s(\cdot)\}_{s=1}^{S}$, where $e_s : \{1, \ldots, J\} \to [\delta, 1 - \delta]$ and $\delta \in (0, 1/2)$ enforces positivity. At the beginning of stage s, the design computes $e_s(\cdot)$ from $\mathcal{H}_{s-1}$ and assigns each newly enrolled subject according to

$$\Pr(T_i = 1 \mid \mathcal{G}_{i-1}, Z_i = z) = e_s(z), \quad z \in \{1, \ldots, J\}. \tag{1}$$

2.3 LLM-based counterfactual predictions

To use the full baseline record W_i, we construct prompt-level few-shot predictions $Y_i^{\dagger}(t; \mathcal{S}_{i,t}^{(b)})$, $b = 1, \ldots, B$, where $\mathcal{S}_{i,t}^{(b)}$ is a demonstration set sampled from accrued trial data using only information available before prediction. Each demonstration set consists of m arm-t subjects with observed outcomes, drawn without replacement from the eligible pool, and the target subject never appears in its own demonstration set. We aggregate these predictions as $Y_i^{\dagger}(t) = B^{-1} \sum_{b=1}^{B} Y_i^{\dagger}(t; \mathcal{S}_{i,t}^{(b)})$. The design uses this aggregated signal only after centering and calibration. Algorithm 1 in Appendix A.4 summarizes the full procedure in chronological order, and Appendix A.5 gives the prompt templates.

For each arm t, define $\mu_t^{\dagger}(z) = \mathbb{E}\{Y^{\dagger}(t) \mid Z = z\}$ and $\widetilde{Y}_i^{\dagger}(t) = Y_i^{\dagger}(t) - \mu_t^{\dagger}(Z_i)$. Within stratum z, define the projection slope

$$\omega_t(z) = \frac{\mathbb{C}\mathrm{ov}\{Y(t), \widetilde{Y}^{\dagger}(t) \mid Z = z\}}{\mathbb{V}\mathrm{ar}\{\widetilde{Y}^{\dagger}(t) \mid Z = z\}}, \tag{2}$$

with $\omega_t(z) = 0$ if the denominator is zero. The effective residual variance is

$$\tilde{v}_t(z) = \mathbb{V}\mathrm{ar}\{Y(t) - \omega_t(z)\widetilde{Y}^{\dagger}(t) \mid Z = z\} = v_t(z)\{1 - \rho_t^2(z)\}, \tag{3}$$

where $v_t(z) = \mathbb{V}\mathrm{ar}\{Y(t) \mid Z = z\}$ and $\rho_t(z) = \mathbb{C}\mathrm{orr}\{Y(t), Y^{\dagger}(t) \mid Z = z\}$. Thus, calibrated auxiliary predictions reduce the residual uncertainty driving allocation updates when they are predictive within arm–stratum cells.

2.4 Two design objectives

A CARA policy $\boldsymbol{e}$ shapes treatment assignments across stages and strata. We use two oracle objectives to motivate the proposed updating rules.

Objective I. Efficiency maximization. The first objective chooses allocation probabilities to minimize the large-sample variance of the final ATE estimator. Appendix B shows that this variance decomposes as $\sigma^2(\boldsymbol{e}) = \mathcal{V}(\boldsymbol{e}) + C_{\mathrm{inv}}$, where only $\mathcal{V}(\boldsymbol{e})$ depends on the allocation policy. The oracle design therefore solves

$$\min_{\boldsymbol{e}} \ \mathcal{V}(\boldsymbol{e}) \quad \text{s.t.} \quad e_s(z) \in [\delta, 1 - \delta], \quad s \leq S, \quad z \in \{1, \ldots, J\}. \tag{4}$$

Appendix C shows that the solution decomposes across strata and does not depend on the stage index. For each stratum z, the optimal allocation has the LLM-enhanced Neyman form

$$e_{\mathrm{I}}^{*}(z)=\frac{\sqrt{\tilde{v}_1(z)}}{\sqrt{\tilde{v}_1(z)}+\sqrt{\tilde{v}_0(z)}}. \tag{5}$$

This rule replaces raw within-stratum variances with effective residual variances after calibrated auxiliary adjustment. Here and below, allocation rules are truncated to $[\delta, 1-\delta]$.

Objective II. Welfare improvement under an efficiency constraint. Many trials also prioritize participant welfare, such as reducing the probability of failure outcomes. Let $\mathcal{F}$ denote a failure set and define $\mu_t^{\mathrm{F}}(z)=\Pr\{Y(t)\in\mathcal{F}\mid Z=z\}$. The population expected failure rate under policy $\boldsymbol{e}$ is

$$\mathcal{L}_{\mathrm{F}}(\boldsymbol{e})=\sum_{s=1}^{S} r_s\mathbb{E}\left[e_s(Z)\mu_1^{\mathrm{F}}(Z)+\{1-e_s(Z)\}\mu_0^{\mathrm{F}}(Z)\right]. \tag{6}$$

The welfare-oriented oracle design solves

$$\min_{\boldsymbol{e}}\ \mathcal{L}_{\mathrm{F}}(\boldsymbol{e})\quad \text{s.t.}\quad e_s(z)\in[\delta,1-\delta],\quad s\le S,\quad z\in\{1,\dots,J\},\quad \mathcal{V}(\boldsymbol{e})\le C, \tag{7}$$

where $C>0$ controls the allowed loss in statistical precision. For binary outcomes with $Y=1$ indicating success, one may take $\mathcal{F}=\{0\}$, so that $\mu_t^{\mathrm{F}}(z)=1-\mathbb{E}\{Y(t)\mid Z=z\}$.

The next section gives an implementable stagewise procedure that estimates these primitives from the evolving history, updates allocation probabilities predictably, and supports valid final inference using realized design-known propensities.

3 Our Proposed Method: LLM-enhanced CARA

In this section, we propose an LLM-enhanced CARA design that uses few-shot LLM auxiliary predictions for adaptive allocation and final inference. The design treats LLM outputs as auxiliary signals rather than plug-in outcomes. Because prompt-level predictions can vary with sampled demonstrations, we aggregate over prompt-sampling randomness in a resampling step inspired by U-statistic symmetrization. We then calibrate the aggregated signal against observed outcomes and convert it into effective residual variances for allocation updates.

Stage 1. Enroll n_1 participants. For each enrolled subject i, construct arm-specific aggregated few-shot LLM auxiliary signals $Y_i^{\dagger}(0)$ and $Y_i^{\dagger}(1)$ using only information available before prediction, as described in Section 2.3. Then assign treatment using a prespecified rule, for example $e_1(z)=1/2$ for all strata z.

Stage $s+1$, for $s=1,\dots,S-1$. At the beginning of stage $s+1$, compute the next allocation rule using $\mathcal{H}_s$.

Step 1. Estimation of calibration primitives. Let $\mathcal{O}_s=\{i\le N_s: Y_i \text{ is observed in } \mathcal{H}_s\}$ and $\mathcal{I}_{t,s}(z)=\{i\in\mathcal{O}_s: T_i=t,\ Z_i=z\}$. Let $\widehat{\mu}_{t,s}(z)$ and $\widehat{\mu}^{\dagger}_{t,s}(z)$ estimate $\mu_t(z)=\mathbb{E}\{Y(t)\mid Z=z\}$ and $\mu_t^{\dagger}(z)$, possibly by regression or cross-fitting. Using $\mathcal{I}_{t,s}(z)$, estimate

$$\widehat{\omega}_{t,s}(z)=\frac{\sum_{i\in\mathcal{I}_{t,s}(z)}\{Y_i-\widehat{\mu}_{t,s}(z)\}\{Y_i^{\dagger}(t)-\widehat{\mu}^{\dagger}_{t,s}(z)\}}{\sum_{i\in\mathcal{I}_{t,s}(z)}\{Y_i^{\dagger}(t)-\widehat{\mu}^{\dagger}_{t,s}(z)\}^2},$$

with $\widehat{\omega}_{t,s}(z)=0$ if the denominator is zero. This is the sample version of the projection slope $\omega_t(z)$ in Eq. (2), computed separately in each arm–stratum cell and analyzed in Appendix B, and Appendix A.3 gives the stabilization used in the implementation. When $|\mathcal{I}_{t,s}(z)|>0$, define

$$\widehat{\tilde{v}}_{t,s}(z)=\frac{1}{|\mathcal{I}_{t,s}(z)|}\sum_{i\in\mathcal{I}_{t,s}(z)}\{Y_i-\widehat{\mu}_{t,s}(z)-\widehat{\omega}_{t,s}(z)(Y_i^{\dagger}(t)-\widehat{\mu}^{\dagger}_{t,s}(z))\}^2.$$

If $|\mathcal{I}_{t,s}(z)|=0$, keep the previous-stage value or use a prespecified default.

For final-stage nuisance estimation, use K-fold cross-fitting, with demonstration pools built outside the target fold, as in Algorithm 1.

Step 2. Allocation update. Update $e_{s+1}(\cdot)$ according to the selected objective and truncate to $[\delta, 1-\delta]$.

Objective I. Efficiency maximization. For each stratum z, set

$$\widehat{e}_{s+1,\mathrm{I}}(z) = \frac{\sqrt{\widehat{\widetilde{v}}_{1,s}(z)}}{\sqrt{\widehat{\widetilde{v}}_{1,s}(z)} + \sqrt{\widehat{\widetilde{v}}_{0,s}(z)}}. \tag{8}$$

Objective II. Welfare improvement under an efficiency constraint. Let $\widehat{\mu}^{\mathrm{F}}_{t,s}(z)$ estimate $\mu^{\mathrm{F}}_t(z)$, and let $\widehat{p}_s(z)$ be the empirical stratum proportion up to stage s. Given a precision bound C_s, choose $\widehat{e}_{s+1,\mathrm{II}}(\cdot)$ by solving

$$\begin{aligned} \min_{e(\cdot)} \quad & \sum_{z=1}^{J} \widehat{p}_s(z)\{e(z)\widehat{\mu}^{\mathrm{F}}_{1,s}(z) + (1-e(z))\widehat{\mu}^{\mathrm{F}}_{0,s}(z)\} \\ \text{s.t.} \quad & e(z) \in [\delta, 1-\delta],\ \forall z, \\ & \sum_{z=1}^{J} \widehat{p}_s(z) \left\{ \frac{\widehat{\widetilde{v}}_{1,s}(z)}{e(z)} + \frac{\widehat{\widetilde{v}}_{0,s}(z)}{1-e(z)} \right\} \le C_s. \end{aligned} \tag{9}$$

By Appendix D, if Eq. (9) is strictly feasible, there is $\lambda^*_s \ge 0$ such that, for each z with $\widehat{p}_s(z) > 0$,

$$\widehat{e}_{s+1,\mathrm{II}}(z) \in \operatorname*{argmin}_{e \in [\delta, 1-\delta]} \left[e\{\widehat{\mu}^{\mathrm{F}}_{1,s}(z) - \widehat{\mu}^{\mathrm{F}}_{0,s}(z)\} + \lambda^*_s \left\{ \frac{\widehat{\widetilde{v}}_{1,s}(z)}{e} + \frac{\widehat{\widetilde{v}}_{0,s}(z)}{1-e} \right\} \right], \tag{10}$$

with λ^*_s chosen so that Eq. (9) is tight when $\lambda^*_s > 0$.

Step 3. Few-shot LLM auxiliary signals and treatment assignment for newly enrolled subjects. Set $\widehat{e}_{s+1}(\cdot) = \widehat{e}_{s+1,\mathrm{I}}(\cdot)$ under Objective I and $\widehat{e}_{s+1}(\cdot) = \widehat{e}_{s+1,\mathrm{II}}(\cdot)$ under Objective II. For each new subject i with $S_i = s+1$, compute $Y_i^\dagger(0)$ and $Y_i^\dagger(1)$ from W_i at enrollment, and then assign treatment by $\Pr(T_i = 1 \mid \mathcal{G}_{i-1}, Z_i = z) = \widehat{e}_{s+1}(z)$.

Statistical inference after stage S. Define $e_i = e_1(Z_i)\mathbb{1}\{S_i = 1\} + \widehat{e}_{S_i}(Z_i)\mathbb{1}\{S_i \ge 2\}$, $e_{i,1} = e_i$, $e_{i,0} = 1 - e_i$, and $O_i = (W_i, T_i, Y_i)$. For each arm t, form

$$\widehat{\varphi}_t(O_i, Y_i^\dagger(t)) = \frac{\mathbb{1}\{T_i = t\}Y_i}{e_{i,t}} + \left(1 - \frac{\mathbb{1}\{T_i = t\}}{e_{i,t}}\right)\left[\widehat{\mu}_t(Z_i) + \widehat{\omega}_t(Z_i)\{Y_i^\dagger(t) - \widehat{\mu}_t^\dagger(Z_i)\}\right].$$

The ATE estimator is

$$\widehat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\{\widehat{\varphi}_1(O_i, Y_i^\dagger(1)) - \widehat{\varphi}_0(O_i, Y_i^\dagger(0))\},$$

with variance estimate

$$\widehat{\mathbb{V}} = \frac{1}{N}\sum_{i=1}^{N}\{\widehat{\varphi}_1(O_i, Y_i^\dagger(1)) - \widehat{\varphi}_0(O_i, Y_i^\dagger(0)) - \widehat{\tau}\}^2.$$

This is an augmented inverse-probability-weighted (AIPW) estimator with the LLM-calibrated working regression $m_t(z, y^\dagger) = \mu_t(z) + \omega_t(z)\{y^\dagger - \mu_t^\dagger(z)\}$ in Eq. (14), where $\mu_t^\dagger(z)$ centers the raw signal and $\omega_t(z)$ sets its direction and scale. It reduces to standard AIPW when $\omega_t \equiv 0$. The same calibration also defines the effective residual variances used for allocation. A plug-in estimate of the population failure rate induced by the realized policy is

$$\widehat{\mathcal{L}}_{\mathrm{F}} = \sum_{s=1}^{S} r_s \sum_{z=1}^{J} \widehat{p}_s(z)\{e_s(z)\widehat{\mu}^{\mathrm{F}}_{1,s}(z) + (1-e_s(z))\widehat{\mu}^{\mathrm{F}}_{0,s}(z)\}.$$

Remark 3.1 (Optional validation-gated safeguard)**.** *In small samples or under possible subgroup miscalibration, the auxiliary signal may spuriously appear predictive within an arm–stratum cell. As an optional finite-sample safeguard, one may replace $\widehat{\omega}_{t,s}(z)$ and $\widehat{\widetilde{v}}_{t,s}(z)$ by validation-gated versions $\widehat{\omega}^{\mathrm{scr}}_{t,s}(z)$ and $\widehat{\widetilde{v}}^{\mathrm{scr}}_{t,s}(z)$ that screen out cells without statistically positive out-of-fold residual-variance reduction. The screened versions then enter both the allocation update in Step 2 and the estimator $\widehat{\tau}$. Appendix A.1 gives the construction. We recommend the safeguard when auxiliary-signal reliability is uncertain. The main procedure and the theory below use the unscreened calibration to isolate the target mechanism, and the screened version is used only in robustness analyses.*

4 Efficiency gains and welfare improvements

Our framework couples a calibrated LLM-enhanced estimator with a predictable CARA allocation policy. The estimator uses few-shot LLM auxiliary signals, and the allocation rule uses estimated effective residual variances. This coupling can yield layered efficiency gains under Objective I and welfare benefits under Objective II when the calibrated signal reduces residual variation.

4.1 Three layers of efficiency gains under Objective I

We formalize efficiency comparisons under Objective I using the allocation-relevant oracle criterion $\mathcal{V}(\boldsymbol{e})$. Appendix B shows that

$$\mathcal{V}(\boldsymbol{e}) = \sum_{s=1}^{S} r_s \mathbb{E}\left[\frac{\tilde{v}_1(Z)}{e_s(Z)} + \frac{\tilde{v}_0(Z)}{1-e_s(Z)}\right], \tag{11}$$

where $\tilde{v}_t(z) = v_t(z)\{1-\rho_t^2(z)\}$ is the effective residual variance in Eq. (3).

Let $\mathcal{V}_{\mathrm{AIPW}}(\boldsymbol{e})$ denote the corresponding allocation-relevant oracle criterion for standard AIPW without auxiliary adjustment. Let $\mathcal{V}^{\mathrm{zs}}$ and $\mathcal{V}^{\mathrm{fs}}$ denote the criteria induced by zero-shot and few-shot auxiliary signals, and let $\boldsymbol{e}_{\mathrm{I}}^{*,\mathrm{zs}}$ and $\boldsymbol{e}_{\mathrm{I}}^{*,\mathrm{fs}}$ denote the corresponding LLM-enhanced Neyman oracle policies. Without the auxiliary signal, the classical Neyman allocation is

$$e_{\mathrm{Ney}}^{*}(z) = \frac{\sqrt{v_1(z)}}{\sqrt{v_1(z)}+\sqrt{v_0(z)}}. \tag{12}$$

The total Objective I gain admits the telescoping identity

$$\begin{aligned}\mathcal{V}_{\mathrm{AIPW}}(\boldsymbol{e}_{\mathrm{Ney}}^{*}) - \mathcal{V}^{\mathrm{fs}}(\boldsymbol{e}_{\mathrm{I}}^{*,\mathrm{fs}}) = &\underbrace{\left\{\mathcal{V}_{\mathrm{AIPW}}(\boldsymbol{e}_{\mathrm{Ney}}^{*}) - \mathcal{V}^{\mathrm{zs}}(\boldsymbol{e}_{\mathrm{Ney}}^{*})\right\}}_{\text{Layer 1: LLM-enhanced efficiency gain}} + \underbrace{\left\{\mathcal{V}^{\mathrm{zs}}(\boldsymbol{e}_{\mathrm{Ney}}^{*}) - \mathcal{V}^{\mathrm{zs}}(\boldsymbol{e}_{\mathrm{I}}^{*,\mathrm{zs}})\right\}}_{\text{Layer 2: design-enhanced efficiency gain}} \\ &+ \underbrace{\left\{\mathcal{V}^{\mathrm{zs}}(\boldsymbol{e}_{\mathrm{I}}^{*,\mathrm{zs}}) - \mathcal{V}^{\mathrm{fs}}(\boldsymbol{e}_{\mathrm{I}}^{*,\mathrm{fs}})\right\}}_{\text{Layer 3: few-shot-enhanced efficiency gain}}.\end{aligned} \tag{13}$$

Layer 1 fixes the allocation at $\boldsymbol{e}_{\mathrm{Ney}}^{*}$ and changes only the estimator, from standard AIPW to the zero-shot calibrated estimator. Layer 2 fixes the zero-shot criterion and changes only the policy, from $\boldsymbol{e}_{\mathrm{Ney}}^{*}$ to $\boldsymbol{e}_{\mathrm{I}}^{*,\mathrm{zs}}$. Layer 3 evaluates each auxiliary under its own optimized policy and captures the additional gain from the few-shot construction. The first two are nonnegative oracle comparisons, while Layer 3 depends on prompt quality and need not be positive.

4.2 Welfare-enhanced benefits under Objective II

Objective II uses $\mathcal{L}_{\mathrm{F}}(\boldsymbol{e})$ as the objective and treats $\mathcal{V}(\boldsymbol{e}) \le C$ as a precision constraint. Calibrated LLM adjustment replaces $v_t(z)$ by $\tilde{v}_t(z) \le v_t(z)$ in the allocation-relevant criterion, so it can reduce $\mathcal{V}(\boldsymbol{e})$ for a fixed allocation and relax the feasible set $\{\boldsymbol{e} : \mathcal{V}(\boldsymbol{e}) \le C\}$. The welfare optimizer can then select more welfare-oriented allocations under the same bound on the allocation-relevant criterion.

This gives a three-layer Objective-II mechanism using the same layer names as in Objective I, but with a different interpretation. Layer 1, the LLM-enhanced efficiency gain, is a constraint-side effect. Calibrated LLM adjustment reduces the allocation-relevant variance criterion and creates precision slack. Layer 2, the design-enhanced welfare gain, is the welfare-side benefit from re-optimizing $\mathcal{L}_{\mathrm{F}}(\boldsymbol{e})$ using this slack. Layer 3, the few-shot-enhanced welfare gain, is the additional welfare-side benefit when the few-shot construction improves or stabilizes the auxiliary signal used in the constraint. Thus, unlike the Objective I decomposition, the Objective-II layers are mechanism components rather than an additive variance decomposition.

5 Theoretical investigation

This section gives large-sample guarantees for the proposed LLM-enhanced CARA procedure under predictable randomization. The proof has two components. The first uses martingale arguments to

handle allocation probabilities that depend on data from earlier stages. The second shows that, under resampling stability, the randomness from sampled few-shot demonstrations does not affect first-order inference. The assumptions below separate the main sources of randomness in the proposed design.

Assumption 5.1 (Regularity conditions)**.** *The baseline records and potential outcomes* $\{(W_i, Y_i(0), Y_i(1))\}_{i=1}^N$ *are independently and identically distributed, and* $(W_i, Y_i(0), Y_i(1))$ *is independent of* $\mathcal{G}_{i-1}$*. For each* $t \in \{0,1\}$*,* $\mathbb{E}\{|Y_i(t)|^4\} < \infty$*, and there exists* $\varepsilon > 0$ *such that* $\tilde{v}_t(z) \geq \varepsilon$ *for all* z *with* $p(z) = \Pr(Z_i = z) > 0$*. Under Objective II,* $C_s \xrightarrow{\mathrm{p}} C > \mathcal{V}(\boldsymbol{e}_{\mathrm{I}}^*)$ *for each* s*, and* $\mu_1^{\mathrm{F}}(z) \neq \mu_0^{\mathrm{F}}(z)$ *for all* z *with* $p(z) > 0$*.*

Assumption 5.1 gives the moment and non-degeneracy conditions needed for martingale central limit arguments and, under Objective II, a unique oracle rule.

Assumption 5.2 (Sequential randomization)**.** *For each stage* s *and stratum* z*,* $e_s(z)$ *is* $\mathcal{H}_{s-1}$*-measurable and satisfies* $e_s(z) \in [\delta, 1-\delta]$*. Conditional on* $(\mathcal{G}_{i-1}, Z_i)$*,* T_i *is generated according to Eq.* (1)*. In particular,* $T_i \perp\!\!\!\perp \{Y_i(0), Y_i(1)\} \mid \mathcal{G}_{i-1}, Z_i$*.*

Assumption 5.2 ensures predictable treatment assignment despite adaptive updates.

Assumption 5.3 (Nuisance estimation)**.** *For each stage* s *and arm* t*, the estimators* $\widehat{\mu}_{t,s}$*,* $\widehat{\mu}_{t,s}^{\mathrm{F}}$*,* $\widehat{\mu}_{t,s}^{\dagger}$*,* $\widehat{\omega}_{t,s}$*, and* $\widehat{\tilde{v}}_{t,s}$ *are consistent for* μ_t*,* μ_t^{F}*,* $\mu_t^{\dagger}$*,* ω_t*, and* $\tilde{v}_t$ *in* L^2*, and so are the cross-fitted* $\widehat{\mu}_t$*,* $\widehat{\mu}_t^{\dagger}$*, and* $\widehat{\omega}_t$ *used in* $\widehat{\tau}$*.*

Assumption 5.3 requires only consistency, because the design-known propensities make the nuisance remainder a finite sum of martingale averages with vanishing coefficients.

Assumption 5.4 (Few-shot LLM resampling stability)**.** *(i) For each subject* i*, arm* t*, and replicate* $b = 1, \ldots, B$*, the prompt-level output* $Y_i^{\dagger}(t; \mathcal{S}_{i,t}^{(b)})$ *is generated before* T_i *is assigned, using only information available at that time, and* T_i *is conditionally independent of* $\{W_i, Y_i(0), Y_i(1), Y_i^{\dagger}(0), Y_i^{\dagger}(1)\}$ *given* $(\mathcal{G}_{i-1}, Z_i)$*. The prompt-level outputs take values in* $[0, M_{\dagger}]$ *for some* $M_{\dagger} < \infty$*. (ii) Conditional on* W_i *and the demonstration pool, the* B *prompt-level outputs are independent with common mean* $\bar{f}_{i,t}(W_i)$*, and there is a stable limiting score* $f_{\mathrm{FS},t}$ *such that* $N^{-1}\sum_{i=1}^N \{\bar{f}_{i,t}(W_i) - f_{\mathrm{FS},t}(W_i)\}^2 = o_p(1)$*. (iii)* $B = B_N \to \infty$ *as* $N \to \infty$*.*

Assumption 5.4 allows the aggregated signal $Y_i^{\dagger}(t)$ to depend on the trial history through the demonstrations, and part (i) requires only that it be predictable, that is, fixed before T_i is assigned. In part (ii), $\bar{f}_{i,t}(W_i)$ is the mean prompt-level output over demonstration draws and decoding randomness, given the pool available to subject i. Parts (ii)–(iii), pool stability and growing B, impose no rate relative to N, and the population quantities in this section are defined with $f_{\mathrm{FS},t}$ as the auxiliary signal.

Lemma 5.5 (Stability of few-shot aggregated signal)**.** *Under Assumption 5.4, the aggregated few-shot signal satisfies* $Y_i^{\dagger}(t) = f_{\mathrm{FS},t}(W_i) + R_{i,t,B}$ *with* $N^{-1}\sum_{i=1}^N R_{i,t,B}^2 = O_p(B^{-1}) + o_p(1) = o_p(1)$ *for each arm* t*.*

Lemma 5.5 bounds the remainder only in mean square. This suffices for Theorem 5.9 because the remainder is fixed before assignment, so its contribution to the estimator is a martingale sum.

Theorem 5.6 (Design strategy consistency)**.** *Under Assumptions 5.1–5.4, for each fixed stage* $s = 2, \ldots, S$ *and each stratum* z *with* $p(z) > 0$*,*

$$\widehat{e}_{s,\mathrm{I}}(z) \xrightarrow{\mathrm{p}} e_{\mathrm{I}}^*(z), \quad \widehat{e}_{s,\mathrm{II}}(z) \xrightarrow{\mathrm{p}} e_{\mathrm{II}}^*(z),$$

where $e_{\mathrm{I}}^*(z)$ *and* $e_{\mathrm{II}}^*(z)$ *are the oracle allocation rules under Objectives I and II.*

Theorem 5.6 justifies the stagewise plug-in updates by showing convergence to the oracle rules.

Theorem 5.7 (Asymptotic normality)**.** *Under Assumptions 5.1–5.4, as* $N \to \infty$*,*

$$\sqrt{N}(\widehat{\tau} - \tau) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_*^2), \quad \widehat{\mathbb{V}} \xrightarrow{\mathrm{p}} \sigma_*^2,$$

where $\sigma_*^2 = \sigma^2(\boldsymbol{e}^*)$ *is the full influence-function variance of Appendix B at the limiting design* $\boldsymbol{e}^*$*.*

Theorem 5.7 justifies inference under the adaptive design by using the full variance σ_*^2 at the limiting design $\boldsymbol{e}^*$, which uses e_1 in stage 1 and the oracle rules of Theorem 5.6 afterwards. Appendix G.1 gives conditions under which the theorem also holds for a fixed B.

Theorem 5.8 (Efficiency gain)**.** *Under Assumptions 5.1–5.4, the allocation-relevant criterion satisfies two oracle comparisons. For any feasible CARA policy* $\boldsymbol{e}$*,*

$$\mathcal{V}(\boldsymbol{e}) \leq \mathcal{V}_{\mathrm{AIPW}}(\boldsymbol{e}).$$

Moreover, if $\boldsymbol{e}_{\mathrm{I}}^*$ *is the Objective I oracle policy and* $\boldsymbol{e}_{\mathrm{Ney}}^*$ *is the classical Neyman policy in Eq.* (12)*, then*

$$\mathcal{V}(\boldsymbol{e}_{\mathrm{I}}^*) \leq \mathcal{V}(\boldsymbol{e}_{\mathrm{Ney}}^*).$$

Theorem 5.8 formalizes the two nonnegative oracle layers of the Objective-I decomposition, namely calibrated residual-variance reduction and allocation using effective residual variances. Section 6 compares signals using the full variance $\sigma^2(\boldsymbol{e})$, because C_{inv} does not depend on $\boldsymbol{e}$ but does depend on the auxiliary signal.

Theorem 5.9 (Asymptotic equivalence under few-shot aggregation)**.** *Suppose Assumptions 5.1–5.4 hold. Let* $\widehat{\tau}$ *be the proposed estimator using the aggregated few-shot signal* $Y_i^\dagger(t)$*, and let* $\widehat{\tau}^*$ *be the same estimator with* $Y_i^\dagger(t)$ *replaced by* $f_{\mathrm{FS},t}(W_i)$ *in the calibration and estimation steps, holding the realized assignments, propensities, and cross-fitting folds fixed. Then*

$$\widehat{\tau} - \widehat{\tau}^* = o_p(N^{-1/2}).$$

Thus, $\widehat{\tau}$ *has the same first-order influence function and asymptotic distribution as* $\widehat{\tau}^*$*.*

Theorem 5.9 formalizes the few-shot layer. Under resampling stability, prompt randomness does not affect first-order inference, although finite-sample gains still depend on prompt quality.

6 Simulation studies

We evaluate the proposed LLM-enhanced CARA procedure on synthetic data calibrated to the BRIGHTEN study [Pratap et al., 2022], a fully remote smartphone-based randomized trial for depression management with structured baseline covariates and free-text surveys.

6.1 Simulation setup

BRIGHTEN synthetic data generation. BRIGHTEN enrolled adults with clinically significant depressive symptoms and randomized participants to app-delivered interventions, with follow-up over 12 weeks. The treatment arm, $T = 1$, is the iPST app, which delivers internet-based problem-solving therapy, and the control arm, $T = 0$, consists of the alternative apps. We use a BRIGHTEN-calibrated synthetic superpopulation of $20{,}000$ subjects with structured baseline covariates X_i, two recovered baseline free-text fields U_i, and arm-specific potential outcomes for the PHQ-9 depression severity score. The data-generating process (DGP) preserves the original BRIGHTEN-calibrated conditional outcome structure and adds an orthogonal text-augmentation component derived from LLM auxiliary predictions. The arm-specific outcomes are recentered to preserve the original BRIGHTEN marginal means and clipped to the valid PHQ-9 range $[0, 27]$. The conditional mean models therefore combine structured covariates with this LLM-derived component, so that the LLM auxiliary signal has nontrivial but imperfect predictive content for the generated outcomes.

For LLM prompting on synthetic subjects, we recover each subject's pre-treatment text by matching the synthetic text representation to the closest entry of a phrase bank built from observed BRIGHTEN free-text responses and their paraphrases. We pair the recovered snippets with the synthetic structured covariates to form a complete baseline record. Qwen3-8B receives a fixed prompt that serializes age, gender, education, working status, marital status, race/ethnicity, and the two free-text responses, and requests integer PHQ-9 predictions. The zero-shot prompt contains only the target subject's record and returns predictions for both arms. The few-shot prompt is arm-specific and adds $m = 5$ demonstrations with the same fields and their arm-t outcomes, drawn without replacement from the other subjects of the superpopulation, one from each of five education categories. Predictions are averaged over $B = 4$ demonstration sets per subject-arm pair. We view this value as $B_N = \max\{4, \lceil \log \log N \rceil\}$ at $N = 2{,}000$, a sequence that diverges as required by Assumption 5.4(iii). Appendix A.5 gives the prompt templates and decoding settings.

Design setup. We run $R = 300$ Monte Carlo replicates. In each replicate, we sample $N = n_1 + n_2 = 2{,}000$ units without replacement from the superpopulation, set $n_1 = n_2$, and randomly permute

Table 1: Population-oracle Objective-I gains, in percent, on the full influence-function variance $\sigma^2(\boldsymbol{e})$, relative to standard AIPW under classical Neyman allocation. $\bar{\rho}_t^2$ is the population-weighted $\rho_t^2(z)$.

Auxiliary signal	$\bar{\rho}_0^2$	$\bar{\rho}_1^2$	Est.	Alloc.	Total
Demo-only	0.014	0.001	0.60	0.00	0.60
TF-IDF+RF	0.086	0.058	6.57	0.01	6.58
TF-IDF/SVD+RF	0.077	0.053	6.22	0.01	6.23
TF-IDF/SVD+Ridge	0.072	0.066	6.46	0.00	6.47
LLM-ZS	0.140	0.018	6.62	0.11	6.73
LLM-FS	0.274	0.001	6.72	0.54	**7.26**
Oracle signal	0.493	0.283	22.32	0.46	22.78

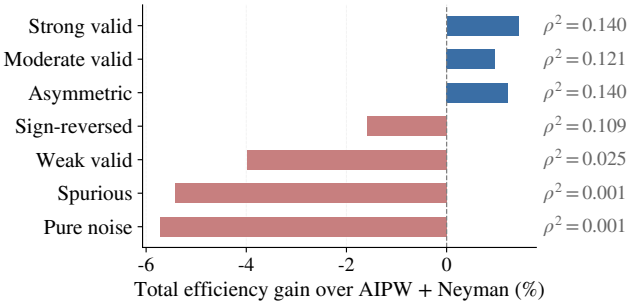

Figure 1: Realized gains under signal scenarios that vary auxiliary-signal strength, direction, and arm-specific reliability. Appendix J gives the construction details.

enrollment order. The strata Z are the 18 education $\times$ working-status cells. Stage 1 uses equal randomization, $e_1(z) \equiv 1/2$. Under Objective I, stage 2 uses stratum-specific allocation probabilities computed from stage-1 estimates and truncated to $[\delta, 1-\delta]$ with $\delta = 0.05$. The classical Neyman, Ethical, and Objective-II welfare allocations are computed from the superpopulation.

We compare zero-shot and few-shot Qwen3-8B prompting with non-LLM text baselines that combine the structured covariates with TF-IDF features, with or without truncated SVD, and fit random forests or ridge regression. Demo-only, a baseline that uses only the structured covariates, and an oracle signal equal to the true conditional mean plus Gaussian noise serve as reference points. The trained non-LLM baselines use 2-fold cross-fitting, so each target prediction comes from a model trained outside the target fold.

Evaluation metrics. For Objective I, Table 1 reports population-oracle gains as percent reductions in the full influence-function variance $\sigma^2(\boldsymbol{e})$ relative to standard AIPW under the classical Neyman allocation, split into estimation and allocation components. All auxiliary signals are evaluated on the same subjects, outcomes, strata, and calibration procedure. Appendix J.1 reports the three-layer oracle decomposition in Eq. (13) in four strata. For Objective II, we report the reduction in the adverse PHQ-9 event rate, where the event is defined as PHQ-9 ≥ 15, corresponding to the two most severe PHQ-9 categories, relative to the Ethical design. Real-data-calibrated synthetic experiments are common in CARA methodology because counterfactual outcomes and repeated adaptive trials are not observed in a single study [Hu et al., 2015, Villar and Rosenberger, 2018, Zhu and Zhu, 2023].

6.2 Simulation results

Table 1 compares auxiliary prediction methods on a common basis. Few-shot LLM prompting gives the largest population-oracle gain among the learned auxiliaries, 7.26%, compared with 6.73% for zero-shot prompting and 6.23%–6.58% for the non-LLM text predictors. Figure 1 shows that realized gains depend on signal reliability. Strong, moderate, and asymmetric valid signals produce positive gains, whereas sign-reversed, weak, spurious, and pure-noise signals give negative realized gains. Because Lemma B.1 shows that the oracle gain depends on $\rho_t^2(z)$ rather than the sign of $\rho_t(z)$, and calibration reverses a sign-reversed signal through $\omega_t(z) < 0$, these negative values reflect finite-sample estimation of the calibration and nuisance quantities.

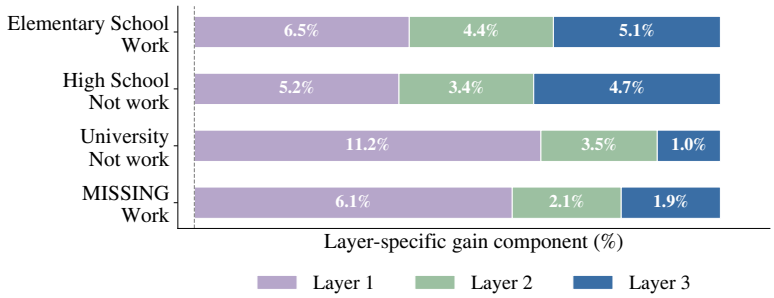

Figure 2: Objective-II mechanism components in four strata. Layer 1 is the LLM-enhanced efficiency gain that relaxes the precision constraint, and Layers 2 and 3 are the design-enhanced and few-shot-enhanced welfare gains in failure-rate reduction. The bars are not an additive decomposition.

Figure 2 shows how Objective II converts precision slack into lower adverse outcome rates. In the four displayed strata, zero-shot calibration lowers the allocation-relevant criterion at the Ethical design by 5.2%–11.2%, which is Layer 1. Under the common precision bound of Appendix J, the welfare optimizer uses this slack to move the treatment probability from the nearly balanced Ethical allocation toward the arm with the lower failure rate, by 0.08–0.15 with zero-shot and 0.18–0.20 with few-shot signals. The failure rate then falls by 2.1%–4.4% of the Ethical rate in Layer 2 and by a further 1.0%–5.1% in Layer 3. Layers 2 and 3 need not track Layer 1 because the welfare gain also scales with the arm difference in failure rates. Over all strata, Layers 2 and 3 equal 1.82% and 0.94%.

Under Objective I, the LLM-enhanced Neyman rule responds to arm asymmetry in effective residual uncertainty, not to overall signal strength. In the predictor-independent experiment of Appendix J.8, $(\rho_0^2, \rho_1^2) = (0.8, 0.1)$ moves the oracle allocation to 0.680 and gives a 6.29% allocation gain, whereas $(0.6, 0.6)$ gives a 30.00% estimation gain and no allocation gain. In the BRIGHTEN-calibrated population, the population-weighted $\sqrt{\tilde{v}_1(Z)/\tilde{v}_0(Z)}$ is 0.947 for zero-shot and 1.039 for few-shot signals, with mean absolute shifts from $e^*_{\mathrm{Ney}}(Z)$ of 0.017 and 0.040, so the allocation gain is small.

In the independently generated structured clinical superpopulation of Appendix J.7, whose outcomes contain none of the evaluated predictions, absolute bias is at most 0.065 and coverage is between 0.94 and 0.96. In Appendix J.9, on the same superpopulation, changing only the auxiliary model from Qwen3-8B to GPT-5-mini gives population-oracle gains of 14.80% and 14.47% and coverage of 0.960 and 0.963. Appendix J also reports arm-by-stratum calibration, safeguard, screening, and finite-B diagnostics.

7 Discussion

We developed an LLM-enhanced CARA framework that treats calibrated arm-specific LLM predictions as auxiliary signals rather than plug-in counterfactual outcomes and links them to allocation and estimation through effective residual variances, while preserving design-known randomization and cross-fitted inference. The method may improve statistical efficiency and participant welfare when auxiliary signals are reliable. In our simulation, few-shot prompting gives the largest population-oracle efficiency gain among the learned auxiliaries and lowers the adverse-event rate under Objective II. Its main limitations concern finite-sample stability and auxiliary-signal reliability. Sparse strata, noisy outcomes, prompt variation, population shift, and query cost can reduce the value of LLM signals, and unreliable or miscalibrated signals can steer allocation away from its intended target.

These limitations call for statistical safeguards, namely validation diagnostics, truncation, optional screening, and finite-B stability checks, as well as operational ones. Because the propensities are design-known, inference remains asymptotically valid with a poor auxiliary signal, so these safeguards mainly protect efficiency. A deployed trial should also prespecify the model version, prompt template, decoding settings, demonstration-selection rule, parsing rule, and update schedule, archive every prompt, output, and allocation probability for reproducibility, and process privacy-sensitive records through an approved local or private deployment. Future work will sharpen safeguards against miscalibrated auxiliary predictions and evaluate the design in prospective trials with delayed outcomes.

Acknowledgments and Disclosure of Funding

This work was supported by NIH grants R01AG096488 and R01AG089512 and NSF grant DMS-2239047. The authors declare no competing interests.

References

Aurélien Bibaut and Nathan Kallus. Demystifying inference after adaptive experiments. *Annual Review of Statistics and Its Application*, 12(1):407–423, 2025.

Jessica Dai, Paula Gradu, and Christopher Harshaw. CLIP-OGD: An experimental design for adaptive Neyman allocation in sequential experiments. *Advances in Neural Information Processing Systems*, 36:32235–32269, 2023.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331, 1998.

Jinyong Hahn, Keisuke Hirano, and Dean Karlan. Adaptive experimental design using the propensity score. *Journal of Business & Economic Statistics*, 29(1):96–108, 2011.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Feifang Hu and William F Rosenberger. *The Theory of Response-adaptive Randomization in Clinical Trials*. John Wiley & Sons, Hoboken, NJ, 2006.

Jianhua Hu, Hongjian Zhu, and Feifang Hu. A unified family of covariate-adjusted response-adaptive designs based on efficiency and ethics. *Journal of the American Statistical Association*, 110(509): 357–367, 2015.

Yanqing Hu and Feifang Hu. Asymptotic properties of covariate-adaptive randomization. *Annals of Statistics*, 40(3):1794–1815, 2012.

Xinwei Ma and Jingshen Wang. Robust inference using inverse probability weighting. *Journal of the American Statistical Association*, 115(532):1851–1860, 2020.

Ojash Neopane, Aaditya Ramdas, and Aarti Singh. Logarithmic Neyman regret for adaptive estimation of the average treatment effect. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 4303–4311. PMLR, 03–05 May 2025.

Jerzy Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Annals of Agricultural Sciences*, 10:1–51, 1923. Translated and edited by D. M. Dabrowska and T. P. Speed, *Statistical Science*, 5(4):465–472, 1990.

Molly Offer-Westort, Alexander Coppock, and Donald P Green. Adaptive experimental design: Prospects and applications in political science. *American Journal of Political Science*, 65(4): 826–844, 2021.

Abhishek Pratap, Ava Homiar, Luke Waninger, Calvin Herd, Christine Suver, Joshua Volponi, Joaquin A Anguera, and Pat Areán. Real-world behavioral dataset from two fully remote smartphone-based randomized clinical trials for depression. *Scientific Data*, 9(1):522, 2022.

David S Robertson, Kim May Lee, Boryana C López-Kolkovska, and Sofía S Villar. Response-adaptive randomization in clinical trials: From myths to practical considerations. *Statistical Science*, 38(2):185–208, 2023.

William F Rosenberger and John M Lachin. *Randomization in Clinical Trials: Theory and Practice*. John Wiley & Sons, Hoboken, NJ, second edition, 2015.

William F Rosenberger, Nigel Stallard, Anastasia Ivanova, Cherice N Harper, and Michelle L Ricks. Optimal adaptive designs for binary response trials. *Biometrics*, 57(3):909–913, 2001.

Xinrui Ruan, Xinwei Ma, Yingfei Wang, Waverly Wei, and Jingshen Wang. Can language models boost the power of randomized experiments without statistical bias? *arXiv preprint arXiv:2510.05545*, 2025.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

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

Sofía S Villar and William F Rosenberger. Covariate-adjusted response-adaptive randomization for multi-arm clinical trials using a modified forward looking Gittins index rule. *Biometrics*, 74(1): 49–57, 2018.

Li-Xin Zhang, Feifang Hu, Siu Hung Cheung, and Wai Sum Chan. Asymptotic properties of covariate-adjusted response-adaptive designs. *Annals of Statistics*, 35(3):1166–1182, 2007.

Jinglong Zhao. Adaptive Neyman allocation. arXiv preprint arXiv:2309.08808, 2023.

Hai Zhu and Hongjian Zhu. Covariate-adjusted response-adaptive designs based on semiparametric approaches. *Biometrics*, 79(4):2895–2906, 2023.

Technical appendices

A Additional methodological details

A.1 Optional validation-gated safeguard

In cell (t, z) at stage s, let $\widehat{\omega}^{\mathrm{cf}}_{t,s}(z)$, $\widehat{\tilde{v}}^{\mathrm{cf}}_{t,s}(z)$, and $\widehat{v}^{\mathrm{cf}}_{t,s}(z)$ be cross-fitted estimates of $\omega_t(z)$, $\tilde{v}_t(z)$, and $v_t(z)$, and let $\widehat{G}^{\mathrm{cf}}_{t,s}(z)$ be the cross-fitted reduction in residual variance from using the auxiliary signal, with standard error $\widehat{\mathrm{se}}\{\widehat{G}^{\mathrm{cf}}_{t,s}(z)\}$. For a prespecified critical value c_α and truncation level $M > 0$, define

$$\widehat{\omega}^{\mathrm{scr}}_{t,s}(z) = \Pi_{[-M,M]}\{\widehat{\omega}^{\mathrm{cf}}_{t,s}(z)\}\mathbb{1}\left[\widehat{G}^{\mathrm{cf}}_{t,s}(z) > c_\alpha \widehat{\mathrm{se}}\{\widehat{G}^{\mathrm{cf}}_{t,s}(z)\}\right].$$

The screened variance input is

$$\widehat{\tilde{v}}^{\mathrm{scr}}_{t,s}(z) = \begin{cases} \widehat{\tilde{v}}^{\mathrm{cf}}_{t,s}(z), & \widehat{G}^{\mathrm{cf}}_{t,s}(z) > c_\alpha \widehat{\mathrm{se}}\{\widehat{G}^{\mathrm{cf}}_{t,s}(z)\}, \\ \widehat{v}^{\mathrm{cf}}_{t,s}(z), & \text{otherwise.} \end{cases}$$

When the gate fails, the screened coefficient is zero and the raw cell variance is used, so the update reverts to the raw variance input of the classical Neyman allocation in that cell.

A.2 Interpretation of the three-layer gain decomposition

Eq. (13) gives the Objective I three-layer accounting identity. This decomposition compares four oracle quantities, namely standard AIPW under the classical Neyman allocation, the zero-shot LLM-enhanced criterion under the same classical Neyman allocation, the zero-shot LLM-enhanced criterion under the zero-shot LLM-enhanced Neyman allocation, and the few-shot LLM-enhanced criterion under the few-shot LLM-enhanced Neyman allocation. The three differences isolate estimation, design, and few-shot effects.

Layer 1. LLM-enhanced efficiency gain. The first layer fixes the classical Neyman allocation and compares the allocation-relevant AIPW criterion with the corresponding LLM-enhanced criterion. Thus, it isolates the gain from calibrated auxiliary adjustment while holding the treatment allocation rule fixed. At the oracle level, this gain is nonnegative because calibrated adjustment replaces the raw within-stratum variance $v_t(z)$ by the effective residual variance $\tilde{v}_t(z) = v_t(z)\{1 - \rho_t^2(z)\} \leq v_t(z)$. It is therefore an estimation gain rather than a design gain. The allocation rule is unchanged, but the influence-function residual variance entering the allocation-relevant criterion is reduced.

Layer 2. Design-enhanced efficiency gain. The second layer fixes the zero-shot LLM-enhanced criterion and compares the classical Neyman allocation with the LLM-enhanced Neyman allocation. This isolates the gain from using effective residual variances rather than raw outcome variances to guide adaptive allocation. At the oracle level, the comparison is nonnegative because $\boldsymbol{e}^{*,\mathrm{zs}}_{\mathrm{I}}$ minimizes the zero-shot allocation-relevant criterion over the admissible allocation class. This layer is therefore a design gain.

Layer 3. Few-shot-enhanced efficiency gain. The third layer contrasts zero-shot and few-shot auxiliary signals under their corresponding LLM-enhanced Neyman allocations. It captures the benefit of the few-shot construction when prompt demonstrations improve the auxiliary signal. Unlike the first two layers, this term is not guaranteed to be nonnegative for arbitrary prompts. Few-shot demonstrations may help when they provide relevant arm- and stratum-specific examples, but they may add noise or induce miscalibration when the examples are uninformative, poorly matched, or spuriously predictive. For this reason, the third layer should be interpreted as a prompt-dependent few-shot component rather than a guaranteed improvement.

Connection to Objective II. The same layer names are used in the Objective II welfare analysis, but the interpretation differs because Objective II minimizes the failure-rate objective $\mathcal{L}_{\mathrm{F}}(\boldsymbol{e})$ subject to the precision constraint $\mathcal{V}(\boldsymbol{e}) \leq C$. In this setting, Layer 1 is a constraint-side effect, in which the LLM-enhanced efficiency gain reduces the allocation-relevant variance criterion and creates precision

slack. Layer 2 is a welfare-side design effect, in which the constrained optimizer uses this slack to move allocation probabilities toward lower-failure arms while maintaining the same bound C. Layer 3 is a few-shot-side welfare effect that arises when the few-shot construction further improves or stabilizes the auxiliary signal used in the constraint, thereby increasing the usable slack and improving the welfare allocation. Therefore, in Objective II, the layers are mechanism components rather than an additive variance decomposition. Layer 1 is measured on the constraint side, whereas Layers 2–3 are measured on the welfare-objective side.

A.3 Implementation details for calibration primitives

Eq. (2) defines the population projection slope, and Section 3 gives its cell-level sample version. In the implementation, the final estimator estimates the corresponding nuisance components with 2-fold cross-fitting, so each target observation uses nuisance fits built outside its fold. These nuisance components are random forests on the categorical baseline covariates, so the fitted $\widehat{\mu}_t$, $\widehat{\mu}_t^{\dagger}$, and $\widehat{\omega}_t$ may vary within strata. For sparse or noisy arm–stratum cells, we stabilize the projection estimate by applying a denominator floor of 0.1 to the auxiliary-signal variance and clipping the estimated projection slope to $[-5, 5]$. Section 5 analyzes the stratum-level calibration of Section 3. In the superpopulation of Section 6, the stratum-level auxiliary-signal variances exceed 0.25 and the population projection slopes lie in $[-4, 4]$, so the floor and the clipping do not bind at the population level.

A.4 Chronological algorithm

Algorithm 1 states when each prompt, nuisance estimate, and allocation update is formed. Two restrictions govern every demonstration used for subject i in stage s. Trial chronology requires each demonstration to have an outcome observed in $\mathcal{H}_{s-1}$, and cross-fitting requires it to lie outside the fold of subject i. The eligible arm-t pool is therefore $\mathcal{P}_{i,t} = \{j \in \mathcal{O}_{s-1} : T_j = t,\ k_j \neq k_i\}$, where k_j is the fold label of subject j. Each aggregated signal $Y_i^{\dagger}(t)$ and each stage-s allocation rule is therefore fixed before treatment assignment, and no current-stage or future outcome enters a current signal or allocation decision. The same stored signal is used in the allocation updates and in the final estimator. All fold labels are drawn before enrollment, independently of all other data, so they belong to $\mathcal{G}_0$.

Algorithm 1 LLM-enhanced CARA with few-shot auxiliary signals

Require: stage sizes $n_1, \ldots, n_S$; folds K; demonstrations per prompt m; resamples B; truncation δ; Objective I or II

Draw fold labels $k_1, \ldots, k_N \in \{1, \ldots, K\}$ independently of all other data.
for $s = 1, \ldots, S$ **do**
  If $s = 1$, set $e_1(z) = 1/2$. Otherwise compute $\widehat{e}_s(\cdot)$ from $\mathcal{H}_{s-1}$ by Steps 1–2 of Section 3 and truncate it to $[\delta, 1 - \delta]$.
  for each subject i enrolled in stage s **do**
    Record W_i.
    for $t \in \{0, 1\}$ and $b = 1, \ldots, B$ **do**
      Draw $\mathcal{S}_{i,t}^{(b)}$: m subjects without replacement from $\mathcal{P}_{i,t}$.
      Query the LLM with the template in Appendix A.5 and parse $Y_i^{\dagger}(t; \mathcal{S}_{i,t}^{(b)})$.
    end for
    Set $Y_i^{\dagger}(t) = B^{-1} \sum_{b=1}^{B} Y_i^{\dagger}(t; \mathcal{S}_{i,t}^{(b)})$ for $t = 0, 1$.
    Assign $T_i \sim \text{Bernoulli}\{e_s(Z_i)\}$ and archive the prompts, outputs, and $e_s(Z_i)$.
  end for
  Add the outcomes observed by the end of stage s to form $\mathcal{H}_s$.
end for
Inference. For each fold k, estimate $\widehat{\mu}_t$, $\widehat{\mu}_t^{\dagger}$, and $\widehat{\omega}_t$ from subjects outside fold k. Compute $\widehat{\varphi}_t$ for subjects in fold k using the stored $Y_i^{\dagger}(t)$ and design propensities $e_{i,t}$, and return $\widehat{\tau}$ and $\widehat{\mathbb{V}}$.

If $\mathcal{P}_{i,t}$ contains fewer than m subjects, as in stage 1, the demonstrations are taken from a prespecified pool. Assumption 5.4(ii) then requires the pools used at different stages to yield the same limiting score, which holds, for example, when one fixed pool is used at every stage, as in Section 6.

A.5 Prompt templates and decoding settings

All prompts for the BRIGHTEN-calibrated superpopulation use the system message "You are a licensed clinical psychologist specializing in depression. Return STRICT JSON only." Each baseline record is serialized by the following context block, where a missing free-text response is shown as "(not provided)".

```
- Age: {age}
- Gender: {gender}
- Education: {education}
- Working: {working}
- Marital status: {marital_status}
- Race/ethnicity: {race}
- App satisfaction / initial expectations: "{free-text response 1}"
- Reason for enrollment: "{free-text response 2}"
```

Zero-shot prompt. One call returns predictions for both arms.

```
You are a licensed clinical psychologist with expertise in depression treatment outcomes. A participant
     enrolled in the BRIGHTEN study, a smartphone-based depression management trial. Using all available
     baseline information - demographics, age, and the participant's own words about their expectations
     and reasons for joining - predict the expected end-of-study PHQ-9 total score (0-27) under two
     conditions.

Clinical guidance:
- PHQ-9 ranges: 0-4 Minimal, 5-9 Mild, 10-14 Moderate, 15-19 Moderately Severe, 20-27 Severe.
- Consider how the participant's language reflects their current severity, coping capacity, and
     motivation.
- Participants expressing desperation, survival language, or severe functional impairment tend toward
     worse outcomes.
- Participants expressing curiosity, specific goals, or active coping strategies tend toward better
     outcomes.
- The iPST app (arm 1) provides problem-solving therapy and typically leads to modest improvement
     relative to control (arm 0).

Participant baseline information:
{context block}

Return STRICT JSON with exactly these keys:
{ "arm_0_phq9": <integer 0-27>, "arm_1_phq9": <integer 0-27> }
where:
- arm_0_phq9 = predicted end-of-study PHQ-9 under control/other arms
- arm_1_phq9 = predicted end-of-study PHQ-9 under iPST app
Return JSON only. No explanation.
```

Few-shot prompt for arm t. The arm name is "control (alternative apps)" for $t = 0$ and "iPST app (treatment)" for $t = 1$, and $j = 1, \ldots, m$ indexes the demonstrations.

```
You are a licensed clinical psychologist with expertise in depression treatment outcomes.
Given a participant's baseline information, predict their end-of-study PHQ-9 total score (0-27) under the
      {arm name} condition.

PHQ-9 ranges: 0-4 Minimal, 5-9 Mild, 10-14 Moderate, 15-19 Moderately Severe, 20-27 Severe.

Here are some example participants with their observed outcomes:

Example {j}:
{context block of demonstration j}
  Condition: {arm name}
  Observed PHQ-9: {outcome of demonstration j}

Now predict for THIS participant:
{context block of the target subject}
  Condition: {arm name}

Return STRICT JSON: {"phq9": <integer 0-27>}
No explanation.
```

Decoding and parsing. We use Qwen3-8B with local weights and its chat template with thinking disabled, and allow up to 128 new tokens. The zero-shot prompt uses greedy decoding, and the few-shot prompt samples at temperature 0.6 with top-p 0.95 and top-k 20. The parser reads the JSON fields, falls back to a regular-expression match of the same fields, and truncates the value to $[0, 27]$. A generation call that returns an error is retried, with at most five attempts in total, and a few-shot resample whose output cannot be parsed is excluded from the average.

B Derivation of $\mathcal{V}(\boldsymbol{e})$ in Eq. (11)

This section derives the oracle influence-function representation of the estimator in Section 3. It also derives the oracle variance functional $\mathcal{V}(\boldsymbol{e})$ in Eq. (11). Throughout this appendix, $Y^\dagger(t)$ denotes the auxiliary signal that defines the population quantities, which is $f_{\mathrm{FS},t}(W)$ under Assumption 5.4, so that $(W_i, Y_i(0), Y_i(1), Y_i^\dagger(0), Y_i^\dagger(1))$, $i = 1, \ldots, N$, are independent and identically distributed.

Fix a nonrandom policy $\boldsymbol{e} = \{e_s(\cdot)\}_{s=1}^S$ with values in $[\delta, 1-\delta]$. For each subject i, define the realized design propensity

$$e_i := e_{S_i}(Z_i), \quad e_{i,1} := e_i, \quad e_{i,0} := 1 - e_i.$$

Recall that $O_i = (W_i, T_i, Y_i)$ is the observed data for subject i at the end of the trial.

For each arm $t \in \{0, 1\}$ and stratum z, recall

$$\mu_t(z) = \mathbb{E}\big[Y(t) \mid Z = z\big], \quad \mu_t^\dagger(z) = \mathbb{E}\big[Y^\dagger(t) \mid Z = z\big],$$

and define the centered auxiliary signal

$$\widetilde{Y}^\dagger(t) := Y^\dagger(t) - \mu_t^\dagger(Z).$$

Define the within-stratum projection slope

$$\omega_t(z) := \frac{\mathbb{C}\mathrm{ov}\big(Y(t), \widetilde{Y}^\dagger(t) \mid Z = z\big)}{\mathbb{V}\mathrm{ar}\big(\widetilde{Y}^\dagger(t) \mid Z = z\big)},$$

with $\omega_t(z) = 0$ if the denominator equals 0.

Define the within-stratum working regression function

$$m_t\big(z, y^\dagger\big) := \mu_t(z) + \omega_t(z)\Big(y^\dagger - \mu_t^\dagger(z)\Big). \tag{14}$$

The corresponding oracle arm-specific influence function is

$$\varphi_t\big(O_i, Y_i^\dagger(t)\big) := m_t\big(Z_i, Y_i^\dagger(t)\big) + \frac{\mathbb{1}\{T_i = t\}}{e_{i,t}}\Big\{Y_i - m_t\big(Z_i, Y_i^\dagger(t)\big)\Big\}. \tag{15}$$

Under Assumptions 5.2 and 5.4(i), $\mathbb{E}[\varphi_t(O_i, Y_i^\dagger(t)) \mid Z_i = z] = \mu_t(z)$, hence

$$\mathbb{E}\Big[\varphi_1\big(O_i, Y_i^\dagger(1)\big) - \varphi_0\big(O_i, Y_i^\dagger(0)\big)\Big] = \tau.$$

Therefore, under Assumption 5.3 with cross-fitting, as shown in Step 1 of Appendix G,

$$\widehat{\tau} - \tau = \frac{1}{N}\sum_{i=1}^N \Big(\varphi_1\big(O_i, Y_i^\dagger(1)\big) - \varphi_0\big(O_i, Y_i^\dagger(0)\big) - \tau\Big) + o_p(N^{-1/2}). \tag{16}$$

Define the within-stratum outcome variance and correlation

$$\begin{aligned} v_t(z) :=& \mathbb{V}\mathrm{ar}\big(Y(t) \mid Z = z\big), \\ \rho_t(z) :=& \mathbb{C}\mathrm{orr}\big(Y(t), Y^\dagger(t) \mid Z = z\big) = \mathbb{C}\mathrm{orr}\big(Y(t), \widetilde{Y}^\dagger(t) \mid Z = z\big), \end{aligned}$$

with $\rho_t(z) = 0$ if $\mathbb{V}\mathrm{ar}(\widetilde{Y}^\dagger(t) \mid Z = z) = 0$. Define the effective residual variance

$$\tilde{v}_t(z) := \mathbb{V}\mathrm{ar}\big(Y(t) - \omega_t(z)\widetilde{Y}^\dagger(t) \mid Z = z\big).$$

Lemma B.1 (Projection identity)**.** *For each $t \in \{0, 1\}$ and stratum z with $p(z) > 0$,*

$$\tilde{v}_t(z) = v_t(z)\big\{1 - \rho_t^2(z)\big\}, \quad \omega_t^2(z)\mathbb{V}\mathrm{ar}\big(\widetilde{Y}^\dagger(t) \mid Z = z\big) = v_t(z)\rho_t^2(z).$$

Proof. Fix t and z. If $\mathbb{V}\mathrm{ar}(\widetilde{Y}^\dagger(t) \mid Z = z) = 0$, then $\omega_t(z) = \rho_t(z) = 0$ and both identities hold trivially, so assume this variance is positive. By definition, $\omega_t(z)$ is the slope of the linear projection of $Y(t)$ on $\widetilde{Y}^\dagger(t)$ given $Z = z$. The projection residual is orthogonal to $\widetilde{Y}^\dagger(t)$ given $Z = z$. Therefore,

$$\tilde{v}_t(z) = \mathbb{V}\mathrm{ar}\big(Y(t) - \mu_t(z) \mid Z = z\big) - \frac{\mathbb{C}\mathrm{ov}^2\big(Y(t), \widetilde{Y}^\dagger(t) \mid Z = z\big)}{\mathbb{V}\mathrm{ar}\big(\widetilde{Y}^\dagger(t) \mid Z = z\big)}.$$

Write

$$\mathbb{C}\mathrm{ov}\big(Y(t), \widetilde{Y}^\dagger(t) \mid Z = z\big) = \rho_t(z)\sqrt{v_t(z)}\sqrt{\mathbb{V}\mathrm{ar}\big(\widetilde{Y}^\dagger(t) \mid Z = z\big)}.$$

Substituting gives $\tilde{v}_t(z) = v_t(z)\{1 - \rho_t^2(z)\}$. For the second identity, use $\omega_t(z) = \frac{\mathbb{C}\mathrm{ov}\left(Y(t), \widetilde{Y}^\dagger(t) \mid Z=z\right)}{\mathbb{V}\mathrm{ar}\left(\widetilde{Y}^\dagger(t) \mid Z=z\right)}$ and rearrange. □

Define

$$\psi_i := \varphi_1\big(O_i, Y_i^\dagger(1)\big) - \varphi_0\big(O_i, Y_i^\dagger(0)\big) - \tau.$$

By Eq. (16), the oracle asymptotic variance of $\sqrt{N}(\widehat{\tau} - \tau)$ is $\mathbb{V}\mathrm{ar}(\psi_i)$. We compute $\mathbb{V}\mathrm{ar}(\psi_i)$ by conditioning on stage and stratum. Recall that $r_s = n_s/N$, so a subject drawn uniformly from the trial has $\Pr(S_i = s) = r_s$.

First,

$$\mathbb{E}[\psi_i \mid Z_i = z, S_i = s] = \tau(z) - \tau,$$

since $\mathbb{E}[m_t(Z, Y^\dagger(t)) \mid Z = z] = \mu_t(z)$ and the inverse-propensity residual term has conditional mean zero. Thus,

$$\mathbb{V}\mathrm{ar}(\psi_i) = \mathbb{E}\big[(\tau(Z) - \tau)^2\big] + \sum_{s=1}^{S} r_s \mathbb{E}\Big[\mathbb{V}\mathrm{ar}(\psi_i \mid Z, S = s)\Big]. \tag{17}$$

Fix stage s and stratum z. Write $e = e_s(z)$. From Eq. (15),

$$\varphi_1 - \mu_1(z) = \omega_1(z)\widetilde{Y}^\dagger(1) + \frac{\mathbb{1}\{T = 1\}}{e}\Big(Y(1) - \mu_1(z) - \omega_1(z)\widetilde{Y}^\dagger(1)\Big),$$

$$\varphi_0 - \mu_0(z) = \omega_0(z)\widetilde{Y}^\dagger(0) + \frac{\mathbb{1}\{T = 0\}}{1 - e}\Big(Y(0) - \mu_0(z) - \omega_0(z)\widetilde{Y}^\dagger(0)\Big).$$

Define the projection residuals

$$\epsilon_t(z) := Y(t) - \mu_t(z) - \omega_t(z)\widetilde{Y}^\dagger(t), \quad \mathbb{V}\mathrm{ar}\big(\epsilon_t(z) \mid Z = z\big) = \tilde{v}_t(z).$$

Lemma B.1 implies $\mathbb{C}\mathrm{ov}(\widetilde{Y}^\dagger(t), \epsilon_t(z) \mid Z = z) = 0$. Under Assumptions 5.2 and 5.4(i), $T \mid (Z = z, S = s)$ is Bernoulli with success probability e and is independent of $(Y(0), Y(1), Y^\dagger(0), Y^\dagger(1))$ given (Z, S).

Let

$$A_1 := \omega_1(z)\widetilde{Y}^\dagger(1), \quad A_0 := \omega_0(z)\widetilde{Y}^\dagger(0), \quad Q_1 := \frac{\mathbb{1}\{T = 1\}}{e}\epsilon_1(z), \quad Q_0 := \frac{\mathbb{1}\{T = 0\}}{1 - e}\epsilon_0(z).$$

Then $\varphi_1 - \mu_1(z) = A_1 + Q_1$ and $\varphi_0 - \mu_0(z) = A_0 + Q_0$.

Step 1. Conditional variances of each arm. A direct calculation using sequential randomization gives

$$\mathbb{V}\mathrm{ar}(Q_1 \mid Z = z, S = s) = \frac{\tilde{v}_1(z)}{e}, \quad \mathbb{V}\mathrm{ar}(Q_0 \mid Z = z, S = s) = \frac{\tilde{v}_0(z)}{1 - e},$$

and $\mathbb{C}\mathrm{ov}(A_1, Q_1 \mid Z = z, S = s) = \mathbb{C}\mathrm{ov}(A_0, Q_0 \mid Z = z, S = s) = 0$ because T is independent of $(\widetilde{Y}^\dagger(t), \epsilon_t(z))$ given (Z, S) and $\mathbb{C}\mathrm{ov}(\widetilde{Y}^\dagger(t), \epsilon_t(z) \mid Z = z) = 0$. Therefore,

$$\begin{aligned}\mathbb{V}\mathrm{ar}(\varphi_1 \mid Z = z, S = s) =& \mathbb{V}\mathrm{ar}(A_1 \mid Z = z) + \mathbb{V}\mathrm{ar}(Q_1 \mid Z = z, S = s) \\ =& \omega_1^2(z)\mathbb{V}\mathrm{ar}(\widetilde{Y}^\dagger(1) \mid Z = z) + \frac{\tilde{v}_1(z)}{e},\end{aligned}$$

$$\mathbb{V}\mathrm{ar}(\varphi_0 \mid Z = z, S = s) = \omega_0^2(z)\mathbb{V}\mathrm{ar}(\widetilde{Y}^\dagger(0) \mid Z = z) + \frac{\tilde{v}_0(z)}{1 - e}.$$

By Lemma B.1, $\omega_t^2(z)\mathbb{V}\mathrm{ar}(\widetilde{Y}^\dagger(t) \mid Z = z) = v_t(z)\rho_t^2(z)$, hence

$$\begin{aligned}\mathbb{V}\mathrm{ar}(\varphi_1 \mid Z = z, S = s) =& \frac{v_1(z)}{e}\Big\{1 - (1 - e)\rho_1^2(z)\Big\}, \\ \mathbb{V}\mathrm{ar}(\varphi_0 \mid Z = z, S = s) =& \frac{v_0(z)}{1 - e}\Big\{1 - e\rho_0^2(z)\Big\}.\end{aligned}$$

Step 2. Cross-arm covariance term. We now compute $\mathbb{C}\mathrm{ov}(\varphi_1, \varphi_0 \mid Z = z, S = s)$ without imposing orthogonality. Using $\varphi_1 - \mu_1(z) = A_1 + Q_1$ and $\varphi_0 - \mu_0(z) = A_0 + Q_0$, we have

$$\mathbb{C}\mathrm{ov}(\varphi_1, \varphi_0 \mid Z = z, S = s) = \mathbb{C}\mathrm{ov}(A_1 + Q_1,\ A_0 + Q_0 \mid Z = z, S = s).$$

Because $\mathbb{1}\{T = 1\}\mathbb{1}\{T = 0\} = 0$ and $\mathbb{E}[\epsilon_t(z) \mid Z = z] = 0$, one has $\mathbb{C}\mathrm{ov}(Q_1, Q_0 \mid Z = z, S = s) = 0$. Moreover,

$$\begin{aligned}\mathbb{C}\mathrm{ov}(A_1, Q_0 \mid Z = z, S = s) =& \omega_1(z)\mathbb{C}\mathrm{ov}\left(\widetilde{Y}^\dagger(1), \frac{\mathbb{1}\{T = 0\}}{1 - e}\epsilon_0(z) \;\middle|\; Z = z, S = s\right) \\ =& \omega_1(z)\mathbb{C}\mathrm{ov}\big(\widetilde{Y}^\dagger(1), \epsilon_0(z) \mid Z = z\big),\end{aligned}$$

since $\mathbb{E}[\mathbb{1}\{T = 0\}/(1 - e) \mid Z = z, S = s] = 1$ and T is independent of $(\widetilde{Y}^\dagger(1), \epsilon_0(z))$ given (Z, S). Similarly,

$$\begin{aligned}\mathbb{C}\mathrm{ov}(Q_1, A_0 \mid Z = z, S = s) =& \omega_0(z)\mathbb{C}\mathrm{ov}\big(\widetilde{Y}^\dagger(0), \epsilon_1(z) \mid Z = z\big), \\ \mathbb{C}\mathrm{ov}(A_1, A_0 \mid Z = z) =& \omega_1(z)\omega_0(z)\mathbb{C}\mathrm{ov}\big(\widetilde{Y}^\dagger(1), \widetilde{Y}^\dagger(0) \mid Z = z\big).\end{aligned}$$

Collecting terms,

$$\mathbb{C}\mathrm{ov}(\varphi_1, \varphi_0 \mid Z = z, S = s) =: \kappa(z), \tag{18}$$

where the stratum-specific cross-arm covariance primitive is

$$\begin{aligned}\kappa(z) =& \omega_1(z)\omega_0(z)\mathbb{C}\mathrm{ov}\big(\widetilde{Y}^\dagger(1), \widetilde{Y}^\dagger(0) \mid Z = z\big) + \omega_1(z)\mathbb{C}\mathrm{ov}\big(\widetilde{Y}^\dagger(1), \epsilon_0(z) \mid Z = z\big) \\ &+ \omega_0(z)\mathbb{C}\mathrm{ov}\big(\widetilde{Y}^\dagger(0), \epsilon_1(z) \mid Z = z\big).\end{aligned} \tag{19}$$

The factors involving e cancel in the expectations above, so $\kappa(z)$ does not depend on e.

Step 3. Conditional variance of the ATE influence contribution. Since $\psi_i = \varphi_1 - \varphi_0 - \tau$ and τ is a constant, we have

$$\mathbb{V}\mathrm{ar}(\psi_i \mid Z = z, S = s) = \mathbb{V}\mathrm{ar}(\varphi_1 - \varphi_0 \mid Z = z, S = s).$$

Therefore,

$$\begin{aligned}\mathbb{V}\mathrm{ar}(\psi_i \mid Z = z, S = s) =& \mathbb{V}\mathrm{ar}(\varphi_1 \mid Z = z, S = s) + \mathbb{V}\mathrm{ar}(\varphi_0 \mid Z = z, S = s) \\ &- 2\mathbb{C}\mathrm{ov}(\varphi_1, \varphi_0 \mid Z = z, S = s) \\ =& \frac{v_1(z)}{e}\Big\{1 - (1 - e)\rho_1^2(z)\Big\} + \frac{v_0(z)}{1 - e}\Big\{1 - e\rho_0^2(z)\Big\} - 2\kappa(z).\end{aligned} \tag{20}$$

Step 4. Oracle variance functional. In Eq. (20), $e = e_s(z)$. Using $\tilde{v}_t(z) = v_t(z)\{1 - \rho_t^2(z)\}$,

$$\frac{v_1(z)}{e}\{1 - (1 - e)\rho_1^2(z)\} = \frac{\tilde{v}_1(z)}{e} + v_1(z)\rho_1^2(z),$$

and

$$\frac{v_0(z)}{1 - e}\{1 - e\rho_0^2(z)\} = \frac{\tilde{v}_0(z)}{1 - e} + v_0(z)\rho_0^2(z).$$

Therefore the full oracle variance is

$$\begin{aligned}\sigma^2(\boldsymbol{e}) =& \mathbb{E}\{(\tau(Z) - \tau)^2\} + \sum_{s=1}^{S} r_s \mathbb{E}\left[\frac{\tilde{v}_1(Z)}{e_s(Z)} + \frac{\tilde{v}_0(Z)}{1 - e_s(Z)}\right] \\ &+ \sum_{s=1}^{S} r_s \mathbb{E}\left[v_1(Z)\rho_1^2(Z) + v_0(Z)\rho_0^2(Z) - 2\kappa(Z)\right].\end{aligned}$$

Equivalently,

$$\sigma^2(\boldsymbol{e}) = \mathcal{V}(\boldsymbol{e}) + C_{\mathrm{inv}},$$

where

$$C_{\mathrm{inv}} = \mathbb{E}\{(\tau(Z) - \tau)^2\} + \sum_{s=1}^{S} r_s \mathbb{E}\left[v_1(Z)\rho_1^2(Z) + v_0(Z)\rho_0^2(Z) - 2\kappa(Z)\right].$$

Since C_{inv} does not depend on the allocation policy $\boldsymbol{e}$, minimizing the full oracle variance $\sigma^2(\boldsymbol{e})$ over $\boldsymbol{e}$ is equivalent to minimizing the allocation-relevant criterion $\mathcal{V}(\boldsymbol{e})$. The estimator's full inference variance is $\sigma^2(\boldsymbol{e})$, while $\mathcal{V}(\boldsymbol{e})$ is used only for allocation optimization and gain decomposition.

C Derivation of the Objective I allocation in Eq. (5)

This section derives the oracle minimizer of Eq. (4).

Proof. Fix a stratum z with $p(z) > 0$ and a stage s. Consider the part of $\sigma^2(\boldsymbol{e})$ in Appendix B that depends on $e_s(z)$. Let $e := e_s(z)$. By Eq. (20), up to the positive factor $r_s p(z)$ and terms that do not depend on e, this part is

$$\frac{v_1(z)}{e}\Big\{1-(1-e)\rho_1^2(z)\Big\}+\frac{v_0(z)}{1-e}\Big\{1-e\rho_0^2(z)\Big\}.$$

Expand each bracket:

$$\frac{v_1(z)}{e}\Big(1-\rho_1^2(z)+e\rho_1^2(z)\Big)=\frac{v_1(z)\{1-\rho_1^2(z)\}}{e}+v_1(z)\rho_1^2(z),$$

$$\frac{v_0(z)}{1-e}\Big(1-\rho_0^2(z)+\rho_0^2(z)(1-e)\Big)=\frac{v_0(z)\{1-\rho_0^2(z)\}}{1-e}+v_0(z)\rho_0^2(z).$$

Dropping the additive constants $v_1(z)\rho_1^2(z)$ and $v_0(z)\rho_0^2(z)$, the minimization over $e \in [\delta, 1-\delta]$ reduces to

$$\min_{e\in[\delta,1-\delta]}\left\{\frac{\tilde{v}_1(z)}{e}+\frac{\tilde{v}_0(z)}{1-e}\right\},\quad \tilde{v}_t(z)=v_t(z)\{1-\rho_t^2(z)\}.$$

On $(0,1)$ the objective is strictly convex because $\tilde{v}_t(z) \geq \varepsilon$ by Assumption 5.1. Setting the derivative to zero gives

$$-\frac{\tilde{v}_1(z)}{e^2}+\frac{\tilde{v}_0(z)}{(1-e)^2}=0 \quad\Longrightarrow\quad \sqrt{\tilde{v}_0(z)}e=\sqrt{\tilde{v}_1(z)}(1-e),$$

hence

$$e_{\mathrm{I}}^*(z)=\frac{\sqrt{\tilde{v}_1(z)}}{\sqrt{\tilde{v}_1(z)}+\sqrt{\tilde{v}_0(z)}}.$$

Because the objective is strictly convex, its minimizer over $[\delta, 1-\delta]$ is this value truncated to $[\delta, 1-\delta]$, which is Eq. (5) under the truncation convention of Section 2.4. □

D Duality and separability for the Objective II allocation update

This section justifies the pointwise characterization in Eq. (10).

Proposition D.1 (Duality and separability for Objective II)**.** *Fix a stage $s \in \{1, \ldots, S-1\}$. Let $\widehat{p}_s(z) \geq 0$ with $\sum_{z=1}^{J} \widehat{p}_s(z) = 1$. Define*

$$J_s(e) := \sum_{z=1}^{J} \widehat{p}_s(z)\Big\{e(z)\widehat{\mu}_{1,s}^{\mathrm{F}}(z)+\big(1-e(z)\big)\widehat{\mu}_{0,s}^{\mathrm{F}}(z)\Big\},$$

and

$$G_s(e) := \sum_{z=1}^{J} \widehat{p}_s(z)\Big\{\frac{\widehat{\tilde{v}}_{1,s}(z)}{e(z)}+\frac{\widehat{\tilde{v}}_{0,s}(z)}{1-e(z)}\Big\}.$$

Consider the constrained problem

$$\begin{aligned}\min_{e(\cdot)}\quad & J_s(e)\\ \text{s.t.}\quad & e(z)\in[\delta,1-\delta],\ \forall z\in\{1,\ldots,J\},\\ & G_s(e)\leq C_s.\end{aligned}\tag{21}$$

Assume Slater's condition, that is, there exists $\bar{e}(\cdot)$ with $\bar{e}(z) \in (\delta, 1-\delta)$ for all z and $G_s(\bar{e}) < C_s$. Then strong duality holds for Eq. (21)*, and there exists $\lambda_s^* \geq 0$ such that any primal optimizer $\widehat{e}_{s+1,\mathrm{II}}(\cdot)$ satisfies, for each z with $\widehat{p}_s(z) > 0$,*

$$\widehat{e}_{s+1,\mathrm{II}}(z) \in \operatorname*{argmin}_{e\in[\delta,1-\delta]} \left\{ e\big(\widehat{\mu}_{1,s}^{\mathrm{F}}(z) - \widehat{\mu}_{0,s}^{\mathrm{F}}(z)\big) + \lambda_s^* \Big(\frac{\widehat{\widehat{v}}_{1,s}(z)}{e} + \frac{\widehat{\widehat{v}}_{0,s}(z)}{1-e} \Big) \right\}. \tag{22}$$

Moreover, λ_s^* satisfies the complementary slackness condition

$$\lambda_s^* \big(G_s(\widehat{e}_{s+1,\mathrm{II}}) - C_s \big) = 0.$$

Proof. For each fixed z, the map $e \mapsto \widehat{\widehat{v}}_{1,s}(z)/e + \widehat{\widehat{v}}_{0,s}(z)/(1-e)$ is convex on $(0,1)$. Therefore $G_s(e)$ is convex and $J_s(e)$ is affine. The box constraint is convex, hence Eq. (21) is a convex program. Slater's condition implies strong duality and KKT optimality.

For $\lambda \geq 0$, the Lagrangian $\mathcal{L}(e,\lambda) = J_s(e) + \lambda\{G_s(e) - C_s\}$ equals

$$\mathcal{L}(e,\lambda) = -\lambda C_s + \sum_{z=1}^{J} \widehat{p}_s(z) \Big[e(z)\widehat{\mu}_{1,s}^{\mathrm{F}}(z) + (1-e(z))\widehat{\mu}_{0,s}^{\mathrm{F}}(z) + \lambda \Big(\frac{\widehat{\widehat{v}}_{1,s}(z)}{e(z)} + \frac{\widehat{\widehat{v}}_{0,s}(z)}{1-e(z)} \Big) \Big].$$

For fixed λ, the minimization over $e(z) \in [\delta, 1-\delta]$ is separable across z, which yields Eq. (22). Complementary slackness follows from KKT conditions. □

Under Assumptions 5.1 and 5.3, Slater's condition holds with probability tending to one. Appendix F gives an interior oracle point whose oracle constraint value is below C, and G_s converges uniformly on $[\delta, 1-\delta]^J$ to its oracle counterpart while $C_s \xrightarrow{\mathrm{p}} C$.

E Proof of Lemma 5.5

Proof. Fix $t \in \{0,1\}$ and decompose the remainder as

$$R_{i,t,B} = Y_i^\dagger(t) - f_{\mathrm{FS},t}(W_i) = \big\{ Y_i^\dagger(t) - \bar{f}_{i,t}(W_i) \big\} + \big\{ \bar{f}_{i,t}(W_i) - f_{\mathrm{FS},t}(W_i) \big\}.$$

By Assumption 5.4(ii), given W_i and the demonstration pool, $Y_i^\dagger(t)$ is the average of B independent prompt-level outputs with common mean $\bar{f}_{i,t}(W_i)$. By Assumption 5.4(i), each output lies in $[0, M_\dagger]$ and therefore has conditional variance at most $M_\dagger^2/4$. The conditional mean square of the first term is B^{-2} times the sum of these B variances, so

$$\mathbb{E}\Big[\frac{1}{N} \sum_{i=1}^{N} \big\{ Y_i^\dagger(t) - \bar{f}_{i,t}(W_i) \big\}^2 \Big] \leq \frac{M_\dagger^2}{4B},$$

and Markov's inequality makes this average $O_p(B^{-1})$. Assumption 5.4(ii) states that the average square of the second term is $o_p(1)$. Since $(a+b)^2 \leq 2a^2 + 2b^2$,

$$\frac{1}{N} \sum_{i=1}^{N} R_{i,t,B}^2 = O_p(B^{-1}) + o_p(1),$$

which is $o_p(1)$ under Assumption 5.4(iii). □

F Proof of Theorem 5.6

Proof. Fix a stage $s \geq 2$. The rule $\widehat{e}_s$ is computed from stage-$(s-1)$ estimates, which are consistent by Assumption 5.3, and $\widehat{p}_{s-1}(z) \xrightarrow{\mathrm{p}} p(z)$ by the law of large numbers. Strata with $p(z) = 0$ are never enrolled, so all statements below concern strata with $p(z) > 0$, and each such stratum has $\widehat{p}_{s-1}(z) > 0$ with probability tending to one.

Objective I. Fix a stratum z with $p(z) > 0$. Assumption 5.3 gives $\widehat{\tilde{v}}_{t,s-1}(z) \xrightarrow{\mathrm{p}} \tilde{v}_t(z)$, and $\tilde{v}_t(z) \geq \varepsilon$ by Assumption 5.1. The map

$$(a, b) \mapsto \frac{\sqrt{a}}{\sqrt{a} + \sqrt{b}}$$

is continuous on $(0, \infty)^2$, and the truncation map $\Pi_{[\delta,1-\delta]}$ is continuous. Hence by the continuous mapping theorem,

$$\widehat{e}_{s,\mathrm{I}}(z) = \Pi_{[\delta,1-\delta]} \left(\frac{\sqrt{\widehat{\tilde{v}}_{1,s-1}(z)}}{\sqrt{\widehat{\tilde{v}}_{1,s-1}(z)} + \sqrt{\widehat{\tilde{v}}_{0,s-1}(z)}} \right) \xrightarrow{\mathrm{p}} e^*_{\mathrm{I}}(z),$$

which is the minimizer of Eq. (4) by Appendix C.

Objective II. For a rule $e(\cdot)$ with values in $[\delta, 1-\delta]$, write $\mathcal{L}_{\mathrm{F}}(e)$ and $\mathcal{V}(e)$ for the criteria in Eqs. (6) and (11) of the policy that uses e at every stage, namely

$$\mathcal{L}_{\mathrm{F}}(e) = \mathbb{E}\big[e(Z)\mu^{\mathrm{F}}_1(Z) + \{1 - e(Z)\}\mu^{\mathrm{F}}_0(Z)\big], \quad \mathcal{V}(e) = \mathbb{E}\left[\frac{\tilde{v}_1(Z)}{e(Z)} + \frac{\tilde{v}_0(Z)}{1 - e(Z)}\right].$$

These are the oracle counterparts of J_s and G_s in Proposition D.1. For any feasible policy of Eq. (7), the stage average $\bar{e} = \sum_s r_s e_s$ has values in $[\delta, 1-\delta]$, satisfies $\mathcal{V}(\bar{e}) \leq \sum_s r_s \mathcal{V}(e_s) \leq C$ by convexity, and has $\mathcal{L}_{\mathrm{F}}(\bar{e}) = \sum_s r_s \mathcal{L}_{\mathrm{F}}(e_s)$ by linearity. Hence Eq. (7) is solved by a rule used at every stage, and e^*_{II} denotes a minimizer of $\mathcal{L}_{\mathrm{F}}(e)$ over rules with $\mathcal{V}(e) \leq C$. Because the design fixes the stage-1 rule, the update in Eq. (9) imposes the bound on each later stage, and the theorem concerns these later stages.

Uniqueness. Because $\tilde{v}_t(z) \geq \varepsilon$, $\mathcal{V}$ is strictly convex in the values $e(z)$, while $\mathcal{L}_{\mathrm{F}}$ is linear. Suppose that two distinct rules e and e' both minimize $\mathcal{L}_{\mathrm{F}}$ subject to $\mathcal{V} \leq C$. Their average $(e + e')/2$ attains the same value of $\mathcal{L}_{\mathrm{F}}$ and has $\mathcal{V} < C$. By continuity of $\mathcal{V}$, the constraint does not bind near $(e+e')/2$, so this average is a local, and by linearity a global, minimizer of $\mathcal{L}_{\mathrm{F}}$ over all rules with values in $[\delta, 1-\delta]$. Because $\mu^{\mathrm{F}}_1(z) \neq \mu^{\mathrm{F}}_0(z)$ for every z by Assumption 5.1, that minimizer takes the value δ when $\mu^{\mathrm{F}}_1(z) > \mu^{\mathrm{F}}_0(z)$ and $1 - \delta$ otherwise. An endpoint of $[\delta, 1-\delta]$ is not the average of two distinct points of the interval, so $e = e'$, a contradiction.

Consistency. Fix a neighborhood of e^*_{II}. By compactness of the set of rules, continuity of $\mathcal{L}_{\mathrm{F}}$ and $\mathcal{V}$, and uniqueness of e^*_{II}, there is $c > 0$ such that every rule outside the neighborhood with $\mathcal{V}(e) \leq C + c$ has $\mathcal{L}_{\mathrm{F}}(e) \geq \mathcal{L}_{\mathrm{F}}(e^*_{\mathrm{II}}) + c$, since otherwise a limit point of violating rules would be a second minimizer. By Appendix C, e^*_{I} minimizes $\mathcal{V}$, and $\mathcal{V}(e^*_{\mathrm{I}}) < C$ by Assumption 5.1, so by continuity some rule with values in $(\delta, 1-\delta)$ has $\mathcal{V} < C$. By convexity, every rule on the segment from e^*_{II} to that rule, other than e^*_{II} itself, has values in $(\delta, 1-\delta)$ and $\mathcal{V} < C$. Continuity of $\mathcal{L}_{\mathrm{F}}$ therefore gives a rule e° on this segment with $\mathcal{V}(e^\circ) < C$ and $\mathcal{L}_{\mathrm{F}}(e^\circ) < \mathcal{L}_{\mathrm{F}}(e^*_{\mathrm{II}}) + c/2$. The plug-in functions J_{s-1} and G_{s-1} of Proposition D.1 converge to $\mathcal{L}_{\mathrm{F}}$ and $\mathcal{V}$ uniformly over rules with values in $[\delta, 1-\delta]$, because these values are bounded away from 0 and 1 and the plug-in estimates are consistent. Since $C_{s-1} \xrightarrow{\mathrm{p}} C$, with probability tending to one $G_{s-1}(e^\circ) < C_{s-1}$, so e° is feasible for Eq. (9) and $J_{s-1}(\widehat{e}_{s,\mathrm{II}}) \leq J_{s-1}(e^\circ)$. Uniform convergence then gives

$$\mathcal{L}_{\mathrm{F}}(\widehat{e}_{s,\mathrm{II}}) \leq \mathcal{L}_{\mathrm{F}}(e^\circ) + o_p(1) < \mathcal{L}_{\mathrm{F}}(e^*_{\mathrm{II}}) + c/2 + o_p(1), \quad \mathcal{V}(\widehat{e}_{s,\mathrm{II}}) \leq C + o_p(1).$$

Hence $\widehat{e}_{s,\mathrm{II}}$ lies in the neighborhood with probability tending to one. Since the neighborhood is arbitrary, $\widehat{e}_{s,\mathrm{II}}(z) \xrightarrow{\mathrm{p}} e^*_{\mathrm{II}}(z)$ for each z with $p(z) > 0$. □

G Proof of Theorem 5.7

Proof. **Step 1. Reduction to the limiting score.** As in Theorem 5.9, a superscript $*$ marks a quantity computed with $f_{\mathrm{FS},t}(W_i)$ in place of $Y_i^\dagger(t)$, so that $\widehat{\varphi}^*_t$, $\widehat{\tau}^*$, and $\widehat{\mathbb{V}}^*$ correspond to $\widehat{\varphi}_t$, $\widehat{\tau}$, and $\widehat{\mathbb{V}}$. The limiting score takes values in $[0, M_\dagger]$ almost surely. Indeed, $\bar{f}_{i,t}(W_i) \in [0, M_\dagger]$ because the prompt-level outputs lie in this interval, so the squared distance from $f_{\mathrm{FS},t}(W_i)$ to $[0, M_\dagger]$ is at most $\{\bar{f}_{i,t}(W_i) - f_{\mathrm{FS},t}(W_i)\}^2$. These squared distances are independent and identically distributed by Assumption 5.1, and their average is $o_p(1)$ by Assumption 5.4(ii), so the strong law of large numbers

forces their mean to be zero. Steps 1–4 of Appendix I show that $\widehat{\tau} - \widehat{\tau}^* = o_p(N^{-1/2})$ and do not use the present theorem. Lemma 5.5 and the bounds in Step 3 there also show that the summands of $\widehat{\mathbb{V}}$ and $\widehat{\mathbb{V}}^*$, namely $\widehat{\varphi}_1 - \widehat{\varphi}_0 - \widehat{\tau}$ and $\widehat{\varphi}_1^* - \widehat{\varphi}_0^* - \widehat{\tau}^*$ at subject i, differ by terms whose average square is $o_p(1)$. Since Step 4 below shows $\widehat{\mathbb{V}}^* = O_p(1)$, the Cauchy–Schwarz inequality gives $\widehat{\mathbb{V}} - \widehat{\mathbb{V}}^* = o_p(1)$. It therefore suffices to prove the theorem for $\widehat{\tau}^*$ and $\widehat{\mathbb{V}}^*$, whose auxiliary signal is the fixed function $f_{\mathrm{FS},t}$ of the baseline record.

Let ψ_i be the oracle influence contribution of Appendix B with $f_{\mathrm{FS},t}(W_i)$ as the auxiliary signal,

$$\psi_i = \varphi_1\big(O_i, f_{\mathrm{FS},1}(W_i)\big) - \varphi_0\big(O_i, f_{\mathrm{FS},0}(W_i)\big) - \tau,$$

and write $D_{i,t} := 1 - \mathbb{1}\{T_i = t\}/e_{i,t}$, so $|D_{i,t}| \le 1/\delta$. Comparing $\widehat{\varphi}_t^*$ with Eq. (15) gives $\widehat{\varphi}_t^* - \varphi_t = D_{i,t}\{\widehat{m}_t^* - m_t\}$, where $\widehat{m}_t^*$ is the working regression in Eq. (14) with the cross-fitted estimates and both functions are evaluated at $(Z_i, f_{\mathrm{FS},t}(W_i))$. The cross-fitted estimates are constant within each stratum–fold cell, so on each cell this difference is an affine function of $f_{\mathrm{FS},t}(W_i)$ with two cell-level coefficients. These coefficients are $o_p(1)$ by Assumption 5.3 and Step 3 of Appendix I. Step 2 there shows that the cell sums of $D_{i,t}$ and of $D_{i,t}f_{\mathrm{FS},t}(W_i)$, divided by N, are $O_p(N^{-1/2})$. Summing over the finitely many cells and both arms gives

$$\widehat{\tau}^* - \tau = \frac{1}{N}\sum_{i=1}^{N} \psi_i + o_p(N^{-1/2}).$$

Step 2. Martingale difference array. By the definition in Section 2.2, $\mathcal{G}_i$ contains all information generated up to subject i, including W_i, $Y_i^\dagger(0)$, $Y_i^\dagger(1)$, T_i, the outcome Y_i even when it is not yet observed, and the stage history $\mathcal{H}_{S_i-1}$. By Algorithm 1, it also contains all fold labels. Hence ψ_i is $\mathcal{G}_i$-measurable. The realized propensity $e_i = e_{S_i}(Z_i)$ is determined by $\mathcal{G}_{i-1}$ and Z_i, and by Assumptions 5.2 and 5.4(i),

$$\mathbb{P}\{T_i = 1 \mid \mathcal{G}_{i-1}, W_i, Y_i(0), Y_i(1)\} = e_i.$$

Because $f_{\mathrm{FS},t}(W_i)$ is a function of W_i, the augmentation term in φ_t has conditional mean zero given $(\mathcal{G}_{i-1}, W_i)$, and $\mathbb{E}[\varphi_t(O_i, f_{\mathrm{FS},t}(W_i)) \mid \mathcal{G}_{i-1}, W_i] = \mathbb{E}[Y_i(t) \mid \mathcal{G}_{i-1}, W_i]$. Because $(W_i, Y_i(0), Y_i(1))$ is independent of $\mathcal{G}_{i-1}$ under Assumption 5.1, taking conditional expectations given $\mathcal{G}_{i-1}$ yields

$$\mathbb{E}[\psi_i \mid \mathcal{G}_{i-1}] = \mathbb{E}\{Y(1) - Y(0)\} - \tau = 0,$$

so $\{\psi_i, \mathcal{G}_i\}_{i=1}^N$ is a martingale difference array.

Step 3. Conditional variance limit. Let $\boldsymbol{e}^*$ be the limiting design, whose stage-1 rule is e_1 and whose stage-s rule for $s \ge 2$ is e_{I}^* under Objective I and e_{II}^* under Objective II. Given $\mathcal{G}_{i-1}$, the rule $e_{S_i}(\cdot)$ is fixed and $(W_i, Y_i(0), Y_i(1))$ has its population distribution. Given also Z_i, T_i is Bernoulli with success probability $e_{S_i}(Z_i)$, independently of $(W_i, Y_i(0), Y_i(1))$. Hence $\mathbb{E}[\psi_i^2 \mid \mathcal{G}_{i-1}] = \mathbb{E}[\{\tau(Z) - \tau\}^2] + \mathbb{E}[\mathbb{V}\mathrm{ar}(\psi_i \mid Z, S = S_i)]$, where the conditional variance is Eq. (20) with $e = e_{S_i}(Z)$ for the realized rule $e_{S_i}(\cdot)$, as in the stage-S_i term of Eq. (17). Averaging over subjects and using $n_s/N = r_s$ shows that $N^{-1}\sum_{i=1}^N \mathbb{E}[\psi_i^2 \mid \mathcal{G}_{i-1}]$ equals the full oracle variance σ^2 of Appendix B evaluated at the realized rules $(\widehat{e}_1, \widehat{e}_2, \ldots, \widehat{e}_S)$. This variance is continuous in rules with values in $[\delta, 1-\delta]$, so Theorem 5.6 gives

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[\psi_i^2 \mid \mathcal{G}_{i-1}] \xrightarrow{\mathrm{p}} \sigma^2(\boldsymbol{e}^*).$$

Assumption 5.1 gives bounded fourth moments of $Y_i(t)$, $f_{\mathrm{FS},t}$ is bounded, and $|D_{i,t}| \le 1/\delta$, so $\sup_i \mathbb{E}\psi_i^4 < \infty$. This implies the conditional Lindeberg condition.

Step 4. Central limit theorem and variance estimator. By Steps 2–3 and the martingale central limit theorem for arrays, which needs no nesting condition because the limit in Step 3 is constant, $N^{-1/2}\sum_{i=1}^N \psi_i \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_*^2)$ with $\sigma_*^2 := \sigma^2(\boldsymbol{e}^*)$, so Step 1 gives $\sqrt{N}(\widehat{\tau}^* - \tau) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_*^2)$. Because $\sup_i \mathbb{E}\psi_i^4 < \infty$, the martingale weak law gives $N^{-1}\sum_{i=1}^N \{\psi_i^2 - \mathbb{E}[\psi_i^2 \mid \mathcal{G}_{i-1}]\} \xrightarrow{\mathrm{p}} 0$, so $N^{-1}\sum_{i=1}^N \psi_i^2 \xrightarrow{\mathrm{p}} \sigma_*^2$. By Step 1, the summand $\widehat{\varphi}_1^* - \widehat{\varphi}_0^* - \widehat{\tau}^*$ of $\widehat{\mathbb{V}}^*$ differs from ψ_i by $\widehat{\tau}^* - \tau$ and,

for each arm, by $D_{i,t}$ times an affine function of $f_{\mathrm{FS},t}(W_i)$ with $o_p(1)$ cell-level coefficients. The average squared difference is therefore $o_p(1)$, because $|D_{i,t}| \le 1/\delta$, $f_{\mathrm{FS},t}$ takes values in $[0, M_\dagger]$, and $\widehat{\tau}^* - \tau = o_p(1)$. The Cauchy–Schwarz inequality then gives $\widehat{\mathbb{V}}^* \xrightarrow{\mathrm{p}} \sigma_*^2$, and Step 1 transfers both limits to $\widehat{\tau}$ and $\widehat{\mathbb{V}}$. The allocation-relevant part of σ_*^2 is $\mathcal{V}(\boldsymbol{e}^*)$, but the full variance used for inference also includes allocation-invariant heterogeneity and covariance terms. □

G.1 Fixed number of prompt resamples

Theorems 5.6–5.8 also hold for a fixed B, such as $B = 4$, if Assumption 5.4(iii) is replaced by two conditions on the stored aggregate itself. First, given W_i and the demonstration pool, the prompt-level outputs are conditionally independent of $(Y_i(0), Y_i(1))$, as when they are computed from W_i and the pool alone. Second, let $\Sigma_i^\dagger$ be the conditional covariance matrix of $(Y_i^\dagger(0), Y_i^\dagger(1))$ given W_i and the demonstration pool, and suppose that there is a fixed matrix-valued function $\Sigma^\dagger$ such that $N^{-1}\sum_{i=1}^N \|\Sigma_i^\dagger - \Sigma^\dagger(W_i)\| = o_p(1)$. Together with Assumption 5.4(ii), this condition states that the first two conditional moments of the stored aggregate stabilize as the pool grows. A fixed demonstration pool, such as the superpopulation used in Section 6, satisfies it because the same rule then generates the stored aggregate for every subject.

Under these conditions, the population quantities $\mu_t^\dagger$, ω_t, $\tilde{v}_t$, κ, and $\sigma^2(\boldsymbol{e})$ are computed for an auxiliary signal with conditional mean $f_{\mathrm{FS},t}(W)$ and conditional covariance $\Sigma^\dagger(W)$ given W, and Assumptions 5.1 and 5.3 refer to these quantities. The proof above applies without the reduction in Step 1, with $Y_i^\dagger(t)$ in place of $f_{\mathrm{FS},t}(W_i)$ and with $(\mathcal{G}_{i-1}, W_i, Y_i^\dagger(0), Y_i^\dagger(1))$ in place of $(\mathcal{G}_{i-1}, W_i)$ in Step 2. Assumption 5.4(i) again gives $\Pr\{T_i = 1 \mid \mathcal{G}_{i-1}, W_i, Y_i^\dagger(0), Y_i^\dagger(1), Y_i(0), Y_i(1)\} = e_i$, so ψ_i is a martingale difference, and the expansion in Step 1 holds because $Y_i^\dagger(t)$ is bounded and fixed given $(\mathcal{G}_{i-1}, W_i, Y_i^\dagger(0), Y_i^\dagger(1))$. By the first condition, $\mathbb{E}[\psi_i^2 \mid \mathcal{G}_{i-1}]$ depends on the stored aggregate only through its conditional mean $\bar{f}_{i,t}$ and covariance $\Sigma_i^\dagger$ given W and the pool. These are $\mathcal{G}_{i-1}$-measurable functions of W because $\mathcal{G}_{i-1}$ contains the pool and the fold labels. Since W_i is independent of $\mathcal{G}_{i-1}$ and the outputs are bounded, subject averages of these functions evaluated at W_i and of their conditional expectations given $\mathcal{G}_{i-1}$ differ by a martingale average that tends to zero. Assumption 5.4(ii) and the second condition therefore give the limit $\sigma^2(\boldsymbol{e}^*)$ in Step 3. Theorem 5.6 follows from Assumption 5.3 as in Appendix F, and Theorem 5.8 holds for any auxiliary signal.

In this fixed-B case, $\widehat{\mathbb{V}}$ estimates the asymptotic variance of the estimator that uses the stored aggregate, and no additional LLM queries are needed. Relative to $f_{\mathrm{FS},t}$, the stored aggregate has an additional conditional variance, which attenuates ω_t and increases $\tilde{v}_t$, so a finite B increases the allocation-relevant criterion but does not affect validity. Theorem 5.9, which compares the stored aggregate with $f_{\mathrm{FS},t}$, still requires $B \to \infty$.

H Proof of Theorem 5.8

Proof. For any feasible policy $\boldsymbol{e}$,

$$\mathcal{V}_{\mathrm{AIPW}}(\boldsymbol{e}) - \mathcal{V}(\boldsymbol{e}) = \sum_{s=1}^{S} r_s \mathbb{E}\left[\frac{v_1(Z)\rho_1^2(Z)}{e_s(Z)} + \frac{v_0(Z)\rho_0^2(Z)}{1 - e_s(Z)}\right] \ge 0.$$

Each term is nonnegative, so $\mathcal{V}(\boldsymbol{e}) \le \mathcal{V}_{\mathrm{AIPW}}(\boldsymbol{e})$. Appendix C shows that $\boldsymbol{e}_{\mathrm{I}}^*$ minimizes $\mathcal{V}(\boldsymbol{e})$ over feasible policies, and the classical Neyman policy, truncated to $[\delta, 1-\delta]$, is feasible. Hence

$$\mathcal{V}(\boldsymbol{e}_{\mathrm{I}}^*) \le \mathcal{V}(\boldsymbol{e}_{\mathrm{Ney}}^*).$$

The argument uses only $\tilde{v}_t(z) = v_t(z)\{1 - \rho_t^2(z)\}$ and Appendix C, so both comparisons hold for any auxiliary signal, including the zero-shot signal that defines Layers 1 and 2 in Eq. (13). □

I Proof of Theorem 5.9

Proof. Fix an arm t, and let $D_{i,t} = 1 - \mathbb{1}\{T_i = t\}/e_{i,t}$ as in Appendix G, with $\mathcal{G}_i$ as defined there. Given $(\mathcal{G}_{i-1}, W_i, Y_i^\dagger(0), Y_i^\dagger(1))$, the remainder $R_{i,t,B} = Y_i^\dagger(t) - f_{\mathrm{FS},t}(W_i)$ of Lemma 5.5 is fixed,

and Assumption 5.4(i) gives $\Pr(T_i = 1 \mid \mathcal{G}_{i-1}, W_i, Y_i^\dagger(0), Y_i^\dagger(1)) = e_i$, so $D_{i,t}$ has conditional mean zero. As in Appendix G, a superscript $*$ marks a quantity computed with $f_{\mathrm{FS},t}(W_i)$ in place of $Y_i^\dagger(t)$.

Step 1. Reduction to stratum–fold sums. The two estimators share $\widehat{\mu}_t$, the assignments, the propensities, and the folds. As in Step 1 of Section 3, the cross-fitted $\widehat{\mu}_t(z)$ and $\widehat{\mu}_t^\dagger(z)$ are averages and $\widehat{\omega}_t(z)$ is a slope, all computed from the arm-t subjects in stratum z outside fold k, so they are constant within each stratum–fold cell. For subjects with $k_i = k$ and $Z_i = z$, write $\widehat{\omega}_t(z)$ and $\widehat{\mu}_t^\dagger(z)$ for the estimates computed from $Y^\dagger(t)$ outside fold k, suppressing their dependence on k, and $\widehat{\omega}_t^*(z)$ and $\widehat{\mu}_t^{\dagger *}(z)$ for their counterparts computed from $f_{\mathrm{FS},t}$. The arm-t contribution of subject i to $\widehat{\varphi}_t - \widehat{\varphi}_t^*$ is

$$\begin{aligned} &D_{i,t}\big[\widehat{\omega}_t(z)\{Y_i^\dagger(t) - \widehat{\mu}_t^\dagger(z)\} - \widehat{\omega}_t^*(z)\{f_{\mathrm{FS},t}(W_i) - \widehat{\mu}_t^{\dagger *}(z)\}\big] \\ &= D_{i,t}\widehat{\omega}_t(z)R_{i,t,B} - D_{i,t}\widehat{\omega}_t(z)\{\widehat{\mu}_t^\dagger(z) - \widehat{\mu}_t^{\dagger *}(z)\} \\ &\quad + D_{i,t}\{\widehat{\omega}_t(z) - \widehat{\omega}_t^*(z)\}\{f_{\mathrm{FS},t}(W_i) - \widehat{\mu}_t^{\dagger *}(z)\}. \end{aligned}$$

Summing over the subjects of the cell, write $\sum_{kz}$ for the sum over i with $k_i = k$ and $Z_i = z$. The arm-t contribution of the cell to $\widehat{\tau} - \widehat{\tau}^*$ is

$$\begin{aligned} &\widehat{\omega}_t(z)\,\frac{1}{N}\sum\nolimits_{kz} D_{i,t}R_{i,t,B} - \widehat{\omega}_t(z)\{\widehat{\mu}_t^\dagger(z) - \widehat{\mu}_t^{\dagger *}(z)\}\,\frac{1}{N}\sum\nolimits_{kz} D_{i,t} \\ &+ \{\widehat{\omega}_t(z) - \widehat{\omega}_t^*(z)\}\Big[\frac{1}{N}\sum\nolimits_{kz} D_{i,t}f_{\mathrm{FS},t}(W_i) - \widehat{\mu}_t^{\dagger *}(z)\,\frac{1}{N}\sum\nolimits_{kz} D_{i,t}\Big]. \end{aligned}$$

Step 2. Martingale bounds. The three cell sums in Step 1 are sums of martingale differences with respect to $\{\mathcal{G}_i\}$. Each summand is $\mathcal{G}_i$-measurable, and its factors other than $D_{i,t}$, including the indicator that subject i belongs to the cell, are fixed given $(\mathcal{G}_{i-1}, W_i, Y_i^\dagger(0), Y_i^\dagger(1))$, under which $D_{i,t}$ has conditional mean zero. For the first sum, the predictable quadratic variation of $N^{-1/2}\sum_{kz} D_{i,t}R_{i,t,B}$ is at most

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\Big[\frac{1 - e_{i,t}}{e_{i,t}} R_{i,t,B}^2 \;\Big|\; \mathcal{G}_{i-1}\Big] \le \frac{1}{\delta}\cdot\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\big[R_{i,t,B}^2 \mid \mathcal{G}_{i-1}\big].$$

Because $R_{i,t,B}^2 \le M_\dagger^2$ by Step 1 of Appendix G, the martingale weak law shows that the right-hand average differs from $N^{-1}\sum_i R_{i,t,B}^2$ by $o_p(1)$, and Lemma 5.5 makes the latter $o_p(1)$. Lenglart's inequality then gives $N^{-1}\sum_{kz} D_{i,t}R_{i,t,B} = o_p(N^{-1/2})$. The same argument with $R_{i,t,B}$ replaced by 1 and by $f_{\mathrm{FS},t}(W_i) \in [0, M_\dagger]$ shows that the other two cell sums are $O_p(N^{-1/2})$.

Step 3. Calibration quantities. The centering estimates $\widehat{\mu}_t^\dagger(z)$ and $\widehat{\mu}_t^{\dagger *}(z)$ differ by the average of $R_{i,t,B}$ over the training subjects of the cell. By Cauchy–Schwarz, this difference is at most the square root of $N^{-1}\sum_i R_{i,t,B}^2$ divided by the fraction of subjects used for training in the cell. That fraction is bounded away from zero in probability because $p(z) > 0$ and $e_{i,t} \ge \delta$, so the difference is $o_p(1)$. The slopes are ratios of cell-level covariances and variances. By Cauchy–Schwarz and bounded moments, their numerators and denominators differ by $O_p(\{N^{-1}\sum_i R_{i,t,B}^2\}^{1/2}) = o_p(1)$. Assumption 5.3 gives $\widehat{\omega}_t(z) = O_p(1)$. If $\mathbb{V}\mathrm{ar}\{f_{\mathrm{FS},t}(W) \mid Z = z\} > 0$, the denominators are bounded away from zero with probability tending to one, so $\widehat{\omega}_t(z) - \widehat{\omega}_t^*(z) = o_p(1)$. If $\mathbb{V}\mathrm{ar}\{f_{\mathrm{FS},t}(W) \mid Z = z\} = 0$, then $f_{\mathrm{FS},t}(W_i) = \widehat{\mu}_t^{\dagger *}(z)$ for every subject in stratum z, so the third term in Step 1 vanishes. The same conclusions hold with the denominator floor and clipping of Appendix A.3, which are Lipschitz in these numerators and denominators.

Step 4. Conclusion. By Steps 2 and 3, each of the finitely many terms in Step 1 is $o_p(N^{-1/2})$. The first is $O_p(1)\,o_p(N^{-1/2})$, the second is $O_p(1)\,o_p(1)\,O_p(N^{-1/2})$, and the third is $o_p(1)\,O_p(N^{-1/2})$ or zero. Summing over arms gives $\widehat{\tau} - \widehat{\tau}^* = o_p(N^{-1/2})$. The argument uses only $N^{-1}\sum_i R_{i,t,B}^2 = o_p(1)$, so no rate condition relating B to N is needed. Because $\widehat{\tau}^*$ admits the influence-function expansion in Step 1 of Appendix G, $\widehat{\tau}$ has the same first-order influence function and the same asymptotic distribution. □

J Additional simulation results

This appendix reports additional simulation results supporting Section 6. We include the construction of the signal-sensitivity scenarios and auxiliary baselines, the Objective-II settings behind Figure 2, the Objective-I efficiency decomposition, welfare failure-reduction results, modeled finite-B sensitivity, robustness diagnostics for the optional validation-gated safeguard, arm-by-stratum calibration diagnostics, an independently generated superpopulation, a predictor-independent signal experiment, and a cross-model comparison.

Additional DGP details. The BRIGHTEN-calibrated synthetic superpopulation has size $20{,}000$. Arm-specific potential outcomes are generated from conditional mean and variance models, then rounded and clipped to $[0, 27]$. The control-arm conditional mean adds standardized zero-shot and few-shot prediction components, each residualized on the BRIGHTEN-calibrated mean, and the treatment-arm mean has no such component. The LLM auxiliary signal therefore has nontrivial but imperfect predictive content in the control arm, and the signal is arm-asymmetric.

Compute resources. LLM auxiliary predictions were generated locally using Qwen3-8B on a GPU server with two NVIDIA RTX 6000 Ada Generation GPUs. The predictions were cached and reused across all Monte Carlo experiments. The main zero-shot and few-shot experiments used 300 replicates with trial size $N = 2{,}000$. With cached predictions, each of these two runs took about 15 hours on the same server. The cross-model comparison in Appendix J.9 queried `gpt-5-mini-2025-08-07` through the OpenAI Responses API on synthetic structured records only.

Signal-sensitivity scenarios. The scenarios of Figure 1 perturb the zero-shot LLM auxiliary predictions to vary signal strength, direction, and arm-specific reliability. Strong valid uses the original arm-specific predictions. Moderate and weak valid mix the original signal with a permuted signal using weights 0.7 and 0.3, and pure noise uses only permuted predictions. The spurious setting reorders the auxiliary predictions by a fixed permutation given by the age ranking, so each subject receives the predictions of another subject, which are not aligned with its arm-specific outcomes. The sign-reversed setting uses $27 - Y^{\dagger}(t)$ with added Gaussian noise of standard deviation 0.5, and the asymmetric setting keeps the original arm-0 signal but uses a permuted arm-1 signal. All perturbed predictions are truncated to the valid PHQ-9 range. Each scenario uses 100 replicates. The reported gains fix the allocation at the LLM-enhanced Neyman rule computed from the perturbed superpopulation, and the calibration and nuisance quantities are estimated in each replicate. The labels in Figure 1 give the population-weighted $\rho_0^2(z)$.

Auxiliary baselines. In Table 1, the Demo-only baseline uses only structured baseline covariates and excludes free-text features and LLM predictions. The TF-IDF+RF baseline adds 300 TF-IDF features of the combined free text to the structured covariates, whereas the TF-IDF/SVD baselines add a 10-dimensional truncated SVD of $1{,}000$ TF-IDF features. The oracle signal is the true conditional mean of $Y(t)$ given the baseline record plus independent $\mathcal{N}(0, 0.5^2)$ noise, truncated to the valid PHQ-9 range.

Objective-II settings. The failure set is $\mathcal{F} = \{Y \geq 15\}$. The Ethical design assigns treatment in stratum z with probability

$$e_{\mathrm{E}}(z) = \frac{\sqrt{1 - \mu_1^{\mathrm{F}}(z)}}{\sqrt{1 - \mu_0^{\mathrm{F}}(z)} + \sqrt{1 - \mu_1^{\mathrm{F}}(z)}},$$

truncated to $[\delta, 1 - \delta]$. The Objective-II allocations minimize the failure rate subject to a precision bound $C(z)$ on $\tilde{v}_1(z)/e(z) + \tilde{v}_0(z)/\{1 - e(z)\}$ in each stratum z, with $\tilde{v}_t$ replaced by v_t when no auxiliary signal is used. In Figure 2, $C(z) = v_1(z)/e_{\mathrm{E}}(z) + v_0(z)/\{1 - e_{\mathrm{E}}(z)\}$, the criterion of standard AIPW under the Ethical design, for every auxiliary signal. Layer 1 is the percent reduction of this criterion at e_{E} from zero-shot calibration. Layers 2 and 3 are the failure-rate reductions, as percentages of the Ethical failure rate, from replacing v_t by the zero-shot $\tilde{v}_t$ and then the zero-shot $\tilde{v}_t$ by the few-shot $\tilde{v}_t$. In Appendix J.2, $C(z)$ is 1.1 times the criterion of the same signal at its LLM-enhanced Neyman rule.

J.1 Objective-I efficiency decomposition

Figure 3 reports the Objective-I allocation-relevant efficiency decomposition in four strata. The dominant components are the LLM-enhanced efficiency gain of Layer 1 and the few-shot-enhanced gain of Layer 3. The design-enhanced gain of Layer 2 is positive but modest, which is expected because the LLM-enhanced Neyman allocation remains close to the classical Neyman allocation in this BRIGHTEN-calibrated setting. With population weights over all 18 strata, the three layers average 8.29%, 0.11%, and 6.69% of the standard-AIPW criterion under classical Neyman allocation. On the full variance $\sigma^2(\boldsymbol{e})$, the few-shot layer is much smaller because the auxiliary signal also changes C_{inv}. It equals the difference between the LLM-FS and LLM-ZS totals in Table 1, 0.53 percentage points.

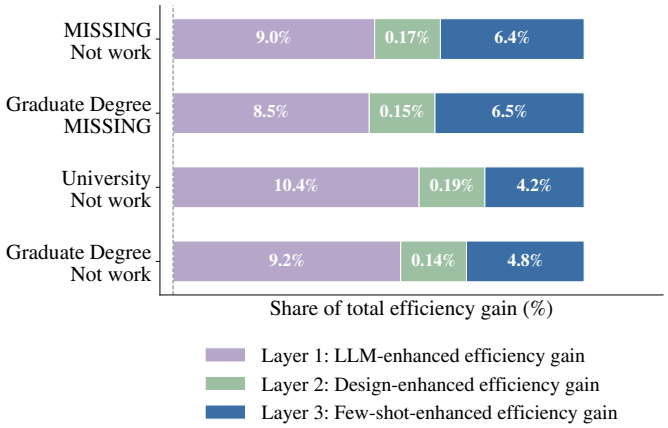

Figure 3: Objective-I efficiency decomposition of Eq. (13) on the allocation-relevant criterion $\mathcal{V}(\boldsymbol{e})$ in four strata. Labels give the layer gains as percentages of the stratum's standard-AIPW criterion under classical Neyman allocation, and segment widths show each layer's share of the stratum total, with a minimum width for readability. Table 1 reports the full-variance counterparts. Layer 1 is the LLM-enhanced efficiency gain from calibrated residual-variance reduction. Layer 2 is the design-enhanced efficiency gain from replacing the classical Neyman allocation with the LLM-enhanced Neyman allocation. Layer 3 is the additional gain from the few-shot construction. The design component is positive but small because the LLM-enhanced Neyman rule is close to the classical Neyman rule in this setting.

J.2 Objective-II failure reduction

Figure 4 shows that few-shot prompting generally yields larger reductions in the adverse PHQ-9 event rate under Objective II than zero-shot prompting. In each replicate, the stratum failure probabilities are estimated from the simulated trial by inverse-probability weighting and combined with the Objective-II and Ethical allocations computed from the superpopulation. Over all strata, the estimated reduction, as a percentage of the estimated Ethical failure rate, is 1.83% with zero-shot and 2.03% with few-shot prompting, each with Monte Carlo standard error 0.23 percentage points over 300 replicates. These results complement Figure 2, which uses a precision bound common to all signals and decomposes the mechanism into efficiency-side constraint relaxation and welfare-side failure-rate reduction.

J.3 Modeled sensitivity to the number of prompt resamples

Figure 5 reports a modeled sensitivity analysis for the number of prompt resamples B. It does not re-query the LLM at each value of B. Independent Gaussian noise with standard deviation $\sigma_{\text{p}}/\sqrt{B}$, with $\sigma_{\text{p}} = 1.5$, is added to the cached zero-shot predictions, which mimics averaging B prompt-level outputs with prompt-level standard deviation σ_{p}, and the Monte Carlo design is rerun with 100 replicates per value of B. As in Figure 1, the gains fix the allocation at the LLM-enhanced Neyman rule computed from the superpopulation with the noise-added signal. The $B = \infty$ reference uses the cached predictions without added noise. With few resamples, the added noise weakens

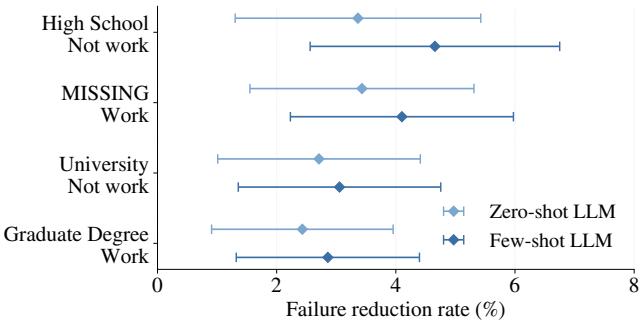

Figure 4: Failure-rate reduction under Objective II in four strata. Points report Monte Carlo means of the estimated percent reduction in the stratum failure rate relative to the Ethical design, with 95% intervals equal to ± 1.96 Monte Carlo standard errors across replicates. Few-shot prompting tends to yield larger reductions than zero-shot prompting, consistent with its stronger control-arm signal.

the calibrated signal and can lead to negative gains. As B increases, the squared within-stratum correlation increases, prediction variability decreases, and the efficiency gain approaches the $B = \infty$ reference. The analysis illustrates the role of the resampling-stability condition in Theorem 5.9.

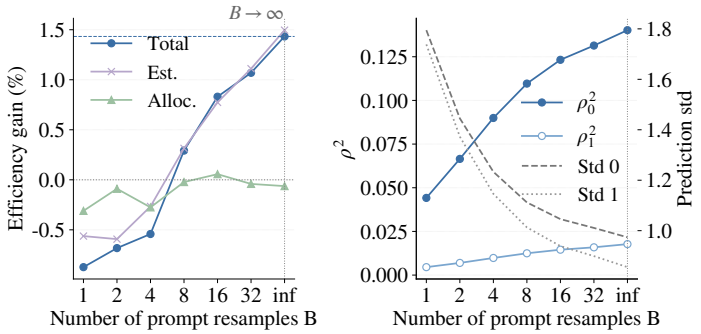

Figure 5: Modeled sensitivity to the number of prompt resamples B. Gaussian noise with variance σ_{p}^2/B is added to cached zero-shot predictions, and the LLM is not re-queried. The left panel reports the total efficiency gain and its estimation and allocation components, and the right panel reports the auxiliary signal quality and prediction variability. The $B = \infty$ reference uses the cached predictions without added noise.

J.4 Stress test for optional validation-gated safeguard

Table 2 evaluates the optional validation-gated safeguard described in Remark 3.1 under four signal regimes. Genuine uses the original zero-shot predictions, pure noise replaces both arm predictions by independent standard normal draws, subgroup miscalibration adds 5 to both arm predictions in five randomly chosen strata, and linear offset adds to both arm predictions of each subject a common offset that increases linearly from 0 to 2 across the superpopulation records. The safeguard is not part of the main calibrated procedure used in the primary experiments. It is a finite-sample robustness option for settings where auxiliary signals may be spurious or miscalibrated. The validation gate follows Appendix A.1 with 2 folds within each arm–stratum cell and $c_\alpha = 1.645$, and it screens out cells with fewer than 8 observations. Shrinkage multiplies $\widehat{\omega}_{t,s}(z)$ by $\max(0, 1 - 2/|r|)$, where r is the ratio of $\widehat{\omega}_{t,s}(z)$ to its standard error. Fallback sets $\widehat{\omega}_{t,s}(z) = 0$ in cells with fewer than 15 observations. Cells whose coefficient is set to zero also use no auxiliary signal in the estimator. The validation gate reduces the harm from unreliable signals, but it does not fully eliminate subgroup miscalibration risk. This is expected because the gate tests out-of-fold residual-variance reduction, which is more effective against pure noise than against systematic subgroup bias. Under the genuine signal, the safeguards change the total gain by at most 0.36 percentage points. Total gains remain close to zero in this configuration because the estimation gain of about 1.5% is offset by the stage-2 allocation estimated from stage-1 data in 18 strata. Absolute bias and coverage are computed under

the oracle LLM-enhanced allocation without a safeguard, so they are reported once per scenario. Coverage is between 96% and 97% in every scenario, including subgroup miscalibration.

Table 2: Stress test of the optional validation-gated safeguard under four signal regimes, with $n_1 = n_2 = 1000$ and 100 replicates. Total gains use the adaptive design with each safeguard. Absolute bias and coverage are computed under the oracle LLM-enhanced allocation without a safeguard, and $|\bar{\omega}|_1$ is the mean absolute stage-1 projection slope. Under subgroup miscalibration, the gain without a safeguard is strongly negative, and the validation gate moves it partly back toward zero.

Scenario	Safeguard	Total gain (%)	$\lvert\text{ATE bias}\rvert$	Coverage	$\lvert\bar{\omega}\rvert_1$
Genuine	None	-0.14	0.059	97.0%	1.86
	Shrinkage	-0.01			0.32
	Val. gate	-0.49			0.51
	Fallback	+0.01			1.08
Pure noise	None	-2.82	0.055	96.0%	1.45
	Shrinkage	-1.78			0.11
	Val. gate	-1.73			0.16
	Fallback	-2.35			0.49
Subgroup miscalibration	None	-20.23	0.058	96.0%	1.86
	Shrinkage	-16.08			0.32
	Val. gate	-14.71			0.51
	Fallback	-20.02			1.08
Linear offset	None	-2.60	0.040	96.0%	1.89
	Shrinkage	-1.54			0.27
	Val. gate	-1.60			0.39
	Fallback	-1.95			0.81

J.5 Screening-in diagnostic

Figure 6 reports how often the validation gate screens in an arm–stratum cell when it is applied to both potential outcomes of $2{,}000$ sampled subjects under the signal regimes of Appendix J.4. Cells with fewer than 10 sampled subjects are not screened in. Genuine signals are screened in more often than pure noise, while pure-noise signals are rejected in most arm–stratum cells. Subgroup miscalibration can still pass the screening rule because the signal may remain correlated with outcomes even when it is biased. This diagnostic explains why validation gating is useful as a finite-sample guardrail but should not be viewed as a complete solution to all forms of auxiliary-signal misspecification.

J.6 Arm-by-stratum calibration diagnostics

Table 3 reports the population calibration primitives in each of the 18 education $\times$ working-status strata for the zero-shot and few-shot signals. It gives the cell size n_z, the squared within-stratum correlation $\rho_t^2(z)$, the projection slope $\omega_t(z)$, and the calibrated residual standard deviation $\sqrt{\tilde{v}_t(z)}$, which is the root-mean-squared error of the calibrated prediction $m_t(Z, Y^\dagger(t))$ for $Y(t)$ within the stratum. The population-weighted $\rho_t^2(z)$ values are 0.140 and 0.018 for the zero-shot signal and 0.274 and 0.0007 for the few-shot signal. Averaged over all strata with population weights, $\sqrt{\tilde{v}_1(Z)/\tilde{v}_0(Z)}$ is 0.947 and 1.039, the LLM-enhanced Neyman allocation $e_{\mathrm{I}}^*(Z)$ averages 0.486 and 0.509, and the mean absolute shift from $e_{\mathrm{Ney}}^*(Z)$ is 0.017 and 0.040.

J.7 Independently generated structured clinical superpopulation

To check inference under an outcome construction that does not involve the evaluated predictions, we use a second synthetic superpopulation of $20{,}000$ subjects whose potential outcomes are generated before the auxiliary predictions and without using any evaluated prediction. Arm-specific random-forest outcome models are fit on the BRIGHTEN trial using only the structured covariates gender, education, working status, marital status, race, age, baseline PHQ-9, and screening score, and

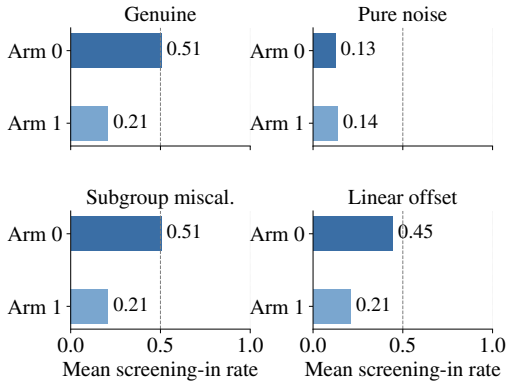

Figure 6: Screening-in rates for the validation-gated safeguard. The gate screens in genuine signals more often than pure noise, indicating that it filters many uninformative auxiliary signals. Subgroup-miscalibrated signals may still pass the gate because they can reduce the out-of-fold residual-variance estimate despite systematic subgroup errors.

Table 3: Population calibration primitives by arm and education × working-status stratum in the BRIGHTEN-calibrated superpopulation.

		LLM-ZS						LLM-FS					
Stratum	n_z	ρ_0^2	ρ_1^2	ω_0	ω_1	$\sqrt{v_0}$	$\sqrt{v_1}$	ρ_0^2	ρ_1^2	ω_0	ω_1	$\sqrt{v_0}$	$\sqrt{v_1}$
Community College / Work	2,578	0.126	0.017	2.68	1.09	6.49	6.27	0.293	0.000	1.62	-0.04	5.84	6.32
Community College / Not work	587	0.084	0.008	2.15	0.76	6.65	6.34	0.252	0.000	1.66	0.02	6.01	6.37
Community College / Missing	388	0.037	0.007	2.14	1.06	6.71	6.48	0.311	0.000	1.62	0.06	5.68	6.50
Elementary School / Work	203	0.109	0.001	2.75	0.31	6.56	5.97	0.304	0.008	1.66	0.27	5.80	5.95
Elementary School / Not work	54	0.326	0.003	2.51	-0.16	6.03	5.08	0.140	0.008	1.36	0.25	6.81	5.07
Elementary School / Missing	55	0.267	0.007	2.92	-0.52	5.62	6.50	0.097	0.000	0.94	-0.05	6.24	6.52
Graduate Degree / Work	1,967	0.162	0.022	2.75	0.94	6.56	6.05	0.280	0.000	1.67	-0.05	6.08	6.12
Graduate Degree / Not work	1,144	0.160	0.015	2.70	0.80	6.42	6.15	0.256	0.000	1.53	-0.04	6.04	6.20
Graduate Degree / Missing	278	0.154	0.007	3.06	0.63	6.69	6.32	0.272	0.004	1.56	-0.18	6.21	6.33
High School / Work	2,183	0.154	0.012	2.81	0.86	6.35	6.11	0.251	0.000	1.65	0.02	5.97	6.15
High School / Not work	807	0.082	0.007	2.29	0.72	6.81	6.03	0.266	0.001	1.81	0.11	6.09	6.05
High School / Missing	254	0.112	0.034	3.13	2.11	6.61	6.13	0.256	0.004	1.83	-0.20	6.05	6.22
University / Work	5,076	0.156	0.024	2.73	1.04	6.33	6.03	0.288	0.000	1.72	0.05	5.82	6.10
University / Not work	1,808	0.180	0.016	3.00	0.85	6.28	5.90	0.265	0.001	1.70	0.09	5.95	5.95
University / Missing	806	0.112	0.032	2.72	1.49	6.40	6.15	0.264	0.001	1.64	0.10	5.83	6.25
Missing / Work	1,151	0.093	0.020	2.59	1.28	6.46	5.97	0.260	0.001	1.64	-0.08	5.84	6.02
Missing / Not work	350	0.163	0.003	3.70	0.51	6.77	6.14	0.272	0.005	1.77	0.22	6.32	6.13
Missing / Missing	311	0.042	0.002	2.16	0.55	6.42	6.14	0.251	0.003	1.54	0.16	5.68	6.14

$Y(t) = \mu_t(X) + \sigma_t(X)\varepsilon_t$ with independent noise, clipped to $[0, 27]$ and rounded. After nonlinear adjustment for the structured covariates, the partial correlation between each μ_t and each evaluated prediction is at most 0.011 in absolute value. The trial uses $N = 2{,}000$ with $n_1 = 1{,}000$, gender strata, and $\delta = 0.1$, with the estimators of Section 6 and random-forest nuisance fits with 50 trees. The zero-shot signal uses a structured-input prompt, and the few-shot signal uses $B = 1$ prompt with three demonstrations drawn at random from the superpopulation. All methods use identical sampled subjects and seeds in each of 50 paired replicates. Table 4 reports the operating characteristics. Absolute bias is at most 0.065 and coverage lies between 0.94 and 0.96 for every auxiliary signal.

Table 4: Operating characteristics in the independently generated superpopulation, with $N = 2{,}000$, 50 paired replicates, and true ATE 0.296. Standard AIPW uses classical Neyman allocation, and the other rows use the proposed estimator under the LLM-enhanced allocation.

Auxiliary signal	Bias	Empirical SD	Mean SE	RMSE	Coverage
None (standard AIPW)	0.000	0.279	0.261	0.276	0.94
LLM-ZS	−0.032	0.239	0.260	0.239	0.94
LLM-FS ($B = 1$)	−0.026	0.263	0.273	0.261	0.96
Oracle signal (μ_t)	−0.065	0.230	0.261	0.237	0.96

J.8 Predictor-independent signal experiment

To isolate the allocation mechanism from any learned predictor, we prespecify (ρ_0^2, ρ_1^2) directly. The population has $20{,}000$ subjects in two balanced strata with $v_0(z) = v_1(z) = 1$ and stratum-level average treatment effect $\tau(z) = 1$ in both strata. Within each stratum, $Y(t) = \mu_t(z) + \sqrt{\rho_t^2}\, Y^\dagger(t) + \sqrt{1-\rho_t^2}\, \varepsilon_t$, where the auxiliary signals $Y^\dagger(0), Y^\dagger(1)$ and the errors $\varepsilon_0, \varepsilon_1$ are mutually orthogonal standardized vectors, so the cross-arm term $\kappa(z)$ in Eq. (18) is zero. Table 5 reports the population-oracle gains on the full variance $\sigma^2(\boldsymbol{e})$. Equal signal changes estimation but not allocation, while arm asymmetry in effective residual uncertainty moves the LLM-enhanced Neyman allocation away from $1/2$ and creates an allocation gain. Realized estimation gains over 20 Monte Carlo replicates are 5.05%, 30.50%, 20.78%, 19.91%, and 23.34% for the five scenarios, close to the population values.

Table 5: Population-oracle Objective-I gains, in percent, with prespecified signal strengths. Classical Neyman allocation is $1/2$ in every scenario.

Scenario	ρ_0^2	ρ_1^2	e_{I}^*	Est.	Alloc.	Total
S1	0.1	0.1	0.500	5.00	0.00	5.00
S2	0.6	0.6	0.500	30.00	0.00	30.00
S3	0.7	0.1	0.634	20.00	4.02	24.02
S4	0.1	0.7	0.366	20.00	4.02	24.02
S5	0.8	0.1	0.680	22.50	6.29	28.79

J.9 Cross-model comparison

To assess sensitivity to the auxiliary model, we compare zero-shot Qwen3-8B with zero-shot GPT-5-mini, version `gpt-5-mini-2025-08-07` with minimal reasoning effort and structured outputs, on the superpopulation of Appendix J.7, with the free-text fields removed and baseline PHQ-9 and screening score included in the prompt. The rows, potential outcomes, strata, seeds, nuisance estimation, allocation rule, estimator, and prompt template are identical, so only the auxiliary model changes. Both models return valid, in-range predictions for every subject and arm. Table 6 reports the results over 300 paired replicates. The paired difference in realized total gain, GPT-5-mini minus Qwen3-8B, is -0.49 percentage points with Monte Carlo standard error 0.047. Model choice changes the calibrated signal quality slightly, while empirical coverage stays close to the nominal 95% level under both models. Because the free text is removed, this comparison isolates model choice rather than text understanding.

Table 6: Matched cross-model comparison with $N = 2{,}000$ and 300 paired replicates. Population-weighted quantities are over both gender strata, and gains are on the full influence-function variance relative to standard AIPW under classical Neyman allocation.

Quantity	Qwen3-8B	GPT-5-mini
Weighted ρ_0^2 / ρ_1^2	0.247 / 0.072	0.241 / 0.071
Weighted $\sqrt{\tilde{v}_1(Z)/\tilde{v}_0(Z)}$	1.152	1.148
$\mathbb{E}\{e_{\mathrm{I}}^*(Z)\}$	0.535	0.534
$\mathbb{E}\lvert e_{\mathrm{I}}^*(Z) - e_{\mathrm{Ney}}^*(Z)\rvert$	0.026	0.025
Population-oracle estimation gain (%)	14.57	14.26
Population-oracle allocation gain (%)	0.23	0.21
Population-oracle total gain (%)	14.80	14.47
Realized total gain (%), $\pm$ Monte Carlo SE	9.30 ± 0.20	8.81 ± 0.20
RMSE, proposed estimator	0.241	0.239
Coverage of 95% intervals	0.960	0.963

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the scope of the method, including calibrated LLM auxiliary signals, predictable CARA allocation, and valid final inference. The theoretical claims are supported by Section 5, and the empirical claims are supported by Section 6 and Appendix J.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 7 discusses finite-sample stability, auxiliary-signal reliability, prompt variability, population shift, query cost, and deployment requirements for reproducibility and privacy. Appendix J reports diagnostics for unreliable or miscalibrated auxiliary signals.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Section 5 states the regularity, sequential randomization, nuisance estimation, and few-shot resampling stability assumptions, and each result states the assumptions it uses. The supporting derivations and proofs are provided in the technical appendices.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 6 describes the BRIGHTEN-calibrated simulation setup, allocation rules, auxiliary prediction methods, Monte Carlo replications, and evaluation metrics. Algorithm 1 states the procedure step by step, Appendix A.5 gives the prompt templates and decoding settings, and Appendix J gives additional DGP, compute, and diagnostic details.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The experiments use the public BRIGHTEN study as the calibration source and use synthetic data for the reported simulations. We provide supplementary materials with the code and scripts needed to reproduce the reported synthetic-data simulations.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 6 reports the trial size, number of Monte Carlo replications, stage sizes, strata, truncation, allocation rules, auxiliary signal sources, and cross-fitting setup. Appendices A.3, A.5, and J provide implementation details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Tables 1 and 5 and Figure 2 report deterministic population-level quantities, which carry no Monte Carlo error. For the Monte Carlo results, Tables 2, 4, and 6 report empirical coverage over replicates, Table 6 and Appendix J.2 report Monte Carlo standard errors for realized gains and for the overall welfare reductions, and Figure 4 shows 95% intervals based on Monte Carlo standard errors for the welfare results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix J reports the GPU hardware, caching strategy for Qwen3-8B auxiliary predictions, the API model used in the cross-model comparison, Monte Carlo sample sizes, and approximate runtime for the main experiments. The supplementary README reports the main software dependencies and few-shot prompt-resampling setting.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses synthetic trial simulations calibrated from an existing public study and does not introduce new human-subject data collection. The paper discusses risks from unreliable auxiliary signals and the need for validation diagnostics, truncation, and safeguards.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 7 states that the method may improve statistical efficiency and participant welfare when auxiliary signals are reliable. It also notes that unreliable or miscalibrated LLM signals can misdirect allocation and describes reproducibility and privacy safeguards for deployment.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release or deploy a high-risk decision system. For the proposed experimental design, Section 3 and Appendix J describe truncation, validation diagnostics, and optional safeguards for unreliable auxiliary signals.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the BRIGHTEN study used for calibration and identifies Qwen3-8B and GPT-5-mini (`gpt-5-mini-2025-08-07`) as the LLMs used to generate auxiliary predictions. Qwen3-8B is released under the Apache License 2.0, and GPT-5-mini was accessed through the OpenAI API under its terms of use. The supplementary materials list the main open-source software dependencies used in the experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The work creates synthetic superpopulations for the simulation experiments and the code that generates and analyzes them. The paper describes their construction in Section 6 and Appendix J.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not use crowdsourcing and does not collect new human-subject data. The reported experiments use synthetic data calibrated from an existing public study.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper reports simulation experiments using synthetic data and does not conduct a new human-subject study. The source study is cited for the BRIGHTEN calibration data.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Qwen3-8B was used to generate zero-shot and few-shot auxiliary outcome predictions from structured covariates and free-text baseline fields, and GPT-5-mini was used in the cross-model comparison. These LLM outputs are used only as auxiliary signals for calibration, allocation, and estimation, not as ground-truth outcomes. Appendix A.5 gives the prompt templates.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.