

Supplementary Materials for “Can Language Models Boost the Power of Randomized Experiments Without Statistical Bias?”

Xinrui Ruan¹ Xinwei Ma² Yingfei Wang³ Waverly Wei⁴ Jingshen Wang¹

¹Division of Biostatistics, University of California, Berkeley

²Department of Economics, University of California, San Diego

³Michael G. Foster School of Business, University of Washington

⁴Department of Data Sciences and Operations, University of Southern California

October 7, 2025

Revised: August 23, 2026

Contents

1	Proof of Results in the Main Paper	2
1.1	Technical Lemmas	2
1.2	Proof of Theorem 1	7
1.3	Proof of Theorem 2	7
1.4	Proof of Corollary 1	9
1.5	Proof of Theorem 3	11
1.6	Proof of Theorem 4	12
2	Motivating Real-world RCTs	16
3	Additional Simulation Study Results	18

1 Proof of Results in the Main Paper

We provide proofs of the results established in the main paper. Throughout the Supplementary Materials, limits are taken with respect to the sample size tending to infinity (i.e., $n \rightarrow \infty$). Convergence in probability and in distribution are denoted by $\xrightarrow{p}$ and $\xrightarrow{d}$, respectively. For two possibly random sequences a_n and b_n with $b_n \geq 0$, $a_n = O(b_n)$ or $a_n \lesssim b_n$ indicates that $|a_n|/b_n$ is bounded. Analogously, we employ $a_n = O_p(b_n)$ or $a_n \lesssim_p b_n$ if $|a_n|/b_n$ is bounded in probability. Asymptotic equivalence between two nonrandom sequences is $a_n \asymp b_n$. Finally, $a_n = o_p(1)$ if $a_n \xrightarrow{p} 0$.

1.1 Technical Lemmas

We first establish the asymptotic properties for the oracle estimator. Recall that $Y_i^\dagger(t)$ denotes the expected few-shot prediction obtained by averaging over the randomness of the demonstrative examples, and its conditional mean is

$$\mu_t^\dagger(x) = \mathbb{E}[Y_i^\dagger(t)|X_i = x].$$

Also recall that

$$\omega_t(x) = \frac{\text{Cov}[Y_i(t), Y_i^\dagger(t)|X_i = x]}{\mathbb{V}[Y_i^\dagger(t)|X_i = x]}.$$

The oracle CALM estimator for the parameter μ_t is defined as

$$\bar{\mu}_{t,\text{CALM}} = \frac{1}{n} \sum_{i=1}^n \varphi_t(i),$$

where

$$\varphi_t(i) = \frac{\mathbb{1}(T_i = t)}{e_t(X_i)} \left(Y_i - \mu_t(X_i) \right) + \mu_t(X_i) + \left(1 - \frac{\mathbb{1}(T_i = t)}{e_t(X_i)} \right) \left(Y_i^\dagger(t) - \mu_t^\dagger(X_i) \right) \omega_t(X_i).$$

To streamline the presentation, let

$$\nu_{i,t} = 1 - \frac{\mathbb{1}(T_i = t)}{e_t(X_i)},$$

which satisfies

$$\mathbb{E}[\nu_{i,t}|X_i] = \mathbb{E}[\nu_{i,t}|X_i, Z_i] = 0, \quad \text{and} \quad \mathbb{V}[\nu_{i,t}|X_i] = \mathbb{V}[\nu_{i,t}|X_i, Z_i] = \frac{1 - e_t(X_i)}{e_t(X_i)} = \lambda_t(X_i).$$

Lemma 1 (Asymptotic properties of the oracle CALM estimator). *Under Assumptions 1–3, the oracle estimator $\bar{\mu}_{t,\text{CALM}}$ is consistent and asymptotically normally distributed. Specifically, we have*

$\bar{\mu}_{t,\text{CALM}} \xrightarrow{p} \mu_t$, and $\sqrt{n} (\bar{\mu}_{t,\text{CALM}} - \mu_t) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_{t,\text{CALM}})$, where

$$\mathbf{V}_{t,\text{AIPW}} = \mathbb{E} \left[\frac{\mathbb{V}[Y_i(t)|X_i]}{e_t(X_i)} + (\mu_t(X_i) - \mu_t)^2 \right], \quad \mathbf{V}_{t,\text{CALM}} = \mathbf{V}_{t,\text{AIPW}} - \mathbb{E} \left[\frac{\mathbb{V}[Y_i(t)|X_i]}{e_t(X_i)} (1 - e_t(X_i)) \rho_t^2(X_i) \right],$$

and

$$\rho_t(x) = \text{Corr}[Y_i(t), Y_i^\dagger(t) | X_i = x].$$

Proof. We prove the result in two steps: (i) consistency and asymptotic normality, and (ii) asymptotic variance computation.

Consistency and asymptotic normality. For any integrable function $h(X_i, Z_i)$, we have

$$\begin{aligned} \mathbb{E} [\nu_{i,t} \cdot h(X_i, Z_i)] &= \mathbb{E} [\mathbb{E} [\nu_{i,t} \cdot h(X_i, Z_i) | X_i, Z_i]] \\ &= \mathbb{E} \left[h(X_i, Z_i) \left(1 - \frac{\mathbb{E}[\mathbf{1}(T_i = t) | X_i, Z_i]}{e_t(X_i)} \right) \right] = 0, \end{aligned}$$

where the last equality uses Assumption 1. Since $Y_i^\dagger(t)$ is a function of (X_i, Z_i) , the term

$$\mu_t(X_i) + \omega_t(X_i) \{Y_i^\dagger(t) - \mu_t^\dagger(X_i)\}$$

is also a function of (X_i, Z_i) . The above argument then implies

$$\mathbb{E}[\varphi_t(i)] = \mathbb{E} \left[\frac{\mathbf{1}(T_i = t) Y_i}{e_t(X_i)} \right] = \mathbb{E}[Y_i(t)] = \mu_t.$$

Thus, $\varphi_t(i)$ is unbiased for μ_t . By Assumption 2 and 3, the sequence $\{\varphi_t(i)\}_{i=1}^n$ is i.i.d. with finite and nonvanishing variance. Therefore, a standard application of the CLT implies

$$\bar{\mu}_{t,\text{CALM}} \xrightarrow{p} \mu_t, \quad \sqrt{n} (\bar{\mu}_{t,\text{CALM}} - \mu_t) \xrightarrow{d} \mathcal{N}(0, \mathbb{V}[\varphi_t(i)]).$$

Asymptotic variance computation. We now compute $\mathbb{V}[\varphi_t(i)]$. By definition,

$$\begin{aligned} \mathbb{V}[\varphi_t(i)] &= \mathbb{E} [\{\varphi_t(i) - \mu_t\}^2] \\ &= \mathbb{E} \left[\left\{ \mu_t(X_i) - \mu_t + \frac{\mathbf{1}(T_i = t)}{e_t(X_i)} \{Y_i - \mu_t(X_i)\} + \nu_{i,t} \omega_t(X_i) \{Y_i^\dagger(t) - \mu_t^\dagger(X_i)\} \right\}^2 \right]. \end{aligned}$$

Let

$$\Delta_t(i) := \frac{\mathbf{1}(T_i = t)}{e_t(X_i)} \{Y_i - \mu_t(X_i)\} + \nu_{i,t} \omega_t(X_i) \{Y_i^\dagger(t) - \mu_t^\dagger(X_i)\}.$$

Then

$$\mathbb{V}[\varphi_t(i)] = \mathbb{E} [\{\mu_t(X_i) - \mu_t\}^2] + \underbrace{\mathbb{E}[\Delta_t(i)^2]}_{\text{(I)}} + 2 \underbrace{\mathbb{E}[\{\mu_t(X_i) - \mu_t\} \Delta_t(i)]}_{\text{(II)}}.$$

Since $\mathbb{E}[\Delta_t(i)|X_i] = 0$, we have

$$(II) = \mathbb{E} [\{\mu_t(X_i) - \mu_t\} \mathbb{E}[\Delta_t(i)|X_i]] = 0.$$

A direct variance-covariance expansion gives

$$(I) = \mathbb{E} \left[\frac{1}{e_t(X_i)} \mathbb{V}[Y_i(t)|X_i] + \frac{1 - e_t(X_i)}{e_t(X_i)} \left\{ \omega_t(X_i)^2 \mathbb{V}[Y_i^\dagger(t)|X_i] - 2\omega_t(X_i) \text{Cov}[Y_i(t), Y_i^\dagger(t)|X_i] \right\} \right].$$

Therefore,

$$\mathbb{V}[\varphi_t(i)] = \mathbb{E} \left[\{\mu_t(X_i) - \mu_t\}^2 + \mathbb{V}[Y_i(t)|X_i] + \frac{1 - e_t(X_i)}{e_t(X_i)} \mathbb{V}[Y_i(t) - \omega_t(X_i) Y_i^\dagger(t)|X_i] \right].$$

By the definition of $\omega_t(x)$,

$$\mathbb{V}[Y_i(t) - \omega_t(X_i) Y_i^\dagger(t)|X_i] = \mathbb{V}[Y_i(t)|X_i] \{1 - \rho_t^2(X_i)\}.$$

This completes the proof. $\square$

For the next lemma, we write $Y_i^\dagger(t) = Y^\dagger(X_i, Z_i, t)$ to emphasize its dependence on the unit's pre-intervention covariates.

Lemma 2 (Convergence of resampling-based few-shot predictions). *Under Assumptions 2–3,*

$$\mathbb{E}_{X_i, Z_i} |Y_i^\dagger(t; \mathcal{I}) - Y_i^\dagger(t)|^2 \lesssim_p \frac{1}{B} + \frac{1}{n},$$

where the expectation is taken with respect to the marginal distribution of (X_i, Z_i) .

Proof. Throughout this proof, we assume that (X_i, Z_i) is statistically independent of the demonstrative examples. Recall that our resampling and aggregation step gives

$$Y_i^\dagger(t; \mathcal{I}) = \frac{1}{B} \sum_{b=1}^B Y^\dagger(X_i, Z_i, t, \mathcal{S}_b^*).$$

Given an unordered sample $\mathcal{I}$, we denote its size by n' . Note that $n' \asymp n$ by our sample splitting procedure. We will also define the symmetrized version: for an unordered sample $\mathcal{S}$ of size m ,

$$Y_{\text{sym}}^\dagger(X_i, Z_i, t, \mathcal{S}) = \frac{1}{m!} \sum_b Y^\dagger(X_i, Z_i, t, \mathcal{S}_b),$$

where the above takes summation over all ordered versions of $\mathcal{S}$: $\mathcal{S}_b$ for $b = 1, 2, \dots, m!$. Now consider the auxiliary quantities

$$\bar{Y}_i^\dagger(t; \mathcal{I}) = \frac{1}{\binom{n'}{m}} \sum_b Y_{\text{sym}}^\dagger(X_i, Z_i, t, \mathcal{S}_b),$$

where $\mathcal{S}_b$ in the above enumerates over all unordered subsamples of size m , and

$$\tilde{Y}_i^\dagger(t; \mathcal{I}) = \frac{1}{\lfloor n'/m \rfloor} \sum_b Y_{\text{sym}}^\dagger(X_i, Z_i, t, \mathcal{S}_b),$$

where the summation takes over some nonoverlapping subsets of m . (This can be done by arbitrarily ordering observations in $\mathcal{I}$ into $O_1, O_2, \dots, O_{n'}$, and let $\mathcal{S}_1 = \{O_1, \dots, O_m\}$, and $\mathcal{S}_2 = \{O_{m+1}, \dots, O_{2m}\}$, ...) The notation $\lfloor n'/m \rfloor$ represents the floor function, which is the largest integer below n'/m . Given the unordered sample $\mathcal{I}$, one immediate observation is that

$$\mathbb{E}[Y_i^\dagger(t; \mathcal{I}) | X_i, Z_i, \mathcal{I}] = \mathbb{E}[\tilde{Y}_i^\dagger(t; \mathcal{I}) | X_i, Z_i, \mathcal{I}] = \bar{Y}_i^\dagger(t; \mathcal{I}).$$

Given the above observation, we immediately have

$$\mathbb{V}[\bar{Y}_i^\dagger(t; \mathcal{I}) | X_i, Z_i] \leq \mathbb{V}[\tilde{Y}_i^\dagger(t; \mathcal{I}) | X_i, Z_i] = \frac{m}{n'} \mathbb{V}[Y_{\text{sym}}^\dagger(X_i, Z_i, t, \mathcal{S}) | X_i, Z_i] \lesssim \frac{1}{n},$$

where the final bound in the above is due to our Assumption 3.

Now consider $Y_i^\dagger(t; \mathcal{I})$. Conditioning on the unordered sample $\mathcal{I}$ and (X_i, Z_i) , its variance is

$$\mathbb{V}[Y_i^\dagger(t; \mathcal{I}) | X_i, Z_i, \mathcal{I}] = \frac{1}{B} \frac{1}{\binom{n'}{m}} \frac{1}{m!} \sum_b \left(Y^\dagger(X_i, Z_i, t, \mathcal{S}_b) - \bar{Y}_i^\dagger(t; \mathcal{I}) \right)^2,$$

and the summation above is taken over all ordered subsets $\mathcal{S}_b$ of size m . Again, by Assumption 3, we have

$$\mathbb{V}[Y_i^\dagger(t; \mathcal{I}) | X_i, Z_i, \mathcal{I}] \lesssim \frac{1}{B}.$$

Combining the previous results with triangle inequality,

$$|Y_i^\dagger(t; \mathcal{I}) - Y_i^\dagger(t)|^2 \leq 2|Y_i^\dagger(t; \mathcal{I}) - \bar{Y}_i^\dagger(t; \mathcal{I})|^2 + 2|\bar{Y}_i^\dagger(t; \mathcal{I}) - Y_i^\dagger(t)|^2 \lesssim_{\text{P}} \frac{1}{B} + \frac{1}{n},$$

Finally, dominated convergence implies

$$\mathbb{E}_{X_i, Z_i} |Y_i^\dagger(t; \mathcal{I}) - Y_i^\dagger(t)|^2 \lesssim_{\text{P}} \frac{1}{B} + \frac{1}{n},$$

where $\mathbb{E}_{X_i, Z_i}[\cdot]$ denotes marginal expectation with respect to X_i and Z_i . □

Lemma 3 (Asymptotic equivalence). *Under Assumptions 1–4 and $B \rightarrow \infty$,*

$$\sqrt{n} (\hat{\mu}_{t, \text{CALM}} - \bar{\mu}_{t, \text{CALM}}) \xrightarrow{\text{P}} 0.$$

Proof. With slight abuse of notation, we define

$$\hat{\mu}_{t,\text{CALM}}^{\mathcal{I}_3} = \frac{1}{|\mathcal{I}_3|} \sum_{i \in \mathcal{I}_3} \hat{\varphi}_t(i; \mathcal{I}_1, \mathcal{I}_2), \quad \bar{\mu}_{t,\text{CALM}}^{\mathcal{I}_3} = \frac{1}{|\mathcal{I}_3|} \sum_{i \in \mathcal{I}_3} \varphi_t(i),$$

and $|\mathcal{I}_3|$ is the sample size of the fold $\mathcal{I}_3$. Note that $|\mathcal{I}_3| \asymp n$. It will suffice to show that

$$\sqrt{n} \left(\hat{\mu}_{t,\text{CALM}}^{\mathcal{I}_3} - \bar{\mu}_{t,\text{CALM}}^{\mathcal{I}_3} \right) \xrightarrow{\mathbb{P}} 0,$$

and the desired conclusion will follow from the cross-fitting step. We recall the difference between these quantities is that the oracle estimator uses $Y_i^\dagger(t)$ and the true nuisance functions $\mu_t(\cdot)$, $\mu_t^\dagger(\cdot)$, and $\omega_t(\cdot)$, whereas the feasible estimator uses the aggregated few-shot prediction $Y_i^\dagger(t; \mathcal{I}_1)$ and the cross-fitted estimates $\hat{\mu}_t(\cdot; \mathcal{I}_2)$, $\hat{\mu}_t^\dagger(\cdot; \mathcal{I}_2)$, and $\hat{\omega}_t(\cdot; \mathcal{I}_2)$.

We first decompose the difference as

$$\sqrt{n} \left(\hat{\mu}_{t,\text{CALM}}^{\mathcal{I}_3} - \bar{\mu}_{t,\text{CALM}}^{\mathcal{I}_3} \right) = \frac{\sqrt{n}}{|\mathcal{I}_3|} \sum_{i \in \mathcal{I}_3} \nu_{i,t} \cdot \left(\hat{\mu}_t(X_i; \mathcal{I}_2) - \mu_t(X_i) \right) \quad (\text{I})$$

$$- \frac{\sqrt{n}}{|\mathcal{I}_3|} \sum_{i \in \mathcal{I}_3} \nu_{i,t} \cdot \hat{\omega}_t(X_i; \mathcal{I}_2) \left(\hat{\mu}_t^\dagger(X_i; \mathcal{I}_2) - \mu_t^\dagger(X_i) \right) \quad (\text{II})$$

$$- \frac{\sqrt{n}}{|\mathcal{I}_3|} \sum_{i \in \mathcal{I}_3} \nu_{i,t} \cdot \mu_t^\dagger(X_i) \left(\hat{\omega}_t(X_i; \mathcal{I}_2) - \omega_t(X_i) \right) \quad (\text{III})$$

$$+ \frac{\sqrt{n}}{|\mathcal{I}_3|} \sum_{i \in \mathcal{I}_3} \nu_{i,t} \cdot Y_i^\dagger(t) \left(\hat{\omega}_t(X_i; \mathcal{I}_2) - \omega_t(X_i) \right) \quad (\text{IV})$$

$$+ \frac{\sqrt{n}}{|\mathcal{I}_3|} \sum_{i \in \mathcal{I}_3} \nu_{i,t} \cdot \hat{\omega}_t(X_i; \mathcal{I}_2) \left(Y_i^\dagger(t; \mathcal{I}_1) - Y_i^\dagger(t) \right), \quad (\text{V})$$

where we recall that $\nu_{i,t} = 1 - \mathbb{1}(T_i = t)/e_t(X_i)$.

We now bound these terms. For the first term, conditional on $\mathcal{I}_1, \mathcal{I}_2$ and the observed covariates $\{X_i, Z_i : i \in \mathcal{I}_3\}$, its variance is

$$\begin{aligned} \mathbb{V}[(\text{I}) | \mathcal{I}_1, \mathcal{I}_2, \{X_i, Z_i\}_{i \in \mathcal{I}_3}] &= \frac{n}{|\mathcal{I}_3|^2} \sum_{i \in \mathcal{I}_3} \frac{1 - e_t(X_i)}{e_t(X_i)} \left(\hat{\mu}_t(X_i; \mathcal{I}_2) - \mu_t(X_i) \right)^2 \\ &\lesssim \frac{1}{|\mathcal{I}_3|} \sum_{i \in \mathcal{I}_3} \left(\hat{\mu}_t(X_i; \mathcal{I}_2) - \mu_t(X_i) \right)^2 \xrightarrow{\mathbb{P}} 0 \end{aligned}$$

by our Assumption 4. Similar arguments apply to (II)–(IV):

$$\begin{aligned} \mathbb{V}[(\text{II}) | \mathcal{I}_1, \mathcal{I}_2, \{X_i, Z_i\}_{i \in \mathcal{I}_3}] &= \frac{n}{|\mathcal{I}_3|^2} \sum_{i \in \mathcal{I}_3} \frac{1 - e_t(X_i)}{e_t(X_i)} \hat{\omega}_t(X_i; \mathcal{I}_2)^2 \left(\hat{\mu}_t^\dagger(X_i; \mathcal{I}_2) - \mu_t^\dagger(X_i) \right)^2 \\ &\lesssim \frac{1}{|\mathcal{I}_3|} \sum_{i \in \mathcal{I}_3} \left(\hat{\mu}_t^\dagger(X_i; \mathcal{I}_2) - \mu_t^\dagger(X_i) \right)^2 \xrightarrow{\mathbb{P}} 0, \end{aligned}$$

where we have used the assumption that the estimated calibration weight $\hat{\omega}_t(X_i; \mathcal{I}_2)$ is bounded. Next for (III),

$$\begin{aligned}\mathbb{V}[(\text{III})|\mathcal{I}_1, \mathcal{I}_2, \{X_i, Z_i\}_{i \in \mathcal{I}_3}] &= \frac{n}{|\mathcal{I}_3|^2} \sum_{i \in \mathcal{I}_3} \frac{1 - e_t(X_i)}{e_t(X_i)} \mu_t^\dagger(X_i)^2 (\hat{\omega}_t(X_i; \mathcal{I}_2) - \omega_t(X_i))^2 \\ &\lesssim \frac{1}{|\mathcal{I}_3|} \sum_{i \in \mathcal{I}_3} (\hat{\omega}_t(X_i; \mathcal{I}_2) - \omega_t(X_i))^2 \xrightarrow{\text{p}} 0.\end{aligned}$$

Finally for (IV),

$$\begin{aligned}\mathbb{V}[(\text{IV})|\mathcal{I}_1, \mathcal{I}_2, \{X_i, Z_i\}_{i \in \mathcal{I}_3}] &= \frac{n}{|\mathcal{I}_3|^2} \sum_{i \in \mathcal{I}_3} \frac{1 - e_t(X_i)}{e_t(X_i)} Y_i^\dagger(t)^2 (\hat{\omega}_t(X_i; \mathcal{I}_2) - \omega_t(X_i))^2 \\ &\lesssim \frac{1}{|\mathcal{I}_3|} \sum_{i \in \mathcal{I}_3} (\hat{\omega}_t(X_i; \mathcal{I}_2) - \omega_t(X_i))^2 \xrightarrow{\text{p}} 0.\end{aligned}$$

It remains to bound the few-shot aggregation error in term (V). By the same conditioning argument,

$$\begin{aligned}\mathbb{V}[(\text{V})|\mathcal{I}_1, \mathcal{I}_2, \{X_i, Z_i\}_{i \in \mathcal{I}_3}] &= \frac{n}{|\mathcal{I}_3|^2} \sum_{i \in \mathcal{I}_3} \frac{1 - e_t(X_i)}{e_t(X_i)} \hat{\omega}_t(X_i; \mathcal{I}_2)^2 (Y_i^\dagger(t; \mathcal{I}_1) - Y_i^\dagger(t))^2 \\ &\lesssim \frac{1}{|\mathcal{I}_3|} \sum_{i \in \mathcal{I}_3} (Y_i^\dagger(t; \mathcal{I}_1) - Y_i^\dagger(t))^2 \\ &\lesssim_{\text{p}} \mathbb{E}_{X_i, Z_i} (Y_i^\dagger(t; \mathcal{I}_1) - Y_i^\dagger(t))^2 \lesssim_{\text{p}} \frac{1}{B} + \frac{1}{n}.\end{aligned}$$

The last line is due to Markov inequality applied to nonnegative random variables and our Lemma 2. $\square$

1.2 Proof of Theorem 1

This theorem follows from Lemmas 1, 2, and 3.

1.3 Proof of Theorem 2

The proof consists of two parts. The first part establishes asymptotic equivalence, and the second computes the asymptotic variance.

Asymptotic equivalence. The main departure of this theorem from Theorem 1 is that it replaces the $\mathcal{I}_2/\mathcal{I}_3$ split with Assumption 5. We start from the decomposition:

$$\sqrt{n} \left(\hat{\mu}_{t, \text{CALM}}^{\mathcal{I}_2} - \bar{\mu}_{t, \text{CALM}}^{\mathcal{I}_2} \right) = \frac{\sqrt{n}}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \nu_{i,t} \cdot \left(\hat{\mu}_t(X_i; \mathcal{I}_2) - \tilde{\mu}_t(X_i) \right) \quad (\text{I})$$

$$+ \frac{\sqrt{n}}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \nu_{i,t} \cdot Y_i^\dagger(t; \mathcal{I}_1) \left(\hat{\omega}_t(X_i; \mathcal{I}_2) - \omega_t(X_i) \right) \quad (\text{II})$$

$$-\frac{\sqrt{n}}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \nu_{i,t} \cdot \left(\hat{\mu}_t^\dagger(X_i; \mathcal{I}_2) \hat{\omega}_t(X_i; \mathcal{I}_2) - \tilde{\mu}_t^\dagger(X_i) \omega_t(X_i) \right) \quad (\text{III})$$

$$+\frac{\sqrt{n}}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \nu_{i,t} \cdot \omega_t(X_i) \left(Y_i^\dagger(t; \mathcal{I}_1) - Y_i^\dagger(t) \right). \quad (\text{IV})$$

The first three terms in the decomposition vanish thanks to Assumption 5. The last term can be handled by conditional variance calculation:

$$\begin{aligned} \mathbb{V}[(\text{IV}) | \mathcal{I}_1, \{X_i, Z_i\}_{i \in \mathcal{I}_2}] &\lesssim \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \frac{1 - e_t(X_i)}{e_t(X_i)} \omega_t(X_i)^2 \left(Y_i^\dagger(t; \mathcal{I}_1) - Y_i^\dagger(t) \right)^2 \\ &\lesssim \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \left(Y_i^\dagger(t; \mathcal{I}_1) - Y_i^\dagger(t) \right)^2 \lesssim_{\text{p}} \mathbb{E}_{X_i, Z_i} \left(Y_i^\dagger(t; \mathcal{I}_1) - Y_i^\dagger(t) \right)^2 \lesssim_{\text{p}} \frac{1}{B} + \frac{1}{n}. \end{aligned}$$

The last line is due to Markov inequality applied to nonnegative random variables and our Lemma 2.

Asymptotic variance calculation. Recall the asymptotic representation:

$$\bar{\mu}_{t, \text{CALM}} = \frac{1}{n} \sum_{i=1}^n \underbrace{\frac{\mathbb{1}(T_i = t)}{e_t(X_i)} \left(Y_i - \tilde{\mu}_t(X_i) \right) + \tilde{\mu}_t(X_i)}_{= A_i} + \underbrace{\left(1 - \frac{\mathbb{1}(T_i = t)}{e_t(X_i)} \right) \left(Y_i^\dagger(t) - \tilde{\mu}_t^\dagger(X_i) \right) \omega_t(X_i)}_{= B_i}.$$

Clearly variance of A_i corresponds to the asymptotic variance of the AIPW approach under misspecification, which is

$$\begin{aligned} \mathbb{V}[A_i] &= \mathbb{V} \left[\frac{\mathbb{1}(T_i = t)}{e_t(X_i)} \left(Y_i - \tilde{\mu}_t(X_i) \right) + \tilde{\mu}_t(X_i) \right] \\ &= \mathbb{V} \left[\frac{\mathbb{1}(T_i = t)}{e_t(X_i)} \left(Y_i - \mu_t(X_i) \right) + \nu_{i,t} (\tilde{\mu}_t(X_i) - \mu_t(X_i)) + (\mu_t(X_i) - \mu_t) \right] \\ &= \mathbb{E} \left[\frac{\mathbb{V}[Y_i(t) | X_i]}{e_t(X_i)} + \lambda_t(X_i) (\tilde{\mu}_t(X_i) - \mu_t(X_i))^2 + (\mu_t(X_i) - \mu_t)^2 \right]. \end{aligned}$$

Notice that the three terms in the second line are conditionally uncorrelated.

To find the asymptotic variance of the CALM approach, we need the variance of B_i as well as the covariance between A_i and B_i . Specifically,

$$\mathbb{V}[B_i] = \mathbb{E} \left[\lambda_t(X_i) \left(Y_i^\dagger(t) - \tilde{\mu}_t^\dagger(X_i) \right)^2 \omega_t(X_i)^2 \right],$$

and the covariance is

$$\begin{aligned} 2\text{Cov}[A_i, B_i] &= 2\mathbb{E} \left[\frac{\mathbb{1}(T_i = t)}{e_t(X_i)} \left(Y_i - \tilde{\mu}_t(X_i) \right) \left(1 - \frac{\mathbb{1}(T_i = t)}{e_t(X_i)} \right) \left(Y_i^\dagger(t) - \tilde{\mu}_t^\dagger(X_i) \right) \omega_t(X_i) \right] \\ &\quad + 2\mathbb{E} \left[\tilde{\mu}_t(X_i) \left(1 - \frac{\mathbb{1}(T_i = t)}{e_t(X_i)} \right) \left(Y_i^\dagger(t) - \tilde{\mu}_t^\dagger(X_i) \right) \omega_t(X_i) \right] \end{aligned}$$

$$\begin{aligned}
&= 2\mathbb{E}\left[\frac{\mathbb{1}(T_i = t)}{e_t(X_i)}\left(Y_i - \tilde{\mu}_t(X_i)\right)\left(1 - \frac{\mathbb{1}(T_i = t)}{e_t(X_i)}\right)\left(Y_i^\dagger(t) - \tilde{\mu}_t^\dagger(X_i)\right)\omega_t(X_i)\right] \\
&= -2\mathbb{E}\left[\lambda_t(X_i)\omega_t(X_i)\mathbb{E}\left[(Y_i - \tilde{\mu}_t(X_i))(Y_i^\dagger(t) - \tilde{\mu}_t^\dagger(X_i))|X_i\right]\right],
\end{aligned}$$

which completes the calculation.

1.4 Proof of Corollary 1

To show this result, we will verify Assumption 5 for the special case of covariate discretization. We define the local constant basis expansion:

$$\mathbf{p}(X_i) = \left(\mathbb{1}(X_i \in \mathcal{X}_1), \mathbb{1}(X_i \in \mathcal{X}_2), \dots, \mathbb{1}(X_i \in \mathcal{X}_J)\right) \in \{0, 1\}^J.$$

Therefore, the nuisance functions and their estimates all have the representation:

$$\begin{aligned}
\tilde{\mu}_t(X_i) &= \mathbf{p}(X_i)^\top \boldsymbol{\beta}_{t,\mu}, & \hat{\mu}_t(X_i; \mathcal{I}_2) &= \mathbf{p}(X_i)^\top \hat{\boldsymbol{\beta}}_{t,\mu}(\mathcal{I}_2), \\
\tilde{\mu}_t^\dagger(X_i) &= \mathbf{p}(X_i)^\top \boldsymbol{\beta}_{t,\mu}^\dagger, & \hat{\mu}_t^\dagger(X_i; \mathcal{I}_2) &= \mathbf{p}(X_i)^\top \hat{\boldsymbol{\beta}}_{t,\mu}^\dagger(\mathcal{I}_2), \\
\omega_t(X_i) &= \mathbf{p}(X_i)^\top \boldsymbol{\beta}_{t,\omega}, & \hat{\omega}_t(X_i; \mathcal{I}_2) &= \mathbf{p}(X_i)^\top \hat{\boldsymbol{\beta}}_{t,\omega}(\mathcal{I}_2).
\end{aligned}$$

A direct consequence is that

$$\|\boldsymbol{\beta}_{t,\mu} - \hat{\boldsymbol{\beta}}_{t,\mu}(\mathcal{I}_2)\| \lesssim_p \frac{1}{\sqrt{n}}, \quad \|\boldsymbol{\beta}_{t,\mu}^\dagger - \hat{\boldsymbol{\beta}}_{t,\mu}^\dagger(\mathcal{I}_2)\| \lesssim_p \frac{1}{\sqrt{n}}, \quad \|\boldsymbol{\beta}_{t,\omega} - \hat{\boldsymbol{\beta}}_{t,\omega}(\mathcal{I}_2)\| \lesssim_p \frac{1}{\sqrt{n}},$$

where we note that $|\mathcal{I}_2| \asymp n$. To verify Assumption 5, we again check each term in the expansion. First,

$$\begin{aligned}
&\frac{1}{\sqrt{|\mathcal{I}_2|}} \sum_{i \in \mathcal{I}_2} \nu_{i,t} \cdot (\hat{\mu}_t(X_i; \mathcal{I}_2) - \tilde{\mu}_t(X_i)) = \frac{1}{\sqrt{|\mathcal{I}_2|}} \sum_{i \in \mathcal{I}_2} \nu_{i,t} \cdot \mathbf{p}(X_i)^\top (\hat{\boldsymbol{\beta}}_{t,\mu}(\mathcal{I}_2) - \boldsymbol{\beta}_{t,\mu}) \\
&\leq \left\| \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \nu_{i,t} \cdot \mathbf{p}(X_i)^\top \right\| \cdot \sqrt{|\mathcal{I}_2|} \|\hat{\boldsymbol{\beta}}_{t,\mu}(\mathcal{I}_2) - \boldsymbol{\beta}_{t,\mu}\| \\
&\lesssim_p \left\| \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \nu_{i,t} \cdot \mathbf{p}(X_i)^\top \right\| \xrightarrow{p} 0,
\end{aligned}$$

by the standard law of large numbers.

Similarly,

$$\begin{aligned}
&\frac{1}{\sqrt{|\mathcal{I}_2|}} \sum_{i \in \mathcal{I}_2} \nu_{i,t} \cdot Y_i^\dagger(t; \mathcal{I}_1) (\hat{\omega}_t(X_i; \mathcal{I}_2) - \omega_t(X_i)) \\
&\leq \left\| \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \nu_{i,t} \cdot Y_i^\dagger(t; \mathcal{I}_1) \mathbf{p}(X_i)^\top \right\| \cdot \sqrt{|\mathcal{I}_2|} \|\hat{\boldsymbol{\beta}}_{t,\omega}(\mathcal{I}_2) - \boldsymbol{\beta}_{t,\omega}\| \\
&\quad + \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \left\| \nu_{i,t} \cdot \mathbf{p}(X_i)^\top \right\| \cdot \left| Y_i^\dagger(t; \mathcal{I}_1) - Y_i^\dagger(t) \right| \cdot \sqrt{|\mathcal{I}_2|} \|\hat{\boldsymbol{\beta}}_{t,\omega}(\mathcal{I}_2) - \boldsymbol{\beta}_{t,\omega}\|
\end{aligned}$$

$$\lesssim_{\mathbf{p}} \left\| \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \nu_{i,t} \cdot Y_i^\dagger(t) \mathbf{p}(X_i)^\top \right\| + \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \left\| \nu_{i,t} \cdot \mathbf{p}(X_i)^\top \right\| \cdot \left| Y_i^\dagger(t; \mathcal{I}_1) - Y_i^\dagger(t) \right|.$$

The first term in the last line clearly converges to zero. The second one can be bounded as

$$\begin{aligned} \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \left\| \nu_{i,t} \cdot \mathbf{p}(X_i)^\top \right\| \cdot \left| Y_i^\dagger(t; \mathcal{I}_1) - Y_i^\dagger(t) \right| &\leq \sqrt{\frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \left\| \nu_{i,t} \cdot \mathbf{p}(X_i)^\top \right\|^2} \sqrt{\frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \left| Y_i^\dagger(t; \mathcal{I}_1) - Y_i^\dagger(t) \right|^2} \\ &\lesssim_{\mathbf{p}} \sqrt{\frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \left| Y_i^\dagger(t; \mathcal{I}_1) - Y_i^\dagger(t) \right|^2} \\ &\lesssim_{\mathbf{p}} \sqrt{\mathbb{E}_{X_i, Z_i} \left| Y_i^\dagger(t; \mathcal{I}_1) - Y_i^\dagger(t) \right|^2}, \end{aligned}$$

which again tends to zero due to Lemma 2.

Finally,

$$\begin{aligned} &\frac{1}{\sqrt{|\mathcal{I}_2|}} \sum_{i \in \mathcal{I}_2} \nu_{i,t} \cdot (\hat{\mu}_t^\dagger(X_i; \mathcal{I}_2) \hat{\omega}_t(X_i; \mathcal{I}_2) - \tilde{\mu}_t^\dagger(X_i) \omega_t(X_i)) \\ &= \frac{1}{\sqrt{|\mathcal{I}_2|}} \sum_{i \in \mathcal{I}_2} \nu_{i,t} \cdot \hat{\omega}_t(X_i; \mathcal{I}_2) (\hat{\mu}_t^\dagger(X_i; \mathcal{I}_2) - \tilde{\mu}_t^\dagger(X_i)) \\ &+ \frac{1}{\sqrt{|\mathcal{I}_2|}} \sum_{i \in \mathcal{I}_2} \nu_{i,t} \cdot \tilde{\mu}_t^\dagger(X_i) (\hat{\omega}_t(X_i; \mathcal{I}_2) - \omega_t(X_i)). \end{aligned}$$

Each of the terms can be bounded via

$$\begin{aligned} &\frac{1}{\sqrt{|\mathcal{I}_2|}} \sum_{i \in \mathcal{I}_2} \nu_{i,t} \cdot \hat{\omega}_t(X_i; \mathcal{I}_2) (\hat{\mu}_t^\dagger(X_i; \mathcal{I}_2) - \tilde{\mu}_t^\dagger(X_i)) \\ &\leq \left\| \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \nu_{i,t} \cdot \omega_t(X_i) \mathbf{p}(X_i)^\top \right\| \cdot \sqrt{|\mathcal{I}_2|} \left\| \hat{\beta}_{t,\mu}^\dagger(\mathcal{I}_2) - \beta_{t,\mu}^\dagger \right\| \\ &\quad + \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \left\| \nu_{i,t} \cdot \mathbf{p}(X_i)^\top \right\| \cdot \left\| \mathbf{p}(X_i)^\top \right\| \cdot \left\| \hat{\beta}_{t,\omega}(\mathcal{I}_2) - \beta_{t,\omega} \right\| \cdot \sqrt{|\mathcal{I}_2|} \left\| \hat{\beta}_{t,\mu}^\dagger(\mathcal{I}_2) - \beta_{t,\mu}^\dagger \right\| \\ &\lesssim_{\mathbf{p}} \left\| \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \nu_{i,t} \cdot \omega_t(X_i) \mathbf{p}(X_i)^\top \right\| + \frac{1}{|\mathcal{I}_2|^2} \sum_{i \in \mathcal{I}_2} \left\| \nu_{i,t} \cdot \mathbf{p}(X_i)^\top \right\| \cdot \left\| \mathbf{p}(X_i)^\top \right\| \xrightarrow{\mathbf{p}} 0, \end{aligned}$$

and

$$\begin{aligned} &\frac{1}{\sqrt{|\mathcal{I}_2|}} \sum_{i \in \mathcal{I}_2} \nu_{i,t} \cdot \tilde{\mu}_t^\dagger(X_i) (\hat{\omega}_t(X_i; \mathcal{I}_2) - \omega_t(X_i)) = \frac{1}{\sqrt{|\mathcal{I}_2|}} \sum_{i \in \mathcal{I}_2} \nu_{i,t} \cdot \tilde{\mu}_t^\dagger(X_i) \cdot \mathbf{p}(X_i)^\top \left(\hat{\beta}_{t,\omega}(\mathcal{I}_2) - \beta_{t,\omega} \right) \\ &\leq \left\| \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \nu_{i,t} \cdot \tilde{\mu}_t^\dagger(X_i) \cdot \mathbf{p}(X_i)^\top \right\| \cdot \sqrt{|\mathcal{I}_2|} \left\| \hat{\beta}_{t,\omega}(\mathcal{I}_2) - \beta_{t,\omega} \right\| \\ &\lesssim_{\mathbf{p}} \left\| \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \nu_{i,t} \cdot \tilde{\mu}_t^\dagger(X_i) \cdot \mathbf{p}(X_i)^\top \right\| \xrightarrow{\mathbf{p}} 0, \end{aligned}$$

which closes the proof.

1.5 Proof of Theorem 3

The first conclusion of the theorem, asymptotic equivalence, follows from Lemmas 1, 2, and 3. We therefore only calculate the asymptotic variance. Recall that the oracle estimator takes the form

$$\begin{aligned}\bar{\tau}_{t,t',\text{CALM}} &= \frac{1}{n} \sum_{i=1}^n \underbrace{\left[\frac{\mathbb{1}(T_i = t)}{e_t(X_i)} (Y_i - \mu_t(X_i)) + \mu_t(X_i) \right] - \left[\frac{\mathbb{1}(T_i = t')}{e_{t'}(X_i)} (Y_i - \mu_{t'}(X_i)) + \mu_{t'}(X_i) \right]}_{= A_i} \\ &\quad + \underbrace{\nu_{i,t} \cdot (Y_i^\dagger(t) - \mu_t^\dagger(X_i))}_{= B_i} \omega_t(X_i) + \underbrace{-\nu_{i,t'} \cdot (Y_i^\dagger(t') - \mu_{t'}^\dagger(X_i))}_{= C_i} \omega_{t'}(X_i).\end{aligned}$$

Clearly the asymptotic variance of the AIPW approach is simply the variance of A_i . Also recall that

$$\nu_{i,t} = 1 - \frac{\mathbb{1}(T_i = t)}{e_t(X_i)},$$

with

$$\mathbb{E}[\nu_{i,t}|X_i] = \mathbb{E}[\nu_{i,t}|X_i, Z_i] = 0, \quad \text{and} \quad \mathbb{V}[\nu_{i,t}|X_i] = \mathbb{V}[\nu_{i,t}|X_i, Z_i] = \frac{1 - e_t(X_i)}{e_t(X_i)} = \lambda_t(X_i).$$

In addition,

$$\mathbb{E}\left[\nu_{i,t} \frac{\mathbb{1}(T_i = t)}{e_t(X_i)} \middle| X_i\right] = \mathbb{E}\left[\nu_{i,t} \frac{\mathbb{1}(T_i = t)}{e_t(X_i)} \middle| X_i, Z_i\right] = -\lambda_t(X_i),$$

and for $t \neq t'$,

$$\mathbb{E}\left[\nu_{i,t'} \frac{\mathbb{1}(T_i = t)}{e_t(X_i)} \middle| X_i\right] = \mathbb{E}\left[\nu_{i,t'} \frac{\mathbb{1}(T_i = t)}{e_t(X_i)} \middle| X_i, Z_i\right] = 1,$$

as the indicators will be mutually exclusive. Finally, we also have for $t \neq t'$

$$\mathbb{E}[\nu_{i,t}\nu_{i,t'}|X_i] = \mathbb{E}[\nu_{i,t}\nu_{i,t'}|X_i, Z_i] = -1.$$

We start with the variances of B_i and C_i , which take the form

$$\mathbb{V}[B_i|X_i] = \lambda_t(X_i)\mathbb{V}[Y_i^\dagger(t)|X_i], \quad \mathbb{V}[C_i|X_i] = \lambda_{t'}(X_i)\mathbb{V}[Y_i^\dagger(t')|X_i].$$

Their covariance is

$$\mathbb{Cov}[B_i, C_i|X_i] = \mathbb{Cov}[Y_i^\dagger(t), Y_i^\dagger(t')|X_i].$$

The above calculation gives the form of the matrix

$$\Sigma^\dagger(X_i) = \begin{bmatrix} \lambda_t(X_i)\mathbb{V}[Y_i^\dagger(t)|X_i] & \mathbb{Cov}[Y_i^\dagger(t), Y_i^\dagger(t')|X_i] \\ \mathbb{Cov}[Y_i^\dagger(t), Y_i^\dagger(t')|X_i] & \lambda_{t'}(X_i)\mathbb{V}[Y_i^\dagger(t')|X_i] \end{bmatrix}.$$

We next compute the additional covariances, where we add an extra negative sign to match the expressions in the main paper:

$$\begin{aligned} -\mathbb{Cov}[A_i, B_i|X_i] &= \lambda_t(X_i)\mathbb{Cov}[Y_i(t), Y_i^\dagger(t)|X_i] + \mathbb{Cov}[Y_i(t'), Y_i^\dagger(t)|X_i] \\ -\mathbb{Cov}[A_i, C_i|X_i] &= \mathbb{Cov}[Y_i(t), Y_i^\dagger(t')|X_i] + \lambda_{t'}(X_i)\mathbb{Cov}[Y_i(t'), Y_i^\dagger(t')|X_i], \end{aligned}$$

which corresponds to the matrix

$$\sigma(X_i) = \begin{bmatrix} \lambda_t(X_i)\mathbb{Cov}[Y_i(t), Y_i^\dagger(t)|X_i] + \mathbb{Cov}[Y_i(t'), Y_i^\dagger(t)|X_i] \\ \mathbb{Cov}[Y_i(t), Y_i^\dagger(t')|X_i] + \lambda_{t'}(X_i)\mathbb{Cov}[Y_i(t'), Y_i^\dagger(t')|X_i] \end{bmatrix}.$$

This completes the variance calculation.

1.6 Proof of Theorem 4

We start by providing the regularity conditions needed for CATE estimation, which are omitted from the main paper to conserve space. The first assumption below imposes smoothness for certain nuisance functions. Let $f(\cdot)$ denote the density function of the structured covariates. We fix an evaluation point x .

Assumption CATE.1 (Smoothness of nuisance functions). *The functions $f(\cdot)$, $\mu_t^\dagger(\cdot)$, $\Sigma^\dagger(\cdot)$ and $\sigma(\cdot)$ are continuous in a neighborhood of x . $\mu_t(\cdot)$ is s -times continuously differentiable in a neighborhood of x with $s \geq 1$. In addition, $f(x) > 0$ and $\Sigma^\dagger(x)$ is nonsingular.*

We also require the true potential outcomes to have bounded fourth conditional moments, which helps establish the asymptotic normality of our CATE estimator.

Assumption CATE.2 (Bounded fourth moments). *There exists some $c > 0$ such that $\mathbb{E}[Y_i(t)^4|X_i] \leq c$ for all $t = 1, 2, \dots, k$.*

The next assumption concerns the kernel function. In particular, we allow the use of both second- and higher-order kernels, where the latter can be appealing for bias reduction.

Assumption CATE.3 (Kernel function). *The kernel function $k(\cdot)$ is continuous, compactly supported, and integrates to 1. In addition, its order is at least $s + 1$.*

As the CATE estimator is local in nature, we assume the nonparametric estimates are uniformly consistent in a local neighborhood of the evaluation point of interest.

Assumption CATE.4 (Consistent nuisance function estimation). *The cross-fitted function estimates satisfy: for some $\epsilon > 0$,*

$$\begin{aligned} \sup_{|x'-x| \leq \epsilon} |\hat{\mu}_t(x'; \mathcal{I}) - \mu_t(x')| &\xrightarrow{P} 0, \quad \sup_{|x'-x| \leq \epsilon} |\hat{\mu}_t^\dagger(x'; \mathcal{I}) - \mu_t^\dagger(x')| \xrightarrow{P} 0, \\ \sup_{|x'-x| \leq \epsilon} |\hat{\omega}_{\text{ATE}}(x'; \mathcal{I}) - \omega_{\text{ATE}}(x')| &\xrightarrow{P} 0, \end{aligned}$$

for each data fold $\mathcal{I}$ and treatment arm $t, t' \in \{1, 2, \dots, k\}$.

We are ready to prove the theorem. We first show the formula of the variance comparison.

Variance calculation. From the calculation in Section 3.2 of the main paper, one has

$$\begin{aligned} V_{t,t', \text{CALM}}(x) &= nh^p \cdot \mathbb{V}[\bar{\tau}_{t,t', \text{CALM}}(x) | \mathbf{X}] \\ &= nh^p \cdot \mathbb{V}[\bar{\tau}_{t,t', \text{AIPW}}(x) | \mathbf{X}] + nh^p \sum_{i=1}^n \kappa_i(x)^2 \left(\omega(X_i)^\top \Sigma^\dagger(X_i) \omega(X_i) - 2\omega(X_i)^\top \sigma(X_i) \right). \end{aligned}$$

We now further expand on the last term:

$$\begin{aligned} &nh^p \sum_{i=1}^n \kappa_i(x)^2 \left(\omega(X_i)^\top \Sigma^\dagger(X_i) \omega(X_i) - 2\omega(X_i)^\top \sigma(X_i) \right) \\ &= nh^p \sum_{i=1}^n \frac{1}{h^{2p}} k\left(\frac{X_i - x}{h}\right)^2 \left(\omega(X_i)^\top \Sigma^\dagger(X_i) \omega(X_i) - 2\omega(X_i)^\top \sigma(X_i) \right) \bigg/ \left(\sum_{i=1}^n \frac{1}{h^p} k\left(\frac{X_i - x}{h}\right) \right)^2 \\ &= \sum_{i=1}^n \frac{1}{nh^p} k\left(\frac{X_i - x}{h}\right)^2 \left(\omega(X_i)^\top \Sigma^\dagger(X_i) \omega(X_i) - 2\omega(X_i)^\top \sigma(X_i) \right) \bigg/ \left(\sum_{i=1}^n \frac{1}{nh^p} k\left(\frac{X_i - x}{h}\right) \right)^2. \end{aligned}$$

Convergence of both the numerator and the denominator follows from standard mean and variance calculations. To illustrate, consider the numerator. Its mean is

$$\begin{aligned} &\mathbb{E} \left[\sum_{i=1}^n \frac{1}{nh^p} k\left(\frac{X_i - x}{h}\right)^2 \left(\omega(X_i)^\top \Sigma^\dagger(X_i) \omega(X_i) - 2\omega(X_i)^\top \sigma(X_i) \right) \right] \\ &= \int_{\mathcal{X}} \frac{1}{h^p} k\left(\frac{u - x}{h}\right)^2 \left(\omega(u)^\top \Sigma^\dagger(u) \omega(u) - 2\omega(u)^\top \sigma(u) \right) f(u) du \\ &= \int_{(\mathcal{X} - x)/h} k(v)^2 \left(\omega(x + hv)^\top \Sigma^\dagger(x + hv) \omega(x + hv) - 2\omega(x + hv)^\top \sigma(x + hv) \right) f(x + hv) dv \\ &= \kappa_2 f(x) \left(\omega(x)^\top \Sigma^\dagger(x) \omega(x) - 2\omega(x)^\top \sigma(x) \right) + o(1). \end{aligned}$$

On the other hand, the variance satisfies

$$\begin{aligned} &\mathbb{V} \left[\sum_{i=1}^n \frac{1}{nh^p} k\left(\frac{X_i - x}{h}\right)^2 \left(\omega(X_i)^\top \Sigma^\dagger(X_i) \omega(X_i) - 2\omega(X_i)^\top \sigma(X_i) \right) \right] \\ &\lesssim \frac{1}{n} \mathbb{V} \left[\frac{1}{h^p} k\left(\frac{X_i - x}{h}\right)^2 \left(\omega(X_i)^\top \Sigma^\dagger(X_i) \omega(X_i) - 2\omega(X_i)^\top \sigma(X_i) \right) \right] \lesssim \frac{1}{nh^p} \rightarrow 0. \end{aligned}$$

As such,

$$\sum_{i=1}^n \frac{1}{nh^p} k\left(\frac{X_i - x}{h}\right)^2 \left(\omega(X_i)^\top \Sigma^\dagger(X_i) \omega(X_i) - 2\omega(X_i)^\top \sigma(X_i) \right) \xrightarrow{P} \kappa_2 f(x) \left(\omega(x)^\top \Sigma^\dagger(x) \omega(x) - 2\omega(x)^\top \sigma(x) \right).$$

The same analysis applies to the denominator, so that

$$\sum_{i=1}^n \frac{1}{nh^p} k\left(\frac{X_i - x}{h}\right) \xrightarrow{P} f(x).$$

Asymptotic equivalence. We now show that the CALM estimator of the CATE is asymptotically equivalent to the oracle one. To reduce notational burden, we focus on one treatment arm, leading to the expansion:

$$\sqrt{nh^p} \left(\hat{\mu}_{t,\text{CALM}}^{\mathcal{I}_3} - \bar{\mu}_{t,\text{CALM}}^{\mathcal{I}_3} \right) = \sqrt{nh^p} \sum_{i \in \mathcal{I}_3} \kappa_i(x; \mathcal{I}_3) \nu_{i,t} \left(\hat{\mu}_t(X_i; \mathcal{I}_2) - \mu_t(X_i) \right) \quad (\text{I})$$

$$- \sqrt{nh^p} \sum_{i \in \mathcal{I}_3} \kappa_i(x; \mathcal{I}_3) \nu_{i,t} \left(\hat{\mu}_t^\dagger(X_i; \mathcal{I}_2) - \mu_t^\dagger(X_i) \right) \hat{\omega}_t(X_i; \mathcal{I}_2) \quad (\text{II})$$

$$- \sqrt{nh^p} \sum_{i \in \mathcal{I}_3} \kappa_i(x; \mathcal{I}_3) \nu_{i,t} \mu_t^\dagger(X_i) \left(\hat{\omega}_t(X_i; \mathcal{I}_2) - \omega_t(X_i) \right) \quad (\text{III})$$

$$+ \sqrt{nh^p} \sum_{i \in \mathcal{I}_3} \kappa_i(x; \mathcal{I}_3) \nu_{i,t} \left(\hat{\omega}_t(X_i; \mathcal{I}_2) - \omega_t(X_i) \right) Y_i^\dagger(t) \quad (\text{IV})$$

$$+ \sqrt{nh^p} \sum_{i \in \mathcal{I}_3} \kappa_i(x; \mathcal{I}_3) \nu_{i,t} \hat{\omega}_t(X_i; \mathcal{I}_2) \left(Y_i^\dagger(t; \mathcal{I}_1) - Y_i^\dagger(t) \right). \quad (\text{V})$$

Due to sample splitting, it is easy to see that they all are conditional mean 0. We therefore consider their conditional variance. Let $\mathbf{X}(\mathcal{I}_3)$ denote all structured covariates from fold $\mathcal{I}_3$. Also define the local sup-norm $\|\cdot\|_{\infty, \epsilon} = \sup_{| \cdot - x | \leq \epsilon} \|\cdot\|$. Starting from (I),

$$\begin{aligned} \mathbb{V}[(\text{I}) | \mathcal{I}_1, \mathcal{I}_2, \mathbf{X}(\mathcal{I}_3)] &= nh^p \sum_{i \in \mathcal{I}_3} \kappa_i(x; \mathcal{I}_3)^2 \lambda_t(X_i) \left(\hat{\mu}_t(X_i; \mathcal{I}_2) - \mu_t(X_i) \right)^2 \\ &\lesssim \|\hat{\mu}_t(\cdot; \mathcal{I}_2) - \mu_t(\cdot)\|_{\infty, \epsilon} \left(nh^p \sum_{i \in \mathcal{I}_3} \kappa_i(x; \mathcal{I}_3)^2 \right). \end{aligned}$$

By our assumption, the first term in the last line vanishes asymptotically. With the same mean and variance calculation, the second term in the above converges to $\kappa_2/f(x)$. As a result, (I) is asymptotically vanishing. The same technique applies to (II), (III) and (IV):

$$\begin{aligned} \mathbb{V}[(\text{II}) | \mathcal{I}_1, \mathcal{I}_2, \mathbf{X}(\mathcal{I}_3)] &= nh^p \sum_{i \in \mathcal{I}_3} \kappa_i(x; \mathcal{I}_3)^2 \lambda_t(X_i) \hat{\omega}_t(X_i; \mathcal{I}_2)^2 \left(\hat{\mu}_t^\dagger(X_i; \mathcal{I}_2) - \mu_t^\dagger(X_i) \right)^2 \\ &\lesssim \|\hat{\mu}_t^\dagger(X_i; \mathcal{I}_2) - \mu_t^\dagger(X_i)\|_{\infty, \epsilon} \left(nh^p \sum_{i \in \mathcal{I}_3} \kappa_i(x; \mathcal{I}_3)^2 \right) \xrightarrow{P} 0. \end{aligned}$$

$$\begin{aligned}
\mathbb{V}[(\text{III})|\mathcal{I}_1, \mathcal{I}_2, \mathbf{X}(\mathcal{I}_3)] &= nh^p \sum_{i \in \mathcal{I}_3} \kappa_i(x; \mathcal{I}_3)^2 \lambda_t(X_i) \mu_t^\dagger(X_i)^2 \left(\hat{\omega}_t(X_i; \mathcal{I}_2) - \omega_t(X_i) \right)^2 \\
&\lesssim \|\hat{\omega}_t(X_i; \mathcal{I}_2) - \omega_t(X_i)\|_{\infty, \epsilon} \left(nh^p \sum_{i \in \mathcal{I}_3} \kappa_i(x; \mathcal{I}_3)^2 \right) \xrightarrow{\mathbb{P}} 0,
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{V}[(\text{IV})|\mathcal{I}_1, \mathcal{I}_2, \mathbf{X}(\mathcal{I}_3)] &= nh^p \sum_{i \in \mathcal{I}_3} \kappa_i(x; \mathcal{I}_3)^2 \lambda_t(X_i) Y_i^\dagger(t)^2 \left(\hat{\omega}_t(X_i; \mathcal{I}_2) - \omega_t(X_i) \right)^2 \\
&\lesssim \|\hat{\omega}_t(X_i; \mathcal{I}_2) - \omega_t(X_i)\|_{\infty, \epsilon} \left(nh^p \sum_{i \in \mathcal{I}_3} \kappa_i(x; \mathcal{I}_3)^2 \right) \xrightarrow{\mathbb{P}} 0.
\end{aligned}$$

Now consider (V). Variance computation yields

$$\begin{aligned}
\mathbb{V}[(\text{V})|\mathcal{I}_1, \mathcal{I}_2, \mathbf{X}(\mathcal{I}_3)] &= nh^p \sum_{i \in \mathcal{I}_3} \kappa_i(x; \mathcal{I}_3)^2 \hat{\omega}_t(X_i; \mathcal{I}_2)^2 \left(Y_i^\dagger(t; \mathcal{I}_1) - Y_i^\dagger(t) \right)^2 \\
&\lesssim_{\mathbb{P}} \frac{1}{nh^p} \sum_{i \in \mathcal{I}_3} k\left(\frac{X_i - x}{h}\right)^2 \left(Y_i^\dagger(t; \mathcal{I}_1) - Y_i^\dagger(t) \right)^2.
\end{aligned}$$

As the kernel function is bounded and supported on a compact set, the above has the bound

$$\begin{aligned}
&\frac{1}{nh^p} \sum_{i \in \mathcal{I}_3} k\left(\frac{X_i - x}{h}\right)^2 \left(Y_i^\dagger(t; \mathcal{I}_1) - Y_i^\dagger(t) \right)^2 \\
&\lesssim_{\mathbb{P}} \mathbb{E} \left[\frac{1}{nh^p} \sum_{i \in \mathcal{I}_3} k\left(\frac{X_i - x}{h}\right)^2 \left(Y_i^\dagger(t; \mathcal{I}_1) - Y_i^\dagger(t) \right)^2 \middle| (X_i, Z_i) : i \in \mathcal{I}_3 \right] \\
&= \frac{1}{nh^p} \sum_{i \in \mathcal{I}_3} k\left(\frac{X_i - x}{h}\right)^2 \mathbb{E} \left[\left(Y_i^\dagger(t; \mathcal{I}_1) - Y_i^\dagger(t) \right)^2 \middle| X_i, Z_i \right],
\end{aligned}$$

where from the proof of Lemma 2, the conditional expectation in the last line is of order $B^{-1} + n^{-1}$ uniformly in X_i and Z_i .

Asymptotic bias order. Given the asymptotically linear representation, we note that the asymptotic bias is determined by

$$\sqrt{nh^p} \mathbb{E} \left[\frac{1}{h^p} k\left(\frac{X_i - x}{h}\right) (\mu_t(X_i) - \mu_{t'}(X_i) - \mu_t(x) + \mu_{t'}(x)) \right],$$

which vanishes as long as $nh^{p+2s} \lesssim 1$.

Asymptotic normality. To establish asymptotic normality, it suffices to verify a conditional Lindeberg condition with fourth moments, which follows from our assumption on the bandwidth sequence and the finite fourth moments of the potential outcomes.

2 Motivating Real-world RCTs

In this section, we present five motivating RCTs that collect unstructured data such as survey responses, video, and audio recordings from participants. These trials span cancer, mental health, and dementia care, as summarized in Table 1, and demonstrate the broad applicability of our proposed methods in real-world settings.

Trial Name	Brief Summary	Patient-Centered Outcomes	Effect Modifiers	Unstructured Pre-treatment Data
A. Precision Clinical Trial Recruitment to Promote Cancer Health Equity [3]	Cluster RCTs on the ALEX Portal testing a virtual Community Health Educator; compares consent and engagement tactics to increase trust, participation, and referrals in Florida.	Referral to NCI-supported trials; engagement with trial education.	Race/ethnicity; language preference; geographic region.	Interaction logs with virtual Community Health Educator.
B. Tailoring Recruitment Communication using Virtual Human Technology for Older Minority Adults [1]	Cluster RCT on the ALEX platform testing a Virtual Human intervention to improve enrollment of 1,363 underrepresented older adults in NIH trials using remote, patient-informed consent and communication.	Enrollment in active NIH-funded trials; engagement with culturally sensitive recruitment materials.	Race/ethnicity; rural vs. urban residence; age; modality of VHT delivery.	Survey responses; interaction with Virtual Human Technology.
C. Researching/Improving Psychotherapy Techniques in Interventions Depression Trial [5]	RCT comparing Cognitive Behavioral Therapy vs. Acceptance and Commitment Therapy for major depressive disorder in 100 adults (California, USA).	Improvement in depressive symptoms; improvement in quality of life.	Treatment modality; age; sex; race/ethnicity; socioeconomic status; pretreatment depressive severity.	Video, audio, and transcripts of digital therapy interactions.
D. BRIGHTEN: Bridging Research Innovations for Greater Health in Technology, Emotion, and Neuroscience [4]	RCT evaluating mobile app-based therapy for depression management among 2,193 U.S. patients; supports symptom tracking and management.	Depression severity via mobile surveys; app engagement; behavioral changes.	Baseline depression severity; demographics; behavioral data.	Open-ended survey responses; user satisfaction; motivation for app use.
E. Care Ecosystem Dementia Randomized Controlled Trial [2]	RCT of a telehealth intervention for 460 patients with dementia and their caregivers.	Cost of care reimbursed by Medicare.	Age; sex; race; ethnicity; dementia severity.	Unstructured text from telephone surveys.

Table 1: Five motivating trials that collect unstructured pre-treatment data.

3 Additional Simulation Study Results

In this section, we present additional simulation study results. First, we provide a simulation study comparing different machine learning methods to estimate the conditional outcome model for CALM and AIPW in Figure 1. Second, we provide a table of the coverage probabilities (Table 2) for all the methods discussed in the main manuscript. Third, we provide example prompts under zero-shot prediction, few-shot prediction, and different prompt engineering techniques.

For both CALM and AIPW, when estimating the conditional mean model, we consider two machine learning methods: random forest (RF) and gradient boosting (GB). We refer to the CALM with zero-shot learning as CALM(zero-shot)+RF and CALM(zero-shot)+GB and the AIPW-based method as AIPW+RF and AIPW+GB. We present the absolute bias and the standard deviation in Figure 1. Figure 1 shows that for both CALM and AIPW, using random forest for conditional outcome estimation yields slightly lower standard deviation and thus higher estimation efficiency. Therefore, in the main manuscript, we adopt the random forest to estimate the conditional outcome model for all the methods in comparison.

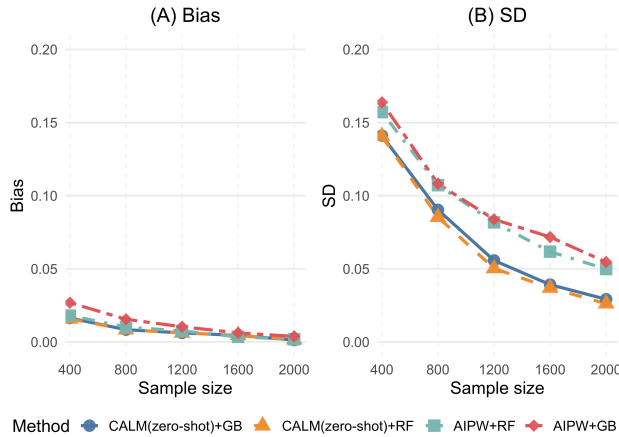


Figure 1: Absolute bias and standard deviation of ATE estimates for CALM (zero-shot) and AIPW-based methods across 300 Monte Carlo simulations. “CALM(zero-shot)+GB” and “CALM(zero-shot)+RF” refer to using gradient boosting and random forest for estimating the conditional mean model, respectively. “AIPW+GB” and “AIPW+RF” refer to the standard AIPW estimator where the conditional mean model is estimated using gradient boosting and random forest, respectively.

In what follows, we provide an example of the zero-shot learning prompt in Figure 2, the few-shot learning prompt in Figure 3, and example prompts under different prompt engineering techniques in Figure 4.

Table 2: Comparison of coverage probabilities.

Method	Coverage probability (SE)	
	$n = 400$	$n = 2,000$
AIPW+RF	0.94(0.01)	0.96(0.01)
AIPW+GB	0.96(0.01)	0.97(0.01)
CALM(zero-shot)+GB	0.95(0.02)	0.95(0.01)
CALM(zero-shot)+RF	0.94(0.02)	0.96(0.01)
CALM(few-shot, $m=6$)	0.94(0.01)	0.95(0.01)
CALM(few-shot, $m=10$)	0.96(0.01)	0.95(0.02)
CALM(few-shot, $m=14$)	0.95(0.02)	0.94(0.01)
AIPW(zero-shot covariate)	0.96(0.01)	0.94(0.01)
AIPW(few-shot covariate, $m=6$)	0.92(0.01)	0.93(0.01)
AIPW(few-shot covariate, $m=10$)	0.90(0.02)	0.92(0.02)
AIPW(few-shot covariate, $m=14$)	0.89(0.02)	0.92(0.01)

Example prompt for zero-shot learning

Instruction: You are a licensed clinical psychologist with expertise in evaluating depression severity. Your task is to classify the depression severity category for a BRIGHTEN study participant based on their demographic characteristics, treatment arm, satisfaction with the study app, and reason for enrollment. Use your professional judgment to interpret the information and determine the most appropriate severity level. Predict the participant’s end-of-study depression severity using exactly one word chosen from the following categories: [Minimal, Mild, Moderate, High, Severe].

Input: The participant is enrolled in the [treatment arm] of the BRIGHTEN study. They are a [Male], [Hispanic/Latino], [45] years old, have completed [High school degree], and are currently [working]. Their marital status is [Married/Partner]. The participant joined the study because [“to gain compensation”]. The participant’s satisfaction level with the app is [“i thought it may be like luminosity. but i was hoping it would help my mood swings”].

Figure 2: Example zero-shot learning prompt. The input example is synthetic and does not reflect real participant information from the study.

Example prompt for few-shot learning
Instruction: You are a licensed clinical psychologist with expertise in evaluating depression severity. Your task is to classify the depression severity category for a BRIGHTEN study participant based on their demographic characteristics, treatment arm, satisfaction with the study app, and reason for enrollment. Use your professional judgment to interpret the information and determine the most appropriate severity level. Based on the following example participant profiles and their depression level, predict the participant’s end-of-study depression severity using exactly one word chosen from the following categories: [Minimal, Mild, Moderate, High, Severe]. Input: Example 1: The participant is enrolled in the [treatment arm] of the BRIGHTEN study. They are a [Male], [Hispanic/Latino], [45] years old, have completed [High school degree], and are currently [working]. Their marital status is [Married/Partner]. The participant joined the study because [“to gain compensation”]. The participant’s satisfaction level with the app is [“i thought it may be like luminosity. but i was hoping it would help my mood swings”] Example 2: The participant is enrolled in the [control arm] of the BRIGHTEN study. They are a [Male], [African-American], [32] years old, have completed [Graduate degree], and are currently [working]. Their marital status is [Married/Partner]. The participant joined the study because [“Interest”]. The participant’s satisfaction level with the app is [“New ways to deal with stress and mood”]. The participant has [Minimal] level of depression. Example 3: The participant is enrolled in the [treatment arm] of the BRIGHTEN study. They are a [Female], [Hispanic/Latino], [18] years old, have completed [University], and are currently [not working]. Their marital status is [Single]. The participant joined the study because [“To be part of a study”]. The participant’s satisfaction level with the app is [“NA”]. The participant has [Moderate] level of depression.

Figure 3: Example few-shot learning prompt. All input examples are synthetic and do not reflect real participant information from the study.

Example prompts under different prompt engineering techniques
(1) Self-consistency. Instruction: You are acting as a licensed clinical psychologist specializing in depression assessment. Your task is to determine the depression severity category for a BRIGHTEN study participant using a self-consistency approach. First, make three independent predictions of depression severity based solely on the participant’s demographic information, treatment arm, satisfaction with the study app, and stated reason for enrollment. For each prediction, consider how these factors might be associated with depressive symptoms, drawing on your professional knowledge of depression severity categories. After completing the three predictions, select the category that appears most frequently among them. This will be your final answer. Predict the participant’s end-of-study depression severity using exactly one word chosen from the following categories: [Minimal, Mild, Moderate, High, Severe]. (2) Role-based. Instruction: You are a licensed clinical psychologist with expertise in evaluating depression severity. Your task is to classify the depression severity category for a BRIGHTEN study participant based on their demographic characteristics, treatment arm, satisfaction with the study app, and reason for enrollment. Use your professional judgment to interpret the information and determine the most appropriate severity level. Predict the participant’s end-of-study depression severity using exactly one word chosen from the following categories: [Minimal, Mild, Moderate, High, Severe]. (3) Decomposition. Instruction: You are a licensed clinical psychologist tasked with classifying the depression severity category for a BRIGHTEN study participant. Break down your reasoning into the following steps before making the final classification: (1) Mood-related indicators – Based on the participant’s demographic information, satisfaction with the study app, and reason for enrollment, assess potential signs of mood disturbance or stability. (2) Functional status – Evaluate how the participant’s working status, marital status, and other life circumstances may reflect or influence functional impairment. (3) Overall symptom severity – Integrate observations from steps 1 and 2 to estimate the likely severity of depressive symptoms. After completing these three steps, use your professional judgment to determine the most appropriate overall depression severity category. Predict the participant’s end-of-study depression severity using exactly one word chosen from the following categories: [Minimal, Mild, Moderate, High, Severe]. (4) Contrastive Instruction: You are a licensed clinical psychologist classifying depression severity for BRIGHTEN study participants. Compare the target participant to the two example participants below and determine the most appropriate severity category. Focus on differences and similarities in demographics, satisfaction level, and reason for enrollment when making your decision. Borderline example 1: Participant in the treatment arm, Female, White, age 45, College education, Employed, Married, satisfaction with the app “Maybe give me a lil bit of pleasure”, enrollment reason: “Interested in improving mental well-being.” This participant has mild level depression. Borderline example 2: Participant in the control arm, Female, White, age 45, College education, Employed, Married, satisfaction: “Had no hopes for this app.” Enrollment reason: “Persistent sadness, loss of interest, trouble sleeping.” This participant has moderate level depression. Predict the participant’s end-of-study depression severity using exactly one word chosen from the following categories: [Minimal, Mild, Moderate, High, Severe]. Input: The participant is enrolled in the [treatment arm] of the BRIGHTEN study. They are a [female], [Asian], [28] years old, have completed [graduate degree], and are currently [working]. Their marital status is [single]. The participant joined the study because [“\$, gift cards”]. The participant’s satisfaction level with the app is [“i thought it may be like luminosity. but i was hoping it would help my mood swings”].

Figure 4: Example zero-shot learning prompts under four different prompt engineering techniques. The input example is synthetic and does not reflect real participant information from the study.

References

- [1] Bryan, E. G., Chen, H., Vilaro, M., Chu, H., Grillo, G., Te, P. *et al.* (2025). “Developing a Supportive Virtual Human to Deliver Clinical Trial Education for Older Women and Other Populations Historically Excluded From Research,” *Patient Education and Counseling*, 130, 108485.
- [2] Guterman, E. L., Kiekhoefer, R. E., Wood, A. J., Allen, I. E., Kahn, J. G., Dulaney, S. *et al.* (2023). “Care Ecosystem Collaborative Model and Health Care Costs in Medicare Beneficiaries with Dementia: A Secondary Analysis of a Randomized Clinical Trial,” *JAMA Internal Medicine*, 183(11), 1222–1228.
- [3] Hanvey, G., Dede, D., Krieger, J., Ross, K., Amdur, R., and Pereira, D. (2024). “Investigating Inequities in Cancer Clinical Trial Participation: Understanding Demographic and Socioeconomic Determinants of Behavioral Cancer Research Participation,” *Psychosomatic Medicine*, 86(5), A50.
- [4] Pratap, A., Homiar, A., Waninger, L., Herd, C., Suver, C., Volponi, J. *et al.* (2022). “Real-World Behavioral Dataset from Two Fully Remote Smartphone-Based Randomized Clinical Trials for Depression,” *Scientific Data*, 9(1), 522.
- [5] Ramos, F. N., Bernstein, R. A., and Ezawa, I. D. (2024). “Assessing Predictive Factors of Attitudes Toward Peer-Supported Mental Health Interventions in the Metaverse: Mixed Methods Study,” *JMIR XR and Spatial Computing (JMXR)*, 1(1), e57990.