Can language models boost the power of randomized experiments without statistical bias?

Xinrui Ruan[1] Xinwei Ma[2] Yingfei Wang[3] Waverly Wei[4] Jingshen Wang[1]

[1]Division of Biostatistics, University of California, Berkeley
[2]Department of Economics, University of California, San Diego
[3]Michael G. Foster School of Business, University of Washington
[4]Department of Data Sciences and Operations, University of Southern California

December 6, 2025

Abstract

Randomized experiments or randomized controlled trials (RCTs) are gold standards for causal inference, yet cost and sample-size constraints limit power. We introduce **CALM** (*Causal Analysis leveraging Language Models*), a statistical framework that integrates large language models (LLMs) generated insights of RCTs with established causal estimators to increase precision while preserving statistical validity. In particular, CALM treats LLM-generated outputs as auxiliary prognostic information and corrects their potential bias via a heterogeneous calibration step that residualizes and optimally reweights predictions. We prove that CALM remains consistent even when LLM predictions are biased and achieves efficiency gains over augmented inverse probability weighting estimators for various causal effects. In particular, CALM develops a few-shot variant that aggregates predictions across randomly sampled demonstration sets. The resulting U-statistic-like predictor restores i.i.d. structure and also mitigates prompt-selection variability. Empirically, in simulations calibrated to a mobile-app depression RCT, CALM delivers lower variance relative to other benchmarking methods, is effective in zero- and few-shot settings, and remains stable across prompt designs. By principled use of LLMs to harness unstructured data and external knowledge learned during pretraining, CALM provides a practical path to more precise causal analyses in RCTs.

Keywords: Artificial intelligence; Clinical trials; Few-shot learning.

1 Introduction

1.1 Background and motivation

Reliably designed randomized controlled trials (RCTs) are widely regarded as the gold standard for evaluating the effectiveness of biomedical, behavioral, socioeconomic or business interventions. By leveraging randomization and controlled conditions, RCTs minimize bias and facilitate the credible estimation of causal effects. Additionally, subgroup analyses are frequently employed to explore treatment effect heterogeneity, thereby supporting the advancement of individualized treatment targeting. However, most datasets from RCTs are analyzed in isolation, without incorporating external information. Moreover, standard estimation strategies for average and subgroup effects often rely on regression analyses using only a few hand-picked pre-intervention covariates. This approach leaves significant baseline information unexploited, particularly variables of multi-modality nature or those subject to substantial missingness. These limitations collectively compromise statistical efficiency, a critical drawback given that RCTs are often constrained by their sample sizes due to high implementation costs.

Recent advances in artificial intelligence (AI), large-scale pre-trained large language models (LLMs) in particular, offer potential opportunities to boost the statistical power of RCTs by leveraging the external knowledge obtained during training. LLMs are deep neural networks trained with self-supervision on broad and heterogeneous corpora, producing general representations that can be rapidly adapted to many downstream tasks with little or no additional labeled data [5]. Importantly, multimodal LLMs now span the biomedical and behavioral spectrum, with representative examples including GPT-4 [38], Gemini [20], and LLaMA [50], meaning that they potentially encode latent biological or behavioral information about the treatments [5, 21, 33]. When LLM-generated predictions are integrated into established causal inference methods in a statistically principled manner, we hypothesize that this can reduce the variance of treatment effect estimates and boost the power to detect treatment effect heterogeneity.

As an added benefit, LLMs are naturally well-suited for extracting rich information from the unstructured data increasingly collected in modern RCTs. Unstructured data, such as clinical notes, medical images, participant narratives, and transcripts of subject-provider interactions, are now standard in RCTs across oncology, mental health [41, 45], dementia [22], and healthcare delivery [7, 23] (see Table 1 in the Supplementary Materials for motivating examples). Historically, incorporating such data into causal analyses has often required ad hoc feature-engineering pipelines that can discard nuanced information, introduce human biases [26], or create additional high-dimensional covariates. By contrast, multimodal LLMs offer a more flexible and principled way to process these complex unstructured inputs.

Yet, developing rigorous and statistically valid causal inference methods for RCTs that integrate LLM outputs remains a major challenge. This is because naively plugging LLM-based predictions into causal estimators can induce bias and lead to problematic statistical inference, as LLMs lack guarantees on prediction accuracy and may rely on spurious correlations. Therefore, any causal

inference method that incorporates LLM-based predictions must be explicitly robust to potential prediction errors.

1.2 Contribution

In this manuscript, we propose a **C**ausal **A**nalysis leveraging **L**anguage **M**odels (CALM) framework, to address the statistical challenges mentioned above. In particular, our contributions can be organized as follows.

CALM introduces a novel statistical framework for valid inference on a broad class of causal parameters by leveraging LLM prediction in randomized experiments. Section 2 focuses on its applications to mean potential outcomes, and Section 3 extends the framework to average and heterogeneous average treatment effects. We demonstrate that, even when LLM predictions are biased, CALM remains both consistent and more efficient than augmented inverse propensity weighting (AIPW) estimators. This advantage stems from the heterogeneous calibration strategy introduced in Section 2 (Steps 3 and 3′ in particular), which both corrects bias in LLM-generated predictions and leverages LLM's intrinsic knowledge and serialized input data to sharpen causal estimation precision.

Another key strength and novelty of CALM is its ability to incorporate few-shot predictions from LLMs into causal effect estimation (Section 2.2), a challenge that has not been systematically discussed and addressed in the existing literature. Although few-shot learning is well established in computer science for its improved prediction performance and rapid task adaptation [6, 19, 37, 46], integrating it into existing causal frameworks is challenging. This is because few-shot predictions are inherently correlated, violating the independence and identical distribution (aka i.i.d.) assumption required by many classical methods. Moreover, few-shot predictions are sensitive to the choice of demonstrative samples, introducing additional variability into causal effect estimation. To address these challenges, CALM augments few-shot learning with resampling-based prediction aggregation (Step 2′ in Section 2.2). We theoretically show that averaging predictions across multiple small, randomly drawn sets of demonstrative samples yields an LLM-based predictor with a U-statistics-like structure, which is not only robust to demonstrative sample selection but also behaves in an i.i.d. manner in large samples. The asymptotic properties of the resulting estimator with this averaged predictor are established in Theorem 2 of Section 4.

This manuscript also contributes to the literature on causal inference using potentially biased auxiliary information [3, 15, 34, 56]. Specifically, we demonstrate in Theorem 1 of Section 4 that our CALM estimator remains consistent even when the zero-shot predictions are biased for the mean potential outcomes. This robustness arises from a novel estimating equation that incorporates residualized LLM predictions, which are mean-zero by construction. Our procedure relies on four nuisance function estimates: the conditional expectation of the potential outcomes (as in classical augmented inverse probability weighting method [8, 16, 36]), the conditional mean of the zero-shot LLM predictions (for the residualization step), and the conditional variance and covariance of the true and predicted potential outcomes (used to construct the optimal weighting scheme).

By employing sample splitting and Neyman orthogonalization, we achieve asymptotically normal distributional approximations and valid statistical inference under mild L_2-type conditions on these nuisance function estimates.

In addition to its methodological and theoretical advances, our work provides systematic empirical evidence on the practical utility of CALM. Using synthetic populations calibrated to the BRIGHTEN study with structured and unstructured data (Section 5.2), we demonstrate that CALM yields more accurate estimates of causal effects, achieving lower bias and reduced variance compared to benchmark methods (Section 5.3, Figure 3). In particular, CALM's efficiency gain is independent of whether unstructured data are used. This indicates that the improvement does not stem from incorporating additional prognostic information from unstructured inputs, but rather from the intrinsic information gain provided by the LLM itself (Section 5.4, Figure 4). We further demonstrate that CALM remains effective in both zero- and few-shot settings without domain-specific fine-tuning, adapts well to heterogeneous covariate strata (Section 5.4, Figure 5), and avoids the under-coverage issues that arise when LLM predictions are naively incorporated as covariates in benchmark methods (Figure 3). In the BRIGHTEN case study in Section 6, CALM not only yields tighter confidence intervals than benchmark methods, but also detects subgroup treatment effects (e.g., among female Hispanic participants) that benchmark methods miss, indicating its potential to enhance the discovery of treatment heterogeneity. Finally, robustness checks across multiple LLMs and diverse prompt-engineering strategies in Section 6 confirm that CALM's empirical performance is stable to design choices, thereby reducing practical barriers to applying LLMs in real-world experiment analysis.

1.3 Related literature

Our work relates to the literature on improving the efficiency of randomized experiments through covariate adjustment. Classical approaches such as regression adjustment and stratification are well known to reduce variance without compromising statistical validity [18, 31], but they are limited in their ability to utilize unstructured data and thus may not achieve full efficiency. More recently, double/debiased machine learning [10] leverages flexible machine learning together with cross-fitting. However, existing flexible machine learning tools, such as random forests and boosting, are primarily limited to structured data and are not directly applicable to unstructured data [11, 22, 41]. In addition, the inclusion of many covariates may lead to biases in the asymptotic normal distribution and therefore renders standard normal-based inference invalid [9].

Another line of related work is prediction-powered inference (PPI), recently proposed by Angelopoulos *et al.* [3]. Originally developed for problems with missing labels, PPI leverages black-box predictions of unobserved outcomes to improve the efficiency of parameter estimation and has since been extended to randomized experiments [12, 40]. Subsequent refinements include the use of optimally tuned parameters [4], stratified tuning parameters [17], and cross-fitted machine learning predictions [58]. Unlike our approach, these methods do not always guarantee efficiency gains and may even suffer efficiency loss when auxiliary predictions are inaccurate. Additionally, they are not

designed to incorporate correlated predictions generated by LLMs using few-shot learning.

Our work is also motivated by recent advances in LLMs, which have demonstrated remarkable predictive performance across diverse tasks. Early models focused on natural language tasks, such as text classification, question answering, and summarization, using models like BERT [13] and GPT-2 [43]. More recent models, such as GPT-3, PaLM, and LLaMA, exhibit strong in-context learning abilities, encompassing zero-shot and few-shot prediction, that make them adaptable to new tasks with reduced supervision [6, 14, 50]. In scientific and biomedical domains, domain-adapted models such as BioGPT and Galactica leverage specialized corpora to improve relevance and accuracy [35, 48]. At the same time, multimodal LLMs such as CLIP, Flamingo, BLIP-2, LLaVA, and GPT-4V extend these capabilities beyond text, enabling joint reasoning over text, images, and audio [2, 29, 32, 38, 42]. However, most LLM-based prediction pipelines place little emphasis on uncertainty quantification or statistical guarantees, which are crucial for causal inference in high-stakes biomedical settings.

1.4 Problem setup: Causal parameters in randomized experiments

In this section, we introduce the notation used throughout the manuscript and then define the causal parameters of interest. Suppose we have a randomized experiment or a randomized controlled trial (RCT) consisting of n subjects. Following the Neyman-Rubin causal model, let $T \in \{1, \ldots, k\}$ denote the randomized treatment assignment, $Y(t)$ the potential outcome under treatment $t \in \{1, \ldots, k\}$, and Y the observed outcome.

For each subject in the experiment, we can observe pre-treatment covariate information. Let $X \in \mathcal{X} \subseteq \mathbb{R}^p$ represent structured pre-treatment covariates that are routinely adopted for classical statistical analysis in randomized experiments, and let $Z \in \mathcal{Z}$ denote any variables that may contain heavy missing or unstructured pre-treatment data that are traditionally challenging to incorporate. We note that CALM's efficiency gain does not depend on the presence of Z and its ability to conveniently handle unstructured data is simply an additional benefit. We denote the observed experimental data as $\{O_i\}_{i=1}^n$, where $O_i := (Y_i, T_i, X_i, Z_i)$.

Furthermore, in RCTs, treatments are often randomly assigned according to a known propensity score $e_t(X) := \mathbb{P}(T = t \mid X)$ based on the structured covariates X. This setup includes, as special cases, the completely randomized design where $e_t(X) \equiv e_t$ is constant across all units, as well as stratified randomized designs where $e_t(X)$ varies across strata defined by discretized X.

This manuscript aims to provide valid and efficient statistical inference for several causal parameters, including (i) mean effect of treatment $t \in \{1, \ldots, k\}$:

$$\mu_t = \mathbb{E}(Y(t));$$

(ii) the average treatment effect (ATE) between treatment arms t and t',

$$\tau_{t,t'} = \mathbb{E}(Y(t) - Y(t')), \quad t \neq t', \quad t, t' \in \{1, \ldots, k\};$$

(iii) the conditional average treatment effect (CATE) given covariates $X = x$, for $x \in \mathcal{X}$:

$$\tau_{t,t'}(x) = \mathbb{E}(Y(t) - Y(t') \mid X = x),$$

which plays a central role in evaluating treatment effect heterogeneity.

2 CALM for estimating $\mathbb{E}(Y(t))$

In this section, we provide a detailed description along with simple heuristics of our proposed Causal Analysis leveraging Language Models (CALM) framework for $\mathbb{E}(Y(t))$, leveraging either zero-shot or few-shot learning with pre-trained LLMs. In section 3, we extend CALM to ATE and CATE estimation.

We first briefly overview and compare zero-shot and few-shot LLM-based predictions. Zero-shot prediction for subject i utilizes a prompt that solely includes serialized covariates (X_i, Z_i) and treatment t, without providing any additional samples (also referred to as examples in the computer science community) collected in the trial data. The zero-shot prediction thus relies entirely on the LLM's pre-trained knowledge and generalization ability. In contrast, few-shot prediction adopt a prompt that incorporates a few demonstrative examples, denoted as $\{(X_i, Z_i, Y_i)\}_{i \in \mathcal{S}}$ with $\mathcal{S}$ being a small subset of $[n] = \{1, 2 \ldots, n\}$. In this case, the LLM can adapt its prediction more closely to the pattern of the demonstrative examples.

2.1 CALM with zero-shot learning

We start with describing the step-by-step procedure for estimating $\mathbb{E}(Y(t))$ using CALM with zero-shot learning. Alongside the detailed steps, we also provide simple heuristics and implementation guidance for practical use.

Step 1 (*Sample splitting*) Randomly split the RCT data $\{O_i\}_{i=1}^n$ into two equally sized, non-overlapping folds $\mathcal{I}_1$ and $\mathcal{I}_2$.

More generally, one can also employ K-fold sample splitting; we use $K = 2$ here for illustration.

Step 2 (*Counterfactual prediction with LLM zero-shot learning*) Construct a prompt for a large language model (LLM) that instructs it to predict counterfactual outcome(s) based on contextual information describing data in the RCT, *without* including any examples from the dataset. With this zero-shot prompt, query LLM to predict counterfactual outcomes under treatment arm $t \in \{1, \ldots, k\}$ for each subject i based on *serialized* X_i and Z_i:

$$X_i, Z_i \xrightarrow[\texttt{LLM zero-shot}]{\texttt{serialization}} Y_i^\dagger(t), \quad \text{for } 1 \le i \le n,$$

by providing both structured X_i and unstructured Z_i covariate information.

In the above step, we construct a structured prompt that includes three components to enable zero-shot prediction from a pre-trained LLM. *First*, we provide an instruction that describes the task: predicting counterfactual outcomes under treatment t using structured (X) and unstructured (Z) covariates. To ensure the LLM correctly interprets the data, the instruction also includes natural language descriptions of input and output variables. *Second*, we provide the input data X_i and Z_i, where structured covariates are serialized into natural language, and unstructured textual data (e.g., free-text notes) are presented directly. Other modalities, such as image data, can be uploaded separately if supported by the model interface. *Third*, we specify the desired output by requesting predictions in a fixed format for counterfactual outcomes $Y_i^\dagger(t)$. We provide an example of our zero-shot prompt design in the Supplementary Materials.

As LLM-based counterfactual predictions may be noisy yet contain valuable information, we must calibrate between observed data and predicted counterfactuals, allowing us to selectively augment observed outcomes and enhance causal effect estimation efficiency. To achieve this, we introduce the following step:

Step 3 (*Heterogeneous calibration between observed outcome and LLM predictions***)** For each fold $\ell \in \{1, 2\}$, we estimate: (i) the conditional mean of the observed outcome under treatment t, $\mu_t(x) := \mathbb{E}(Y \mid T = t, X = x)$; (ii) the conditional mean of the LLM-predicted counterfactual outcome under treatment t, $\mu_t^\dagger(x) := \mathbb{E}\left(Y^\dagger(t) \mid X = x\right)$; and (iii) the heterogeneous calibration weight function,

$$\omega_t(x) := \frac{\mathbb{C}\text{ov}\left(Y, Y^\dagger(t) \mid T = t, X = x\right)}{\mathbb{V}\text{ar}\left(Y^\dagger(t) \mid T = t, X = x\right)}. \tag{1}$$

We denote the resulting estimates as $\widehat{\mu}_t^{\mathcal{I}_\ell}(\cdot)$, $\widehat{\mu}_t^{\dagger\mathcal{I}_\ell}(\cdot)$, and $\widehat{\omega}_t^{\mathcal{I}_\ell}(\cdot)$, respectively.

To provide details for Step 3, we clarify the definition of $\mu_t^\dagger(x)$, offer heuristic interpretations of $\omega_t(x)$, and outline their respective estimation strategies. To start with, $Y^\dagger(t)$ can be viewed as a function of structured covariates X and unstructured covariates Z, denoted by $f_{\theta,t}(X, Z)$. Here, θ contains the pre-trained parameters of the LLM used in Step 1, and $f_{\theta,t}(\cdot)$ represents the mapping implemented by the LLM. The conditional expectation $\mu_t^\dagger(x) := \mathbb{E}(Y^\dagger(t) \mid X = x)$ is therefore taken with respect to Z, conditional on X. A more detailed discussion is provided in Section 4.

Next, as LLM-based prediction can be noisy, the construction of the calibration weight $\omega_t(x)$ is the key to our method and is motivated by three reasons. First, while structured covariate X explains part of the variation in the potential outcome $Y(t)$, the remaining variation could potentially be explained from unstructured data Z that the LLM maps into $Y^\dagger(t)$. Second, because LLMs are pretrained on vast and diverse corpora, tabular trial data used in classical causal analysis may offer limited added signal. When serialized to text (as in our Step 2), an LLM can potentially uncover additional patterns linking outcomes and covariates, so our counterfactual predictions $Y^\dagger(t)$ can carry useful information about the trial outcome. Third, the predictive power of $Y^\dagger(t)$ may not be uniform across subjects: heterogeneity in subject features mean that its predictions may be more

informative for some subjects than for others. To incorporate all three reasons, $\omega_t(x)$ selectively extracts relevant information from the LLM-based predictions to enhance estimation efficiency and also allows such calibration to differ across subjects based on their characteristics. Whenever LLM based predictions are highly predictive of trial outcomes, a higher calibration weight is desired.

Lastly, the functions $\mu_t(x)$, $\mu_t^\dagger(x)$, and $\omega_t(x)$ can be estimated with flexible machine learning methods, following the double machine learning framework [10], using tools such as random forests, kernel regression, or neural networks.

Step 4 (*CALM estimator for* $\mathbb{E}(Y(t))$**)** Let (ℓ_1, ℓ_2) be any permutation of $(1, 2)$. We define for $i \in \mathcal{I}_{\ell_2}$:

$$\widehat{\varphi}_t\left(Y_i, T_i, X_i, Y_i^\dagger(t)\right) = \frac{\mathbf{1}\{T_i = t\}Y_i}{e_t\left(X_i\right)} + \left(1 - \frac{\mathbf{1}\{T_i = t\}}{e_t\left(X_i\right)}\right)\left(\widehat{\mu}_t^{\mathcal{I}_{\ell_1}}(X_i) + \underbrace{\widehat{\omega}_t^{\mathcal{I}_{\ell_1}}\left(X_i\right)\left(Y_i^\dagger(t) - \widehat{\mu}_t^{\dagger\mathcal{I}_{\ell_1}}(X_i)\right)}_{\text{Calibrated residuals with LLM}}\right). \tag{2}$$

Then, the fold-specific CALM estimator on fold ℓ_2 is:

$$\widehat{\mu}_{t,\texttt{CALM}}^{\mathcal{I}_{\ell_2}} = \frac{1}{|\mathcal{I}_{\ell_2}|}\sum_{i \in \mathcal{I}_{\ell_2}} \widehat{\varphi}_t\left(Y_i, T_i, X_i, Y_i^\dagger(t)\right).$$

Aggregating over all permutations, we define our proposed CALM estimator as:

$$\widehat{\mu}_{t,\texttt{CALM}} = \sum_{\ell=1}^{2} \frac{|\mathcal{I}_\ell|}{n}\widehat{\mu}_t^{\mathcal{I}_\ell}.$$

Step 4 constitutes the core of the CALM estimator. Here, we modify the classical augmented inverse propensity weighting (AIPW) method[1] by incorporating centered LLM-based predictions and calibration weights. Because the LLM-based predictions are explicitly centered, the calibrated residual terms introduce no additional bias and are introduced solely to enhance estimation efficiency. Furthermore, since the calibration weights $\omega_t(x)$ vary across covariates x, the information borrowed from the LLM-based predictions is adaptively tailored to subject characteristics, further improving efficiency. In the last step, we provide the CALM-based statistical inference procedure:

Step 5 (*CALM-based statistical inference***)** We estimate the variance of $\widehat{\mu}_{t,\texttt{CALM}}$ by:

$$\widehat{\mathsf{V}}_{t,\texttt{CALM}} = \frac{1}{n}\sum_{\ell=1}^{2}\sum_{i \in \mathcal{I}_\ell}\left[\widehat{\varphi}_t(Y_i, T_i, X_i, Y_i^\dagger(t)) - \widehat{\mu}_{t,\texttt{CALM}}\right]^2,$$

and construct the $(1-\alpha)$ confidence interval for $\widehat{\mu}_{t,\texttt{CALM}}$ as $\left[\widehat{\mu}_{t,\texttt{CALM}} \pm z_{1-\alpha/2}\sqrt{\widehat{\mathsf{V}}_{t,\texttt{CALM}}/n}\right]$.

[1]Recall that the AIPW influence function is given by

$$\widehat{\varphi}_{t,\texttt{AIPW}}(Y_i, T_i, X_i) = \frac{\mathbf{1}\{T_i = t\}Y_i}{e_t(X_i)} + \left(1 - \frac{\mathbf{1}\{T_i = t\}}{e_t(X_i)}\right)\widehat{\mu}_t^{\mathcal{I}_{\ell_1}}(X_i).$$

Rigorous theoretical analyses of $\widehat{\mu}_{t,\texttt{CALM}}$ and $\widehat{\mathsf{V}}_{t,\texttt{CALM}}$ are provided in Section 4.1, where we will also show that CALM has *improved* estimation efficiency compared to the AIPW estimator (Theorem 2). In addition to theoretically analyzing CALM's efficiency gains, we also provide a practical approach for empirically testing whether CALM offers efficiency improvements over AIPW. We formalize this idea in the following remark:

Remark 1 (Test of efficiency improvement of CALM over AIPW) *To evaluate the potential of efficiency gain of $\widehat{\mu}_{t,\texttt{CALM}}$ relative to the classical AIPW estimator, it is possible to test whether the conditional covariance is uniformly zero:*

$$\mathcal{H}_0 : \ \gamma_t(x) := \mathbb{C}\mathrm{ov}(Y, Y^\dagger(t) \mid T = t, X = x) = 0 \quad \forall\, x \in \mathcal{X}.$$

For concreteness, suppose $\gamma_t(x)$ is estimated via the kernel estimator:

$$\widehat{\gamma}_t(x) = \frac{\sum_{i:T_i=t} \kappa_i^x Y_i Y_i^\dagger(t)}{\sum_{i:T_i=t} \kappa_i^x} - \frac{\sum_{i:T_i=t} \kappa_i^x Y_i}{\sum_{i:T_i=t} \kappa_i^x} \cdot \frac{\sum_{i:T_i=t} \kappa_i^x Y_i^\dagger(t)}{\sum_{i:T_i=t} \kappa_i^x},$$

where $\kappa_i^x = K(h^{-1}(X_i - x))$ for some kernel function K and some bandwidth (sequence) h. We then construct the test statistic:

$$T = \sup_x |T(x)| = \sup_x \left| \widehat{\gamma}_t(x) / \widehat{\sigma}_t(x) \right|,$$

where the supremum can be taken on a very fine grid, and $\widehat{\sigma}_t(x)$ is the standard error of $\widehat{\gamma}_t(x)$. Large values of T provide evidence against $\mathcal{H}_0$, indicating that the CALM estimator achieves efficiency gains relative to AIPW. To compute critical values for a level-α test, let $G(x)$ be a centered Gaussian process with the same covariance structure with $T(x)$, and then compute the $1-\alpha$ quantile of $\sup_x |G(x)|$ via simulation. A detailed description of the implementation is provided in Section 2 of the Supplementary Materials.

2.2 CALM with few-shot learning

In this section, we extend zero-shot-based CALM to enable few-shot learning, which is known to yield more robust and accurate predictions, particularly in settings where domain-specific examples help reduce prediction uncertainty and improve generalization [6]. To compare, zero-shot learning, despite being convenient with overall good performance in general-purpose settings, relies entirely on the model's pretrained knowledge and lacks the flexibility to adapt to specific downstream prediction tasks. For example, its performance can degrade under significant distributional shifts between pretraining data and the observed experimental data, or when relevant pretraining data is absent. Few-shot learning, on the other hand, addresses this limitation by instructing LLM-based predictions on a small set of demonstrative examples embedded in the prompt, enabling adaptation to task-specific patterns and potentially improving LLM's predictive performance relative to the zero-shot setting.

However, two statistical challenges arise when integrating few-shot predicted counterfactuals into CALM. *First*, due to limitations in prompt length (a.k.a. LLM token size limit) and the model's tendency to forget earlier inputs, typically only a small number of demonstrative examples can be included in the prompt. Since an LLM relies on these few examples to conduct outcome predictions, the randomness in example selection and ordering of those examples introduces additional variability that must be accounted for for valid statistical inference. *Second*, the few-shot learning-based LLM counterfactual predictions are inherently correlated due to the inclusion of randomly selected examples, and thus violate the i.i.d. assumptions required for zero-shot CALM.

To address the challenges explained above, we propose a novel resampling-based few-shot CALM. The procedure consists of the following steps:

Step 1′ *(Three-way sample splitting)* Randomly split the RCT data $\{O_i\}_{i=1}^n$ into three equally sized, non-overlapping folds $\mathcal{I}_1$, $\mathcal{I}_2$ and $\mathcal{I}_3$.

As shall be made clear in our theoretical investigation, the three-way splitting introduced here is to partially mitigate the second challenge on correlated counterfactual predictions with few-shot learning. For illustration, let the data fold indices (ℓ_1, ℓ_2, ℓ_3) be $(1, 2, 3)$ in the first iteration, then cyclically rotate to $(2, 3, 1)$ and $(3, 1, 2)$ to repeat Steps 2 and 3. We next describe how Step 2 in Section 2.1 can be adapted to few-shot learning.

Step 2′ *(Robust few-shot counterfactual predictions with resampling-based aggregation)* For a small fixed positive integer m, let

$$\mathcal{S}^*(\mathcal{I}_{\ell_1}) := \big\{(X^*_{j,\ell_1}, Z^*_{j,\ell_1}, T^*_{j,\ell_1}, Y^*_{j,\ell_1}), j = 1, \ldots, m\big\}$$

be a sample randomly drawn without replacement from the arm t data in $\mathcal{I}_{\ell_1}$. Next, for subject $i \in \mathcal{I}_{\ell_2} \bigcup \mathcal{I}_{\ell_3}$, construct a few-shot learning prompt that instructs an LLM to predict counterfactual outcomes based on *serialized* X_i, Z_i, and the randomly selected subsample $\mathcal{S}^*(\mathcal{I}_{\ell_1})$. We then query an LLM to predict counterfactual outcomes under arm t with this few-shot prompt:

$$(X_i, Z_i), \mathcal{S}^*(\mathcal{I}_{\ell_1}) \xrightarrow[\texttt{LLM few-shot}]{\texttt{Serialization}} Y_i^\dagger\big(t; \mathcal{S}^*(\mathcal{I}_{\ell_1})\big), \quad \text{for } i \in \mathcal{I}_{\ell_2} \bigcup \mathcal{I}_{\ell_3}.$$

By repeating the above procedure B times with independently resampled subsets $\mathcal{S}^*_b(\mathcal{I}_{\ell_1})$, $b = 1, \ldots, B$, we obtain the aggregated counterfactual prediction $Y^\dagger_{i,\texttt{FS}}(t; \mathcal{I}_{\ell_1})$ for each $i \in \mathcal{I}_{\ell_2} \cup \mathcal{I}_{\ell_3}$:

$$Y^\dagger_{i,\texttt{FS}}(t; \mathcal{I}_{\ell_1}) \;=\; \frac{1}{B}\sum_{b=1}^{B} Y_i^\dagger\big(t; \mathcal{S}^*_b(\mathcal{I}_{\ell_1})\big), \quad \text{for } i \in \mathcal{I}_{\ell_2} \bigcup \mathcal{I}_{\ell_3}.$$

where $\mathcal{I}_{\ell_1}$ in $Y^\dagger_{i,\texttt{FS}}(t; \mathcal{I}_{\ell_1})$ emphasizes that the few-shot prediction is based on examples drawn from that set.

Step 2′ simultaneously tackles the two statistical challenges of integrating few-shot predictions into the CALM framework discussed above. For the first challenge, it repeatedly resamples small,

random sets of demonstrative examples and aggregates the resulting few-shot predictions. This averaging procedure effectively neutralizes the variability introduced by the random selection and ordering of demonstrative examples, which would otherwise persist even as the sample size grows. As a result, both the aggregated predictions and the causal effect estimator built upon them become largely insensitive to the specific choice of demonstrative examples, thereby delivering more robust performance.

To address the second challenge, Step 2′ explicitly separates the data used for selecting demonstrative examples ($\mathcal{I}_{\ell_1}$) from that used for querying ($\mathcal{I}_{\ell_2} \cup \mathcal{I}_{\ell_3}$). By resampling demonstration sets from $\mathcal{I}_{\ell_1}$, the aggregated few-shot predictions $\{Y^{\dagger}_{i,\texttt{FS}}(t;\mathcal{I}_{\ell_1})\}_{i\in\mathcal{I}_{\ell_2}\cup\mathcal{I}_{\ell_3}}$ approximate a U-statistic-like structure whose limit is a deterministic function of (X, Z), akin to the zero-shot case. Consequently, these predictions are asymptotically conditionally independent across subjects given $\mathcal{I}_{\ell_1}$. This construction restores the i.i.d. structure required for applying standard asymptotic theory, thereby enabling valid downstream inference for μ_t.

Step 3′ *(Heterogeneous calibration)* We estimate $\mu_t(\cdot)$, $\mu^{\dagger}_t(\cdot)$, and $\omega_t(\cdot)$, using data fold $\mathcal{I}_{\ell_2}$, where $Y^{\dagger}_i(t)$ is replaced by the aggregated predictions $Y^{\dagger}_{i,\texttt{FS}}(t;\mathcal{I}_{\ell_1})$ obtained in Step 2′. We denote the resulting estimates as $\widehat{\mu}^{\mathcal{I}_{\ell_2}}_{t,\texttt{FS}}(\cdot)$, $\widehat{\mu}^{\dagger\mathcal{I}_{\ell_2}}_{t,\texttt{FS}}(\cdot)$ and $\widehat{\omega}^{\mathcal{I}_{\ell_2}}_{t,\texttt{FS}}(\cdot)$, respectively.

Step 4′ *(CALM estimator for $\mathbb{E}(Y(t))$ with few-shot learning)* We define for $i \in \mathcal{I}_{\ell_3}$:

$$\widehat{\varphi}_{t,\texttt{FS}}\big(Y_i, T_i, X_i, Y^{\dagger}_i(t)\big) = \frac{\mathbf{1}\{T_i = t\}\, Y_i}{e_t(X_i)} + \left(1 - \frac{\mathbf{1}\{T_i = t\}}{e_t(X_i)}\right) \times \left(\widehat{\mu}^{\mathcal{I}_{\ell_2}}_{t,\texttt{FS}}(X_i) + \underbrace{\widehat{\omega}^{\mathcal{I}_{\ell_2}}_{t,\texttt{FS}}(X_i)\big(Y^{\dagger}_{i,\texttt{FS}}(t;\mathcal{I}_{\ell_1}) - \widehat{\mu}^{\dagger\mathcal{I}_{\ell_2}}_{t,\texttt{FS}}(X_i)\big)}_{\text{Calibrated residuals with LLM}}\right).$$

Then, the fold-specific CALM estimator on fold ℓ_3 is:

$$\widehat{\mu}^{\mathcal{I}_{\ell_3}}_{t,\texttt{CALM,FS}} = \frac{1}{|\mathcal{I}_{\ell_3}|} \sum_{i\in\mathcal{I}_{\ell_3}} \widehat{\varphi}_{t,\texttt{FS}}\left(Y_i, T_i, X_i, Y^{\dagger}_i(t)\right).$$

Finally, aggregating over the three fold permutations $(\ell_1, \ell_2, \ell_3) \in \{(1,2,3),(2,3,1),(3,1,2)\}$, the proposed CALM estimator under few-shot setting is defined as

$$\widehat{\mu}_{t,\texttt{CALM,FS}} = \sum_{\ell=1}^{3} \frac{|\mathcal{I}_\ell|}{n} \widehat{\mu}^{\mathcal{I}_\ell}_{t,\texttt{CALM,FS}}.$$

The last two steps naturally extend Steps 3 and 4 from the zero-shot setting in Section 2.1. As part of our theoretical investigation, we rely on U-statistic methods [51] to show that the aggregated few-shot predictor $Y^{\dagger}_{i,\texttt{FS}}$ no longer depends on the demonstrative examples in large samples, which is crucial for establishing the asymptotic properties of $\widehat{\mu}_{t,\texttt{CALM,FS}}$. This makes the framework particularly suitable for analyzing complex black-box predictors, such as LLM-based outcomes. A rigorous theoretical analysis of $\widehat{\mu}_{t,\texttt{CALM,FS}}$ is given in Section 4.2.

2.3 Estimation strategy of the heterogeneous calibration weights

In this section, we describe estimation strategies for the heterogeneous calibration weight function $\omega_t(x)$ adopted in Step 3 for zero-shot-based CALM. These strategies naturally extend to the few-shot setting in Step 3′, by replacing $Y_i^\dagger(t)$ with $Y_{i,\texttt{FS}}^\dagger(t;\mathcal{I}_{\ell_1})$. To estimate $\omega_t(x)$, we may first employ a pre-specified machine learning method to estimate (i) the conditional covariance between the observed and predicted counterfactuals, that is $\gamma_t(x) = \mathbb{C}\text{ov}\left(Y, Y^\dagger(t) \mid T = t, X = x\right)$, and (ii) the conditional variance of the predicted counterfactuals, that is $\nu_t(x) = \mathbb{V}\text{ar}\left(Y^\dagger(t) \mid T = t, X = x\right)$, using data fold $\mathcal{I}_\ell$. Denote the corresponding estimators as $\widehat{\gamma}_t^{\mathcal{I}_\ell}(x)$ and $\widehat{\nu}_t^{\mathcal{I}_\ell}(x)$. Then, a natural plug-in estimator for $\omega_t(x)$ is given by $\widehat{\omega}_t^{\mathcal{I}_\ell}(x) := \widehat{\gamma}_t^{\mathcal{I}_\ell}(x)/\widehat{\nu}_t^{\mathcal{I}_\ell}(x)$, for $k \in \{1, 2\}$.

Yet, in practice, the machine learning methods used to estimate the conditional means $\mu_t(x)$ and $\mu_t^\dagger(x)$ may rely on underlying modeling assumptions that deviate from the true data-generating mechanism, a phenomenon referred to as model misspecification. Such discrepancies can arise in parametric settings, where the assumed functional form (e.g., linearity) is incorrect, as well as in semi-/non-parametric settings with additional structural restrictions (such as piecewise linearity or additive separability). To address this potential model misspecification, we introduce a robust calibration strategy that accounts for heterogeneity in predictive power across subjects while guaranteeing efficiency gains relative to the classical AIPW estimator.

To formalize this idea, we "coarsen" the covariates X into a discrete representation $X^{\texttt{C}}$. Coarsening may be achieved, for instance, by discretizing continuous features, aggregating categories, or grouping baseline health indicators into clinically meaningful strata. For simplicity, we then define the propensity-score adjustment weight as $\lambda_t(x) := \left(\frac{1}{e_t(x)} - 1\right)$. Then, the robust heterogeneous calibration weight is estimated as

$$\widehat{\omega}_{t,\texttt{R}}^{\mathcal{I}_{\ell_1}}(x^{\texttt{C}}) = \frac{\sum_{i\in\mathcal{I}_{\ell_2}:T_i=t,X_i^{\texttt{C}}=x^{\texttt{C}}} \lambda_t(X_i)\left(Y_i - \widehat{\mu}_t^{\mathcal{I}_{\ell_1}}(X_i)\right)\left(Y_i^\dagger(t) - \widehat{\mu}_t^{\dagger\mathcal{I}_{\ell_1}}(X_i)\right)}{\sum_{i\in\mathcal{I}_{\ell_2}:T_i=t,X_i^{\texttt{C}}=x^{\texttt{C}}} \lambda_t(X_i)\left(Y_i^\dagger(t) - \widehat{\mu}_t^{\dagger\mathcal{I}_{\ell_1}}(X_i)\right)^2},$$

where (ℓ_1, ℓ_2) is a permutation of $(1, 2)$. Finally, to construct the CALM estimator, it suffices to replace $\widehat{\omega}_t^{\mathcal{I}_{\ell_1}}(X_i)$ in Equation 2 with $\widehat{\omega}_{t,\texttt{R}}^{\mathcal{I}_{\ell_1}}(X_i^{\texttt{C}})$, and we denote the resulting CALM estimator for μ_t as $\widehat{\mu}_{t,\texttt{R},\texttt{CALM}}$. Theoretical guarantees of robust efficiency gains are provided in Corollary 1 of Section 4.

3 CALM for estimating other causal parameters

In this section, we extend the CALM framework introduced in Section 2 for estimating $\mathbb{E}(Y(t))$ to accommodate other important causal parameters, including the average treatment effect (ATE) $\tau_{t,t'}$ and the conditional average treatment effect (CATE) $\tau_{t,t'}(x)$ between two treatment arms t and t'.

3.1 CALM for estimating ATE

We begin by describing the estimation of the ATE under the zero-shot setting; the extension to the few-shot setting follows directly from Section 2.2. Steps 1 and 2 remain the same as in Section 2.1, yielding predicted counterfactuals for both treatment arms $\{Y_i^\dagger(t), Y_i^\dagger(t')\}_{i=1}^n$.

The calibration step differs from that used for the mean effect in Section 2.1, since the predicted counterfactuals across treatment arms $Y^\dagger(t)$ and $Y^\dagger(t')$ are correlated. We define a new heterogeneous calibration weight as

$$\boldsymbol{\omega}_{\texttt{ATE}}(X) = \begin{pmatrix} \omega_{t,\texttt{ATE}}(X) \\ \omega_{t',\texttt{ATE}}(X) \end{pmatrix} = \Sigma_V^{-1}(X)\mathbb{C}\text{ov}(V, Z \mid X),$$

which is specifically tailored for ATE estimation. Here, the 2×2 matrix $\Sigma_V(X)$ denotes the conditional variance of the vector $V = \big(\sqrt{\frac{1-e_t(X)}{e_t(X)}}Y^\dagger(t), \sqrt{\frac{1-e_{t'}(X)}{e_{t'}(X)}}Y^\dagger(t')\big)'$ conditional on X, and the 2×1 vector $\mathbb{C}\text{ov}(V, Z \mid X)$ is the conditional covariance between V and $Z = \sqrt{\frac{1-e_t(X)}{e_t(X)}}\frac{\mathbf{1}\{T=t\}}{e_t(X)}Y + \sqrt{\frac{1-e_{t'}(X)}{e_{t'}(X)}}\frac{\mathbf{1}\{T=t'\}}{e_{t'}(X)}Y$ conditional on X. In the special case of a balanced design, where subjects are assigned equally across all k arms, the propensity scores further reduce to $e_t(X) \equiv 1/k$ for all $t \in \{1, \ldots, k\}$. In this setting, the propensity score adjustment terms vanish, and the heterogeneous calibration weights simplify accordingly.

Step 3″ *(Heterogeneous calibration for ATE)* For each fold $k \in \{1, 2\}$, we estimate $\mu_t(\cdot)$, $\mu_{t'}(\cdot)$, $\mu_t^\dagger(\cdot)$, $\mu_{t'}^\dagger(\cdot)$, and $\boldsymbol{\omega}_{\texttt{ATE}}(\cdot)$. We denote the resulting estimates as $\widehat{\mu}_t^{\mathcal{I}_\ell}(\cdot)$, $\widehat{\mu}_{t'}^{\mathcal{I}_\ell}(\cdot)$, $\widehat{\mu}_t^{\dagger\mathcal{I}_\ell}(\cdot)$, $\widehat{\mu}_{t'}^{\dagger\mathcal{I}_\ell}(\cdot)$ and $\widehat{\boldsymbol{\omega}}_{\texttt{ATE}}^{\mathcal{I}_\ell}(\cdot)$, respectively.

The CALM estimator for the average treatment effect is then defined as

$$\widehat{\tau}_{t,t',\texttt{CALM}} = \widehat{\mu}_{t,\texttt{CALM},\texttt{ATE}} - \widehat{\mu}_{t',\texttt{CALM},\texttt{ATE}},$$

where $\widehat{\varphi}_{t,\texttt{CALM},\texttt{ATE}}(Y_i, T_i, X_i, Y_i^\dagger(t))$ and $\widehat{\varphi}_{t',\texttt{CALM},\texttt{ATE}}(Y_i, T_i, X_i, Y_i^\dagger(t'))$, together with the mean potential outcome estimators $\widehat{\mu}_{t,\texttt{CALM},\texttt{ATE}}$ and $\widehat{\mu}_{t',\texttt{CALM},\texttt{ATE}}$, are computed following Step 4 of Section 2.1. The only modification is that the original calibration weights $\widehat{\omega}_t(\cdot)$ and $\widehat{\omega}_{t'}(\cdot)$ are replaced by their ATE-specific counterparts $\widehat{\omega}_{t,\texttt{ATE}}(\cdot)$ and $\widehat{\omega}_{t',\texttt{ATE}}(\cdot)$. The theoretical properties of this estimator, together with its comparison to the corresponding AIPW-based estimator, are presented in Corollary 2 of Section 4.3. Finally, we can estimate the variance of $\widehat{\tau}_{t,t',\texttt{CALM}}$ by:

$$\widehat{\mathbb{V}}_{t,t',\texttt{CALM}} = \frac{1}{n}\sum_{\ell=1}^{2}\sum_{i\in\mathcal{I}_\ell}\Big[\widehat{\varphi}_{t,\texttt{CALM},\texttt{ATE}}(Y_i, T_i, X_i, Y_i^\dagger(t)) - \widehat{\varphi}_{t',\texttt{CALM},\texttt{ATE}}(Y_i, T_i, X_i, Y_i^\dagger(t')) - \widehat{\tau}_{t,t',\texttt{CALM}}\Big]^2.$$

3.2 CALM for estimating CATE

Building on the ATE estimation in Section 3.1, we now describe estimation of the CATE $\tau_{t,t'}(x)$. A standard AIPW-based CATE estimator regresses the influence function estimators $\widehat{\varphi}_{t,\texttt{AIPW}}(Y_i, T_i, X_i)$

and $\widehat{\varphi}_{t',\texttt{AIPW}}(Y_i, T_i, X_i)$ on the covariates X [25] (or on a subset of X; see 1). For concreteness, suppose the regression is implemented using a generic linear smoother. Then the AIPW-based CATE estimator can be written as

$$\widehat{\tau}_{t,t',\texttt{AIPW}}(x) = \sum_{i=1}^{n} w_i(x; \mathbf{X}) \left\{ \widehat{\varphi}_{t,\texttt{AIPW}}(Y_i, T_i, X_i) - \widehat{\varphi}_{t',\texttt{AIPW}}(Y_i, T_i, X_i) \right\},$$

where $w_i(x; \mathbf{X})$ are regression weights, such as those from kernel or spline smoothing.

Within our CALM framework, the AIPW-based CATE estimator can be further improved. After obtaining the influence function estimators $\widehat{\varphi}_{t,\texttt{CALM},\texttt{ATE}}(Y_i, T_i, X_i, Y_i^\dagger(t))$ and $\widehat{\varphi}_{t',\texttt{CALM},\texttt{ATE}}(Y_i, T_i, X_i, Y_i^\dagger(t'))$ from Section 3.1, we regress them on X to obtain

$$\widehat{\tau}_{t,t',\texttt{CALM}}(x) = \sum_{\ell=1}^{2} \sum_{i \in \mathcal{I}_\ell} w_i(x; \mathbf{X}) \left\{ \widehat{\varphi}_{t,\texttt{CALM},\texttt{ATE}}(Y_i, T_i, X_i, Y_i^\dagger(t)) - \widehat{\varphi}_{t',\texttt{CALM},\texttt{ATE}}(Y_i, T_i, X_i, Y_i^\dagger(t')) \right\}.$$

A natural variance estimator is

$$\widehat{\mathbb{V}}_{t,t',\texttt{CALM}}(x) = \sum_{\ell=1}^{2} \sum_{i \in \mathcal{I}_\ell} w_i(x; \mathbf{X})^2 \left\{ \widehat{\varphi}_{t,\texttt{CALM},\texttt{ATE}}(Y_i, T_i, X_i, Y_i^\dagger(t)) - \widehat{\varphi}_{t',\texttt{CALM},\texttt{ATE}}(Y_i, T_i, X_i, Y_i^\dagger(t')) - \widehat{\tau}_{t,t',\texttt{CALM}}(x) \right\}^2.$$

In Corollary 3 of Section 4.3, we establish the asymptotic properties of $\widehat{\tau}_{t,t',\texttt{CALM}}(x)$ and compare it with $\widehat{\tau}_{t,t',\texttt{AIPW}}(x)$, focusing on the case where the linear smoother is chosen to be a kernel regression estimator.

4 Theoretical investigations

In this section, we examine the theoretical properties of the CALM methods, described in Sections 2 and 3, that integrate zero-shot and few-shot LLM predictions for enhanced treatment effect estimation and statistical inference. We first lay out the main assumptions, and then establish the asymptotic properties of the CALM estimator for various causal parameters under zero-shot or few-shot learning settings, including consistency, asymptotic normality, and valid statistical inference.

4.1 Theoretical properties of CALM with zero-shot learning

We first study the CALM framework using zero-shot LLM predictions. Theoretical results for our method in the few-shot setting will be provided in the next subsection. Specifically, we show in Theorem 1 that the CALM estimator is both consistent and asymptotically normally distributed. As part of this analysis, we also carefully characterize the asymptotic efficiency gain compared to the AIPW estimator. As an extension of the main theorem, we discuss the implications of misspecifying the nuisance functions due to the discretization of the covariate space. Before presenting the

technical details, though, we introduce the following two assumptions, which apply to both zero-shot and few-shot learning settings.

Assumption 1 (Treatment assignment) *(i) Treatment assignment $T \in \{1, \ldots, k\}$ is independent of the potential outcomes and the unstructured data, conditional on the structured covariates; that is, $T \perp\!\!\!\perp \{Y(1), \ldots, Y(k), Z\} \mid X$. (ii) The propensity score $e_t(x) = \mathbb{P}(T = t|X = x)$ is bounded away from 0 and 1 for all $t \in \{1 \ldots, k\}$.*

This first assumption is standard in the analysis of randomized experiments, which requires that treatment assignment is independent of $Y(t)$ and Z once conditioning on a set of pre-determined covariates. Typically referred to as unconfoundedness, the validity of this assumption follows from the experimental design. The next assumption is concerned with the sampling scheme, which holds in typical experimental settings where participants are recruited and assigned to treatment independently.

Assumption 2 (Random sampling) *The observed data $\{(Y_i, T_i, X_i, Z_i)\}_{i=1}^n$ are independently and identically distributed from an unknown distribution P with common support.*

We now introduce a high-level assumption stating that the estimated nuisance functions, $\widehat{\mu}_t(\cdot)$, $\widehat{\mu}_t^\dagger(\cdot)$, and $\widehat{\omega}_t(\cdot)$, are consistent. Depending on the specific form of the estimators, such as k-nearest neighbors, kernel- or series-based methods, regression trees, or random forests, more primitive conditions are available in the literature. Such high-level consistency assumptions are commonly employed in the analysis of semiparametric estimators with Neyman-orthogonal estimating equations. Together with sample splitting in Step 1 and cross-fitting in Step 3, they help facilitate the establishment of consistency and asymptotic normality by relaxing the Donsker conditions [51] which may restrict the complexity of the nuisance function classes, thereby enabling the use of flexible nonparametric and machine learning methods [10]. Let $\|\cdot\|_{L_2}$ denote the L_2-norm with respect to the distribution of X.

Assumption 3 (Cross-fitted nuisance function estimates under zero-shot setting) *The cross-fitted function estimates obtained in Step 3 are consistent: $\|\widehat{\mu}_t^{\mathcal{I}_\ell} - \mu_t\|_{L_2}$, $\|\widehat{\mu}_t^{\dagger\mathcal{I}_\ell} - \mu_t^\dagger\|_{L_2}$ and $\|\widehat{\omega}_t^{\mathcal{I}_\ell} - \omega_t\|_{L_2} \xrightarrow{p} 0$ for data folds $\ell \in \{1, 2\}$ and treatment $t \in \{1, \ldots, k\}$.*

We are now ready to state our first main theoretical result: the CALM estimator of μ_t is both consistent and asymptotically normally distributed. The theorem also characterizes the estimator's asymptotic variance. Thanks to the incorporation of zero-shot LLM predictions, our method delivers efficiency gains over the classical augmented inverse probability weighting approach by incorporating unstructured data.

Theorem 1 (Asymptotic properties of CALM with zero-shot learning) *Under Assumption 1–3, we have $\widehat{\mu}_{t,\texttt{CALM}} \xrightarrow{p} \mu_t$, and $\sqrt{n}\left(\widehat{\mu}_{t,\texttt{CALM}} - \mu_t\right) \xrightarrow{d} \mathcal{N}\left(0, \mathsf{V}_{t,\texttt{CALM}}\right)$, where:*

$$\mathsf{V}_{t,\texttt{CALM}} = \mathbb{E}\left((\mu_t(X) - \mu_t)^2 + \frac{\mathbb{V}\mathrm{ar}(Y(t) \mid X)}{e_t(X)} \left(1 - (1 - e_t(X))\rho_t^2(X)\right) \right)$$

$$= \mathsf{V}_{t,\texttt{AIPW}} - \mathbb{E}\left(\frac{\mathbb{V}\mathrm{ar}(Y(t) \mid X)}{e_t(X)}(1 - e_t(X))\rho_t^2(X)\right)$$

and $\rho_t(x) = \mathbb{C}\mathrm{orr}(Y(t), Y^\dagger(t) \mid X = x)$.

From the asymptotic variance expression $\mathsf{V}_{t,\texttt{CALM}}$ above, it is clear that CALM yields strictly smaller asymptotic variance, and thus significant efficiency gains, whenever the LLM-based prediction $Y^\dagger(t)$ is correlated with the true potential outcome $Y(t)$ (i.e., when $\rho_t(x)$, or equivalently $\omega_t(x)$, is nonzero). In our empirical analysis, as illustrated in Figure 5, we observe that the calibration weights indeed deviate substantially from zero, indicating efficiency improvements from CALM. Moreover, we gently argue that it is natural to expect a nonzero correlation term from two aspects. First, when the LLM possesses prior knowledge about the intervention applied in the experiments, this correlation reflects the integration of additional relevant domain knowledge into LLM counterfactual predictions. Second, LLM-generated counterfactual predictions inherently incorporate information encoded in unstructured data Z, capturing variation in $Y(t)$ that structured covariates X alone fail to explain. Such unstructured information is typically challenging to exploit effectively with traditional causal inference methods, further highlighting the advantage offered by CALM.

Next, we examine to what extent the conclusions of Theorem 1 continue to hold without Assumption 3. Due to the Neyman orthogonality of our estimating equation, the CALM estimator remains consistent and asymptotically normally distributed. However, the asymptotic variance will take a different form. As expected, it is generally not possible to claim efficiency gains over the AIPW estimator when the nuisance functions used in our method are misspecified. Nevertheless, as we show in the remainder of this section, the CALM estimator remains (weakly) more efficient as long as the AIPW estimator uses the same misspecified models for the nuisance conditional expectations, and the heterogeneous calibration weights are specified in the robust form with a discrete representation of X as described in Section 2.3. We now impose a high-level condition on potentially misspecified conditional mean estimators, analogous to those considered in [24].

Assumption 4 (Misspecified cross-fitted conditional mean estimators under zero-shot setting) *There exists functions* $\widetilde{\mu}_t$ *and* $\widetilde{\mu}_t^\dagger$ *of* X*, such that the cross-fitted conditional mean estimators obtained in Step 3 satisfy* $\|\widehat{\mu}_t^{\mathcal{I}_\ell} - \widetilde{\mu}_t\|_{L_2}$ *and* $\|\widehat{\mu}_t^{\dagger\mathcal{I}_\ell} - \widetilde{\mu}_t^\dagger\|_{L_2} \xrightarrow{p} 0$ *for data folds* $l \in \{1, 2\}$ *and treatment* $t \in \{1, \ldots, k\}$.

For convenience, define the weighted conditional mean squared deviations

$$\sigma_t^2(x^{\texttt{C}}) := \mathbb{E}\big(\lambda_t(X)\{Y(t) - \widetilde{\mu}_t(X)\}^2 \mid X^{\texttt{C}} = x^{\texttt{C}}\big), \quad \sigma_t^{\dagger 2}(x^{\texttt{C}}) := \mathbb{E}\big(\lambda_t(X)\{Y^\dagger(t) - \widetilde{\mu}_t^\dagger(X)\}^2 \mid X^{\texttt{C}} = x^{\texttt{C}}\big),$$

where $\lambda_t(x) = \left(\frac{1}{e_t(x)} - 1\right)$ is the propensity-score adjustment weight, and $X^{\texttt{C}}$ denotes the discretized covariates, described in Section 2.3. Under the stated assumptions, the asymptotic distribution of the CALM estimator under conditional mean misspecification is given below.

Corollary 1 (Asymptotic properties of CALM with misspecified conditional means) *Under Assumptions 1–2 and 4, we have* $\widehat{\mu}_{t,\texttt{R,CALM}} \xrightarrow{p} \mu_t$*, and* $\sqrt{n}\big(\widehat{\mu}_{t,\texttt{R,CALM}} - \mu_t\big) \xrightarrow{d} \mathcal{N}\Big(0, \widetilde{\mathsf{V}}_{t,\texttt{CALM}}\Big)$,

where

$$\begin{aligned}\widetilde{\mathsf{V}}_{t,\texttt{CALM}} &= \mathbb{E}\big((\mu_t(X)-\mu_t)^2 + \mathbb{V}\mathrm{ar}(Y(t)\mid X) + \sigma_t^2(X^{\texttt{C}})\{1-\widetilde{\rho}_t^2(X^{\texttt{C}})\}\big)\\ &= \widetilde{\mathsf{V}}_{t,\texttt{AIPW}} - \mathbb{E}\big(\sigma_t^2(X^{\texttt{C}})\widetilde{\rho}_t^2(X^{\texttt{C}})\big)\,,\end{aligned}$$

with $\widetilde{\mathsf{V}}_{t,\texttt{AIPW}}$ *denoting the asymptotic variance of the AIPW estimator based on the same misspecified models for nuisance function estimation, and*

$$\widetilde{\rho}_t(x^{\texttt{C}}) = \frac{\mathbb{E}\Big(\lambda_t(X)\{Y(t)-\widetilde{\mu}_t(X)\}\{Y^\dagger(t)-\widetilde{\mu}_t^\dagger(X)\}\,\Big|\,X^{\texttt{C}}=x^{\texttt{C}}\Big)}{\sqrt{\sigma_t^2(x^{\texttt{C}})\,\sigma_t^{\dagger 2}(x^{\texttt{C}})}}.$$

From the asymptotic variance expression above, we see that even under misspecified models for $\mu_t(\cdot)$ and $\mu_t^\dagger(\cdot)$, CALM with robust heterogeneous calibration strategy detailed in Section 2.3 remains more efficient than AIPW, provided $\widetilde{\rho}_t(x^{\texttt{C}})(\cdot)$ is not identically zero. This efficiency gains derive not only from leveraging embedded domain knowledge and unstructured data Z via the LLM predictions (as described in Theorem 1), but also from capturing complex nonlinear associations between the potential outcomes $Y(t)$ and structured covariates X, which standard working models fail to account for.

4.2 Theoretical properties of CALM with few-shot learning

In this section, we study the theoretical properties of the few-shot-based CALM estimator. Recall that a unique feature of our method is that it averages over LLM-based predictions generated using small demonstrative examples, which are randomly sampled from a data fold. While this averaging helps reduce the randomness of the LLM-predicted potential outcomes and eliminates dependence on the ordering of the examples, it also complicates the theoretical analysis. To facilitate the discussion, we first introduce a notation that can be viewed as a large-sample analogue of the few-shot predicted outcome $Y_{i,\texttt{FS}}^\dagger(t,\mathcal{I})$. Let $\mathcal{S}=(O_1,O_2,\ldots,O_m)$ denote a random sample from the distribution P (see Assumption 2), and take some covariates $(X,Z)\perp\!\!\!\perp\mathcal{S}$. In line with Step 2′, let $Y^\dagger(t;\mathcal{S})$ denote the few-shot prediction based on (X,Z) and the sample $\mathcal{S}$. We then define

$$\bar{f}_{\theta,\texttt{FS}}((x,z),t) := \mathbb{E}(Y^\dagger(t;\mathcal{S})\mid X=x, Z=z),$$

where, as the notation suggests, the expectation is taken with respect to the random sample $\mathcal{S}$. The subscript θ represents the fixed parameters of the LLM. We then define the nuisance functions $\mu_{t,\texttt{FS}}^\dagger(\cdot)$ and $\omega_{t,\texttt{FS}}(\cdot)$ by replacing $Y^\dagger(t)$ in Step 3 with $\bar{f}_{\theta,\texttt{FS}}((X,Z),t)$; that is, $\mu_{t,\texttt{FS}}^\dagger(x) = \mathbb{E}(\bar{f}_{\theta,\texttt{FS}}((X,Z),t)\mid X=x)$ and

$$\omega_{t,\texttt{FS}}(x) = \mathbb{C}\mathrm{ov}\left(Y(t),\bar{f}_{\theta,\texttt{FS}}((X,Z),t)\mid X=x\right)/\mathbb{V}\mathrm{ar}\left(\bar{f}_{\theta,\texttt{FS}}((X,Z),t)\mid X=x\right).$$

The next assumption is the analogue of Assumption 3 in the few-shot setting, which requires

that the estimated nuisance functions to be consistent.

Assumption 5 (Cross-fitted nuisance function estimates under few-shot setting) *The cross-fitted function estimates obtained in Step 3′ are consistent:* $\|\widehat{\mu}_t^{\mathcal{I}_\ell} - \mu_t\|_{L_2}$, $\|\widehat{\mu}_{t,\texttt{FS}}^{\dagger\mathcal{I}_\ell} - \mu_{t,\texttt{FS}}^{\dagger}\|_{L_2}$ *and* $\|\widehat{\omega}_{t,\texttt{FS}}^{\mathcal{I}_\ell} - \omega_{t,\texttt{FS}}\|_{L_2} \xrightarrow{p} 0$ *for data folds* $\ell \in \{1,2,3\}$ *and treatment* $t \in \{1,\ldots,k\}$.

We are now ready to state the main result of this subsection: the CALM estimator, which incorporates few-shot predicted potential outcomes using an LLM, is consistent, asymptotically normally distributed, and achieves a smaller asymptotic variance than the standard AIPW estimator whenever the LLM-based predictions are correlated with the true potential outcomes.

Theorem 2 (Asymptotic properties of CALM with few-shot learning) *Under Assumptions 1-2 and 5, and the number of resamples in Step 2′ satisfies* $B \to \infty$, *we have* $\widehat{\mu}_{t,\texttt{CALM,FS}} \xrightarrow{p} \mu_t$, *and* $\sqrt{n}\left(\widehat{\mu}_{t,\texttt{CALM,FS}} - \mu_t\right) \xrightarrow{d} \mathcal{N}\left(0, \mathsf{V}_{t,\texttt{CALM,FS}}\right)$, *where:*

$$\begin{aligned}\mathsf{V}_{t,\texttt{CALM,FS}} &= \mathbb{E}\left((\mu_t(X) - \mu_t)^2 + \frac{\mathbb{V}\mathrm{ar}(Y(t) \mid X)}{e_t(X)}\left(1 - (1 - e_t(X))\rho_{t,\texttt{FS}}^2(X)\right)\right) \\ &= \mathsf{V}_{t,\texttt{AIPW}} - \mathbb{E}\left(\frac{\mathbb{V}\mathrm{ar}(Y(t) \mid X)}{e_t(X)}(1 - e_t(X))\rho_{t,\texttt{FS}}^2(X)\right)\end{aligned}$$

and $\rho_{t,\texttt{FS}}(x) = \mathbb{C}\mathrm{orr}(Y(t), \bar{f}_{\theta,\texttt{FS}}((X,Z),t) \mid X = x)$.

The above theorem establishes that the CALM estimator remains consistent and asymptotically normal without requiring additional strong assumptions when coupled with the resampling-based few-shot strategy introduced in Section 2.2. Its asymptotic variance, $\mathsf{V}_{t,\texttt{CALM,FS}}$, has the same structure as $\mathsf{V}_{t,\texttt{CALM}}$ in Theorem 1, except that the zero-shot predictions are replaced by few-shot predictions and the randomness from the demonstrative sample $\mathcal{S}$ is averaged out.

Since obtaining the few-shot CALM estimator $\widehat{\mu}_{t,\texttt{CALM,FS}}$ involves a more elaborate procedure and greater computational resources than constructing the zero-shot estimator $\widehat{\mu}_{t,\texttt{CALM}}$, it is natural to ask whether the few-shot estimator is asymptotically more efficient. Based on the structure of the asymptotic variance, $\widehat{\mu}_{t,\texttt{CALM,FS}}$ achieves higher efficiency whenever the few-shot prediction aligns more closely with the true potential outcome than the zero-shot prediction, that is, when $\rho_{t,\texttt{FS}}^2(x) > \rho_t^2(x)$.

Empirically, literature has documented that few-shot learning outperforms zero-shot learning when the demonstrative sample is informative or when the domain is specialized, while zero-shot learning can remain competitive in tasks closely aligned with the model's pretraining [6, 28, 47, 54]. In line with our empirical evidence demonstrated in Sections 5.3 and 6, the CALM estimator with few-shot learning is consistently more efficient than its zero-shot counterpart across various settings, and its standard deviation decreases as the size of the demonstrative sample grows.

4.3 Theoretical properties of CALM for other causal parameters

In this section, we briefly examine the key theoretical properties of the CALM framework for estimating the ATE and CATE under the zero-shot setting, as introduced in Section 3. We first

state the properties of $\widehat{\tau}_{t,t',\texttt{CALM}}$ in the next corollary.

Corollary 2 (Asymptotic properties of CALM for the ATE) *Suppose Assumptions 1–3 hold, with the calibration weights $\omega_t(\cdot)$ in Assumption 3 replaced by their ATE-specific counterparts $\omega_{t,\texttt{ATE}}(\cdot)$. Further assume that the conditional covariance matrix $\Sigma_V(x)$ defined in Section 3.1 satisfies $\Sigma_V(x) \succ 0$ for all $x \in \mathcal{X}$. Then, using the definitions of V, Z, $\Sigma_V(X)$, and $\mathbb{C}\text{ov}(V, Z \mid X)$ from Section 3.1, we have $\widehat{\tau}_{t,t',\texttt{CALM}} \xrightarrow{p} \tau_{t,t'}$, and $\sqrt{n}\big(\widehat{\tau}_{t,t',\texttt{CALM}} - \tau_{t,t'}\big) \xrightarrow{d} \mathcal{N}\big(0, \mathsf{V}_{t,t',\texttt{CALM}}\big)$, where*

$$\begin{aligned}\mathsf{V}_{t,t',\texttt{CALM}} = \mathbb{E}\Big((\tau_{t,t'}(X) - \tau_{t,t'})^2 + \frac{\mathbb{V}\text{ar}(Y(t) \mid X)}{e_t(X)} + \frac{\mathbb{V}\text{ar}(Y(t') \mid X)}{e_{t'}(X)} \\ - \mathbb{C}\text{ov}(V, Z \mid X)'\Sigma_V^{-1}(X)\mathbb{C}\text{ov}(V, Z \mid X)\Big) \\ = \mathsf{V}_{t,t',\texttt{AIPW}} - \mathbb{E}\big(\mathbb{C}\text{ov}(V, Z \mid X)'\Sigma_V^{-1}(X)\mathbb{C}\text{ov}(V, Z \mid X)\big)\,.\end{aligned}$$

The asymptotic variance result again shows that $\mathsf{V}_{t,t',\texttt{CALM}}$ is smaller than the $\mathsf{V}_{t,t',\texttt{AIPW}}$ whenever the conditional covariance matrix $\Sigma_V(x)$ is positive definite for some x.

We next establish the theoretical properties of the CALM estimator for the CATE. For illustration, we focus on kernel smoothing as the linear smoother described in Section 3.2, without loss of generality. Analogous asymptotic results can also be obtained for other nonparametric regression methods, such as local polynomial regression or K-nearest neighbors.

Corollary 3 (Asymptotic properties of CALM for the CATE with kernel smoothing) *Suppose the assumptions of Corollary 2 hold. Further assume that the linear smoother weights $w_i(x; \mathbf{X})$ in Section 3.2 are generated from kernel weights $\frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^n K\left(\frac{X_j - x}{h}\right)}$ with bandwidth h satisfying $h \to 0$, $nh^p \to \infty$, and $nh^{p+4} \to 0$ for covariate dimension p. Then, under standard conditions for kernel regression [30, 53], for any fixed $x \in \mathcal{X}$, we have $\widehat{\tau}_{t,t',\texttt{CALM}}(x) \xrightarrow{p} \tau_{t,t'}(x)$, and*

$$\sqrt{nh^p}\left(\widehat{\tau}_{t,t',\texttt{CALM}}(x) - \tau_{t,t'}(x)\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\|K\|_2^2 \mathsf{V}_{t,t',\texttt{CALM}}(x)}{f_X(x)}\right),$$

where $f_X(x)$ is the density of X at x, $\|K\|_2 := \left(\int K(u)^2 du\right)^{1/2}$, and

$$\mathsf{V}_{t,t',\texttt{CALM}}(x) = \frac{\mathbb{V}\text{ar}(Y(t) \mid X = x)}{e_t(x)} + \frac{\mathbb{V}\text{ar}(Y(t') \mid X = x)}{e_{t'}(x)} - \mathbb{C}\text{ov}(V, Z \mid X = x)'\Sigma_V^{-1}(x)\mathbb{C}\text{ov}(V, Z \mid X = x).$$

Compared with the AIPW-type CATE estimator using the same kernel weights, the CALM estimator for CATE achieves asymptotic variance reduction. Analogous extensions of Corollary 1 and Theorem 2 to both the ATE and CATE follow directly and are omitted for brevity.

5 Simulation studies

In this section, we evaluate the performance of CALM based on synthetic data calibrated from the BRIGHTEN study [41], which is an RCT with both structured and unstructured data. Before de-

tailing our complete simulation design and results, we briefly summarize the key findings from our simulation studies as follows: First, CALM-based methods consistently exhibit higher estimation efficiency compared to AIPW-based methods, while maintaining statistical inferential validity (Figure 3). Second, when AIPW is augmented with few-shot-learning-based counterfactual predictions as covariates, it shows substantial bias and leads to coverage probabilities declining as the number of few-shot examples grows, leading to systematic under-coverage (Figures 3D). Third, CALM's efficiency gain is present regardless of whether unstructured covariates are incorporated, suggesting that the improvement does not arise simply from adding extra unstructured information but from intrinsic information gain encoded in the pre-trained LLMs (Figure 4). Lastly, the heterogeneous calibration weighting mechanism in CALM selectively leverages LLM-based predictions where they align most closely with observed outcomes. This adaptive borrowing mechanism across covariate strata results in substantial variance reduction compared to other benchmark methods (Figure 5).

5.1 Overview of BRIGHTEN study

The BRIGHTEN (Bridging Research Innovations for Greater Health in Technology, Emotion, and Neuroscience) study [41] comprises fully remote, smartphone-based randomized controlled trials conducted in the United States between 2016 and 2018, with the goal of evaluating the feasibility and effectiveness of delivering mental health care at scale through mobile applications. Recruitment targeted adults ($\geq$18 years old) exhibiting clinically significant depressive symptoms, operationalized as a Patient Health Questionnaire (PHQ-9) score ≥ 5 or a score ≥ 2 on the PHQ functional-impairment item. Eligible participants were required to have an iPhone or Android device. Of the 7,850 individuals screened, 2,193 consented and enrolled. Participants were then randomized to one of the app-delivered interventions and followed for 12 weeks.

The BRIGHTEN study provides a unique data structure that combines both structured covariates and rich unstructured textual baseline surveys, making it particularly suitable for evaluating CALM. Figure 1(A) summarizes the trial's collected data, where we observe that unstructured data are collected at the initiation of the study. Figure 1(B) shows that baseline free-text responses on motivation for app use and follow-up app-satisfaction comments exhibit the highest cosine similarity with discretized PHQ-9 outcomes (trial primary outcome), surpassing demographics and app-usage metrics. Figure 1(B) demonstrates the potential of the unstructured data in the BRIGHTEN study as valuable "prognostic factors" (pre-treatment variables that are predictive of the outcome).

To set the stage for our simulation design, which involves generating synthetic data resembling the BRIGHTEN study, we clarify key notations and their interpretation. We define a binary treatment variable, T_i, indicating assignment either to the intervention group receiving internet-based Problem-Solving Therapy (iPST), denoted as $T_i = 1$, or to a control group assigned to alternative interventions, denoted as $T_i = 0$. We include structured covariates X_i such as demographic and clinical characteristics: sex, education, employment status, marital status, race, and age, etc. The unstructured textual covariates Z_i capture participants' self-reported reasons for enrollment and their satisfaction with the mobile app. The outcome of interest, Y_i, is the PHQ-9 score, a vali-

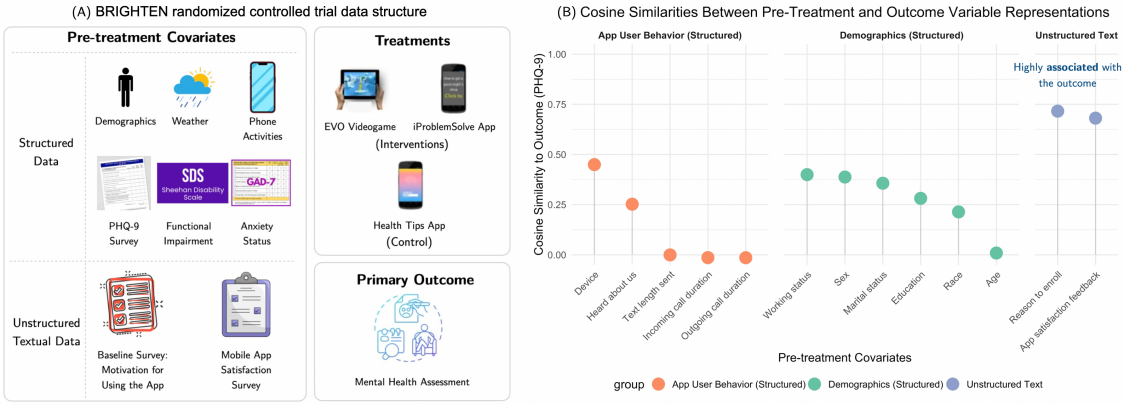


Figure 1: (A) BRIGHTEN study for depression management data structure. (B) Cosine similarity between different pre-treatment covariates and outcomes in BRIGHTEN shows that unstructured data are highly predictive of the primary outcome.

dated measure of depressive symptom severity with higher values indicating more severe depression symptoms [27].

5.2 Simulating BRIGTEN study with structured/unstructured data

In this section, we outline our approach for generating a synthetic superpopulation that closely resembles the BRIGHTEN study population (Figure 2). We provide these details because generating a synthetic population that incorporates both structured and unstructured data from the BRIGHTEN study presents several distinctive challenges. First, it is challenging to capture the relationship between numerical outcomes and covariates when the covariate set comprises both structured covariates (numerical and categorical) and unstructured covariates (free-text survey responses). Second, reproducing the joint covariate distribution requires preserving dependencies between structured and unstructured covariates, which necessitates embedding textual responses into numerical representations that can be modeled alongside structured covariates; however, this embedding step introduces high-dimensionality concerns, particularly given the limited sample size of the real data. Third, because unstructured covariates ultimately need to be recovered into their original natural language format, recovering text from the embedding space requires mimicking the style, diversity, and semantic content of the original survey responses. Together, these difficulties make the construction of a large-scale yet realistic synthetic population from the BRIGHTEN study a non-trivial task. In what follows, we shall illustrate the steps that address the above-mentioned challenges.

First, we preprocess the covariates in the original BRIGHTEN study dataset to prepare for the downstream synthetic data generation. Structured numeric covariates are median-imputed and standardized, while categorical covariates are one-hot encoded. Unstructured covariates, such as survey responses, are first normalized so that missing entries are mapped to the string "No response." We then embed each unstructured covariate separately using the pretrained sentence

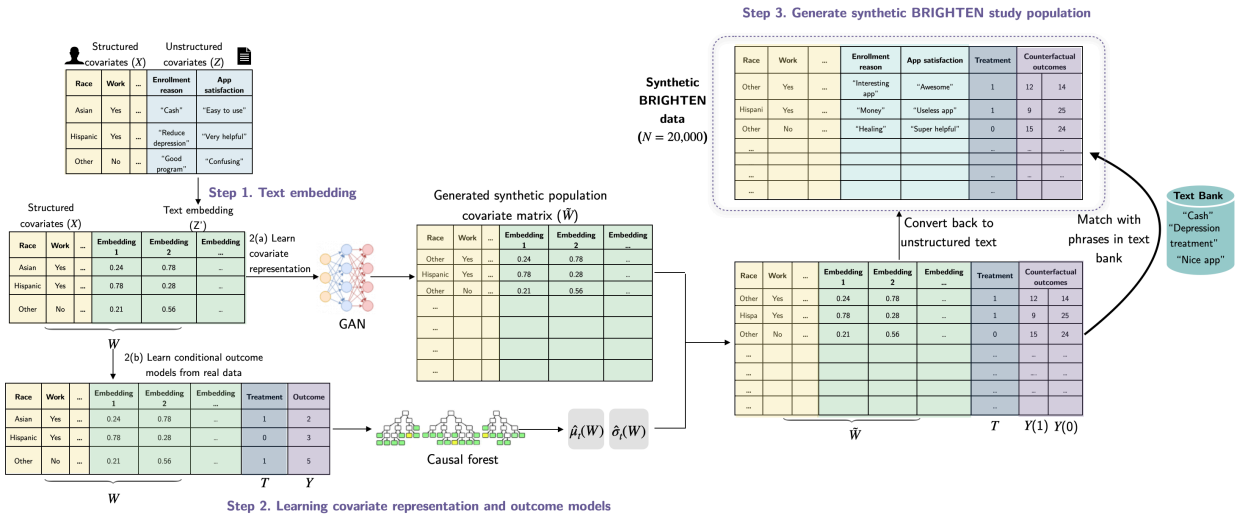


Figure 2: Illustration of the synthetic BRIGHTEN data generation.

transformer `all-MiniLM-L6-v2` [52], applying padding, truncation, mean pooling with attention masks, and l_2 normalization to obtain dense text embeddings. We adopt `all-MiniLM-L6-v2` because it is a compact, contrastively trained transformer model that produces high-quality sentence embeddings with well-preserved semantic structure. Finally, embeddings from the two unstructured covariate columns are then combined with the processed structured covariates to form the final covariate representation, denoted as $W_i = (X_i, Z_i')$, where Z_i'' represents the embedding vectors.

Second, we learn both a generative model for the covariates and predictive models for the outcomes. We first model the joint covariate distribution $p(W)$ using a conditional Generative Adversarial Network (CTGAN) [55]. From the BRIGHTEN dataset, we construct a CTGAN training table that includes both the structured covariates and the two text-embedding blocks obtained from preprocessing. To improve sample efficiency and training stability, each 384-dimensional text block is compressed via principal component analysis (PCA) to 32 components.The structured covariates and PCA-compressed embeddings are then concatenated to form a unified training table, whose schema is validated with `Metadata` before fitting a `CTGANSynthesizer` [39]. This yields a generator G that models $\widehat{p}(W)$, which we use to draw synthetic covariates for downstream simulations. Next, using the embedded BRIGHTEN dataset $\{W_i, T_i, Y_i\}_{i=1}^n$, we fit causal forests to learn the conditional mean and conditional variance, $\mu_t(W) := \mathbb{E}(Y|W, T = t)$, $\sigma_t^2(W) := \mathbb{E}((Y - \mu_t(W)^2))$, $t \in \{0, 1\}$. using the `R` package `grf` [49]. We denote the fitted models as $\widehat{\mu}_t(W)$ and $\widehat{\sigma}_t^2(W)$. Finally, because the original outcome is the PHQ-9 score (0-27), we categorize it into five clinically standard levels of depression severity—minimal, mild, moderate, high, and severe—to simplify prediction tasks for LLMs. LLMs are more effective when predicting discrete, semantically meaningful categories rather than continuous numeric scores. After classifying participants into these categories, we recode the categories numerically from 1 to 5 to represent increasing severity.

Third, we generate the synthetic population of sample size $N = 20,000$. Using the generator learned in the second step, we first generate the synthetic covariate matrix, denoted as $\widetilde{W}_j = G(\nu_j)$,

where $\nu_j \sim N(0, I_r)$, $j = 1, \ldots, N$. We then generate the treatment variable as $\widetilde{T}_j \sim$ Bernoulli(0.5), $j = 1, \ldots, N$, and the potential outcomes as $\widetilde{Y}_j(t) = \widehat{\mu}_t(\widetilde{W}_j) + \widehat{\sigma}_t(\widetilde{W}_j)\varepsilon_j$, $\varepsilon_j \sim N(0, 1)$, $t \in \{0, 1\}$. The observed synthetic outcomes are generated as $\widetilde{Y}_j = \widetilde{T}_j \cdot \widetilde{Y}_j(1) + (1 - \widetilde{T}_j) \cdot \widetilde{Y}_j(0)$, $j = 1, \ldots, N$. The true ATE based on the synthetic population is computed as $\tau := \frac{1}{N} \sum_{j=1}^{N} \widetilde{Y}_j(1) - \widetilde{Y}_j(0)$.

Lastly, because the generated covariates $\widetilde{W}_j$ consist of $(\widetilde{X}_j, \widetilde{Z}'_j)$, where $\widetilde{Z}'_j$ is represented in the numerical embedding space, we recover $\widetilde{Z}'_j$ back into unstructured text. To recover $\widetilde{Z}'_j$, we first construct a phrase bank for each free-text field (impression of the mobile app, reasons for enrolling) using responses from the original BRIGHTEN dataset after normalizing the raw texts. The surviving unique strings form the seed set $\mathcal{S}_k$ for unstructured covariate column k, where $k \in \{1, 2\}$. We then paraphrase each seed using pretrained LLMs–Pre-training with Extracted Gap-sentences for Abstractive Summarization(PEGASUS) and Text-to-Text Transfer Transformer(T5)– to increase lexical and semantic diversity. PEGASUS is a transformer model specialized for text summarization and paraphrase generation, trained with a gap-sentence generation objective that makes it highly effective for producing semantically faithful variations of input text [57]. T5 (Text-to-Text Transfer Transformer) is a general-purpose encoder-decoder transformer that frames all NLP tasks in a unified text-to-text format, enabling effective paraphrase generation among other tasks [44]. Candidate paraphrases are filtered by enforcing length constraints and require cosine similarity in the PCA-compressed embedding space to fall within a semantic band-pass window $[0.65, 0.98]$, relative to the original seed. This procedure yields the final phrase bank $\mathcal{B}_k$ for each unstructured covariate column, with size fixed at 50,000. All phrases in $\mathcal{B}_k$ are then re-embedded using the same encoder applied in the first step and projected through the saved PCA transformation. For each row of $\widetilde{Z}_i$, we perform top-1 nearest neighbor retrieval to identify the closest matching string, thereby recovering natural-language representations $\widetilde{Z}_i$ corresponding to the embedding vectors $\widetilde{Z}'_i$.

5.3 Benchmarking CALM with other methods

Using the synthetic BRIGHTEN super population generated in the previous section, we conduct simulation studies to evaluate using 300 Monte Carlo samples. Each Monte Carlo sample (of various sample sizes) is randomly drawn from this synthetic population. This section focuses on comparing CALM to alternative methods. The subsequent section offers additional insights into CALM's performance.

Within CALM, we compare several variants: zero-shot learning (Section 2.1) and few-shot learning (Section 2.2). For the few-shot variant, we examine prompts containing $m \in \{6, 10, 14\}$ examples, with $B = 200$. CALM employs `GPT-4o-mini` to generate counterfactual outcome predictions. For further context and robustness, the following case study in Section 6 includes additional results using `gemini-2.5-flash` and `GPT-3.5-turbo` as alternative LLMs for generating counterfactual outcomes. An illustrative example of zero- and few-shot learning prompts are provided in the Supplementary Materials Section 4.

For methods in comparison, we consider several variants of AIPW (or, equivalently, the double machine learning method): (1) `AIPW`: the standard AIPW estimator using only structured covari-

ates; (2) `AIPW(zero-shot covariate)`: the AIPW estimator augmented with additional covariates derived from LLM generated counterfactual outcomes with zero-shot learning; (3) `AIPW(few-shot covariate)`: the AIPW estimator augmented with additional covariates derived from LLM generated counterfactual outcomes with few-shot learning, with $m \in 6, 10, 14$ examples included in the prompt. For both the CALM and AIPW methods, we use random forests to estimate conditional outcome models. A comparison of performance using random forests versus gradient boosting is provided in Supplementary Materials Section 4 (Figure 1). The random forest approach exhibits slightly lower bias and standard deviation than the gradient boosting method. We assess performance based on absolute bias, the $\sqrt{n}$-scaled standard deviation (SD) of the ATE estimates, and coverage probability. Results are summarized in Figure 3.

Figure 3 (A)–(C) compares CALM with zero-shot learning against CALM with few-shot learning across different numbers of few-shot examples. The results show that increasing the number of few-shot examples in the prompt leads to progressively greater efficiency gains, with $m = 14$ achieving the lowest standard deviation. We conjecture that this is because few-shot learning provides additional information and guidance to the LLM in generating outcomes, leading to stronger correlation with the observed outcomes and consequently improving efficiency in the estimated treatment effects. The validity of statistical inference of the CALM method is further supported by the coverage probabilities, where the CALM method with zero-shot or few-shot learning achieves nominal level coverage at 95% level.

Figures 3(D)–(F) show that CALM notably achieves significant efficiency improvements compared to all evaluated AIPW-based methods. Furthermore, while standard AIPW, AIPW augmented with zero-shot LLM counterfactual predictions, and CALM with few-shot learning all provide consistent estimates of the ATE, the bias in AIPW augmented by few-shot learning-generated outcomes as covariates increases with the number of examples included in the few-shot prompt. Consequently, coverage probabilities decline, leading to persistent undercoverage. We hypothesize that this increased bias and lowered coverage probability occur due to the reuse of samples when generating synthetic counterfactual outcomes in few-shot learning. In contrast, using zero-shot LLM predictions in AIPW maintains nominal coverage and avoids this bias.

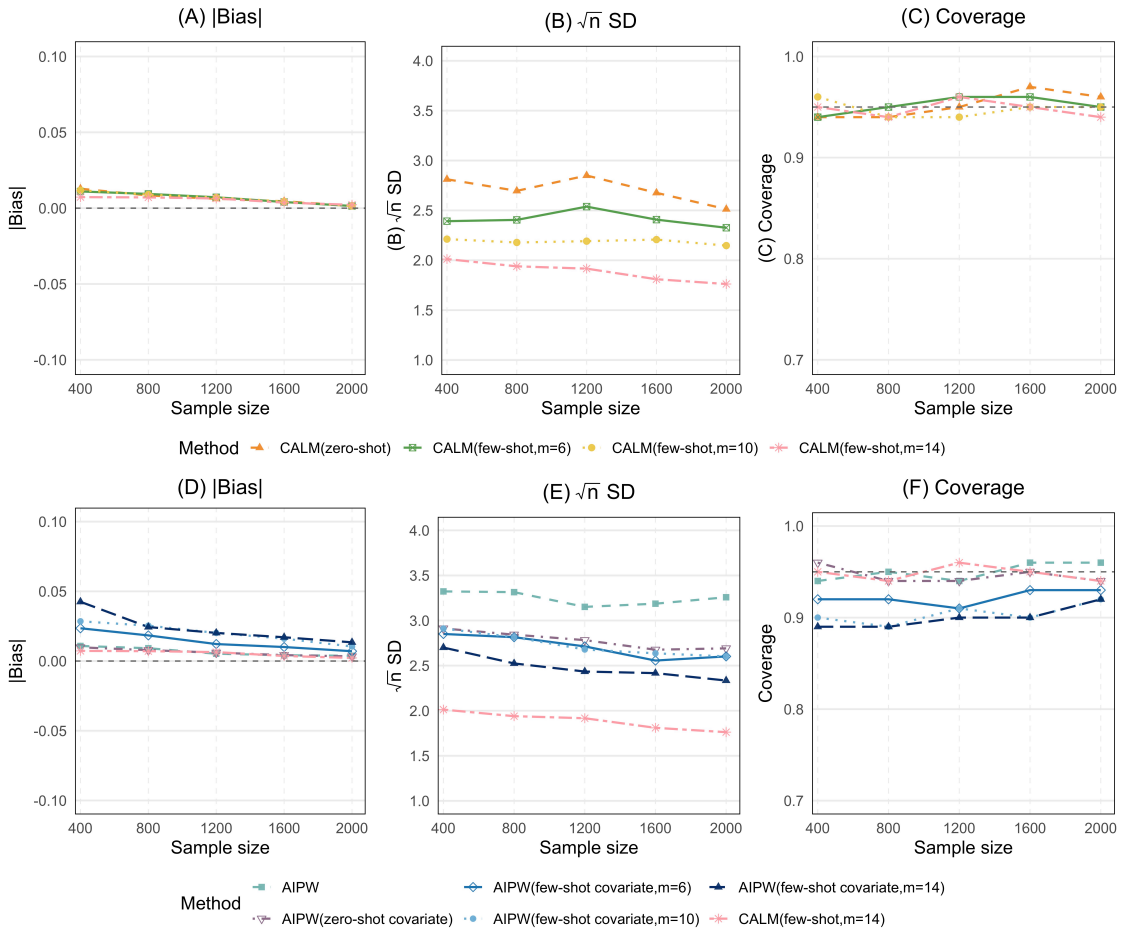


Figure 3: (A)–(C) Comparison of CALM with zero-shot learning with CALM with few-shot learning under different numbers of examples used in the few-shot prompt, $m \in \{6, 10, 14\}$. All four CALM-based methods use random forest for estimating the conditional mean models. (D)–(F) Comparison of CALM with benchmark AIPW-based methods regarding absolute bias, $\sqrt{n}$-scaled standard deviation, and coverage probability of ATE estimates across 300 Monte Carlo simulations. "AIPW" refers to the standard AIPW estimator where the conditional mean model is estimated using a random forest.

5.4 Insights on CALM

In the previous section, the CALM method demonstrated efficiency gains over the benchmark methods. Here, we provide further insights into CALM by addressing two key questions: How informative are pre-trained LLMs, and how informative are LLM-based predictions through $\omega_t(x)$ in Section 5.4.1 and Section 5.4.2 , respectively.

5.4.1 Does pre-trained LLM contain relevant information?

As outlined in the motivation for our proposed method, leveraging LLMs to generate outcome predictions offers two main advantages: (i) the ability to exploit the knowledge embedded in pre-trained LLMs, and (ii) the seamless integration of unstructured covariates. These considerations

naturally raise two questions: How much of the gain can be attributed to the prior knowledge encoded in pre-trained LLMs themselves? And to what extent do unstructured covariates enhance the informativeness of LLM predictions, and consequently improve the efficiency of treatment effect estimation?

In this section, we consider four different variations of the CALM methods with different counterfactual synthetic outcomes generating mechanism: random forest and gradient boosting that only incorporate structured covariates X, zero-shot learning that only incorporates X, and zero-shot learning that incorporates both X and Z (our proposed method in Section 2.1). The comparison among these settings allows us to isolate and assess the additional information contributed by pre-trained LLM and unstructured covariates. To evaluate the informativeness of the predictions, we compare the standard deviation among different methods and the calibration weight in Figure 4.

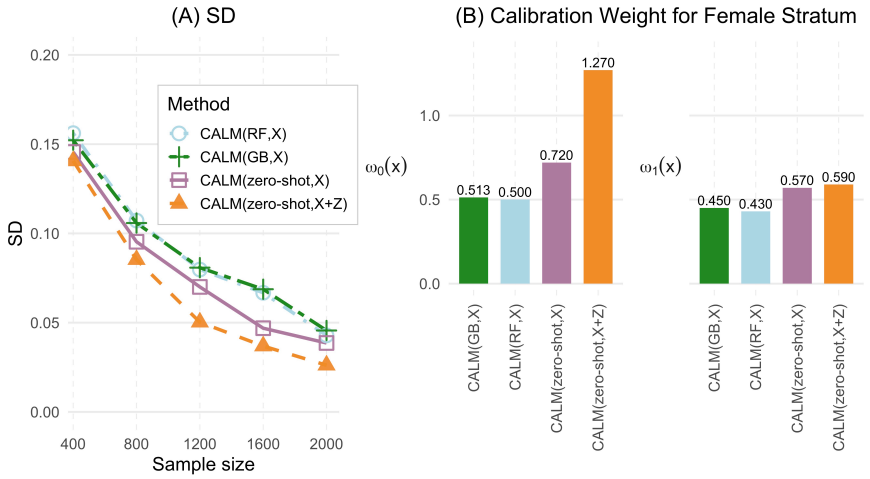


Figure 4: (A) Comparison of the standard deviation under different synthetic outcome generation mechanisms. `CALM(RF,`X`)` and `CALM(GB,`X`)` refer to the CALM-based method where the synthetic counterfactual outcomes are generated by random forest and gradient boosting using only the structured covariates X. `CALM(zero-shot,`X`)` refers to the CALM zero-shot learning method where only the structured covariate X's are included in the prompt. `CALM(zero-shot,`$X + Z$`)` refers to our proposed CALM zero-shot learning method where both the structured covariate X and the unstructured covariates Z are included in the prompt. (B) An example of the calibration weight under different synthetic outcome-generating mechanisms in the female stratum.

Figure 4 demonstrates that our proposed method with zero-shot learning–whether or not unstructured covariates are included in the prompt–yields higher estimation efficiency compared to using standard machine learning methods alone. Moreover, incorporating unstructured covariates into the zero-shot prompt provides an additional efficiency gain. These results suggest that the pre-trained LLM (`gpt-4o-mini` in this setting) encodes knowledge that improves the efficiency of ATE estimation. Even without informative unstructured inputs, the LLM's exposure to large and diverse training corpora may enable it to capture latent associations between covariates and outcomes, which CALM can then exploit to enhance efficiency. Incorporating unstructured covariates into the prompt further amplifies this benefit.

5.4.2 How informative are LLM-based predictions through $\omega_t(x)$?

Next, we provide some insights on the informativeness of LLM-based predictions through the lens of the calibration weights $\omega_t(x)$. Specifically, we examine the relationship between the LLM-predicted outcomes and the observed outcomes across different covariate strata. This analysis is motivated by the inherent structure of the CALM method, which selectively borrows information from distinct regions of the covariate space, depending on where the LLM predictions are most informative.

To better demonstrate the mechanism underlying the efficiency gain, we show the calibration weight $\omega_t(x)$ between the LLM-predicted outcomes and observed outcomes in Figure 5 (A) across strata defined by gender and race. To quantify how these calibration weights translate into improved efficiency, we report the percentage reduction in estimator variance between the standard AIPW method and the CALM method with zero-shot learning in Figure 5 (B). Specifically, we compute the variance reduction as $\frac{\mathbb{V}_{\texttt{AIPW}} - \mathbb{V}_{\texttt{CALM}}}{\mathbb{V}_{\texttt{AIPW}}} \times 100\%$.

Figure 5 suggests substantial heterogeneity in the strength of these correlations–across strata and between treatment arms. In some strata, LLM-predicted outcomes exhibit strong alignment with the observed outcomes (e.g., American Indian/Alaskan Native), while in others the alignment is weaker (e.g., Other race). This pattern suggests that the informativeness of the LLM-based predictions is not uniform across the covariate space. This heterogeneity plays a significant role in our proposed CALM method. Rather than uniformly leveraging the LLM predictions, CALM adaptively borrows strength in regions of the covariate space where the predictions are most reliable. To evaluate the practical implications of this selective borrowing, we compare the variance of the CALM estimator to that of the AIPW estimator across strata. Figure 5 (B) reports the percentage reduction in variance achieved by CALM within each stratum. The results show that the degree of variance reduction achieved by CALM aligns closely with the strength of the calibration weight shown in Figure 5 (A). In strata where LLM predictions are highly aligned with observed outcomes, CALM yields greater variance reductions. Together, these findings suggest that the calibration weight of selectively borrowing information from LLM-generated predictions contribute to the improved estimation efficiency of our proposed method.

6 Case study

In this section, we apply our proposed method to the real BRIGHTEN study dataset. We summarize the takeaways from the case study section as follows: First, the CALM-based methods achieve higher estimation efficiency and are able to detect significant treatment effects in the female Hispanic subgroup compared to the benchmark approaches. Second, the CALM-based methods attain slightly greater efficiency when generating the synthetic outcomes under `GPT-4o-mini` or `Gemini-2.5-flash`, compared to `GPT-3.5-turbo`. Third, CALM's performance is robust to variations in prompt design, demonstrating invariance to different prompting strategies.

We begin by estimating the ATE using CALM and the benchmark methods. The resulting point estimates and 95% confidence intervals are shown in Figure 6 (A). Consistent with the simulation

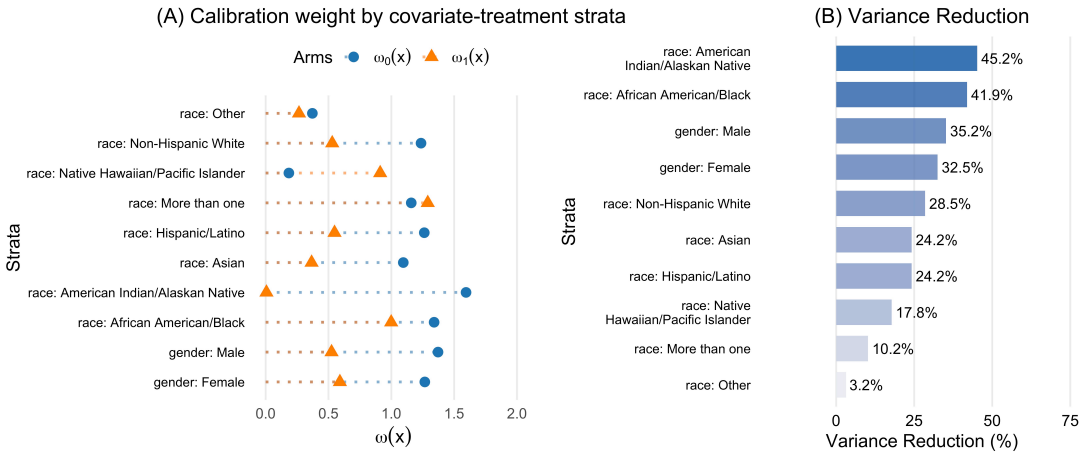


Figure 5: (A) Calibration weight $\omega_t(x)$ for the "race" and "gender" strata using the CALM zero-shot learning method in comparison with the AIPW method ($n = 2,000$) and (B) the variance reduction in comparison with the AIPW estimator. Both the CALM zero-shot method and the AIPW method use random forest for conditional mean estimation.

findings, CALM with few-shot learning yields the narrowest confidence intervals, reflecting higher estimation efficiency. In contrast, AIPW augmented with few-shot learning produces a significantly negative treatment effect estimate, which is likely attributable to the estimation bias documented in Section 5.3. To assess the impact of different LLMs on CALM's performance, we compare results using three models: `GPT-4o-mini`, `GPT-3.5-turbo`, and `Gemini-2.5-flash`, as shown in Figure 6 (B). The figure shows that `GPT-4o-mini` and `Gemini-2.5-flash` achieve comparable efficiency, whereas `GPT-3.5-turbo` yields slightly wider confidence intervals, suggesting reduced estimation efficiency due to less informative LLM-based predictions relative to the other two models.

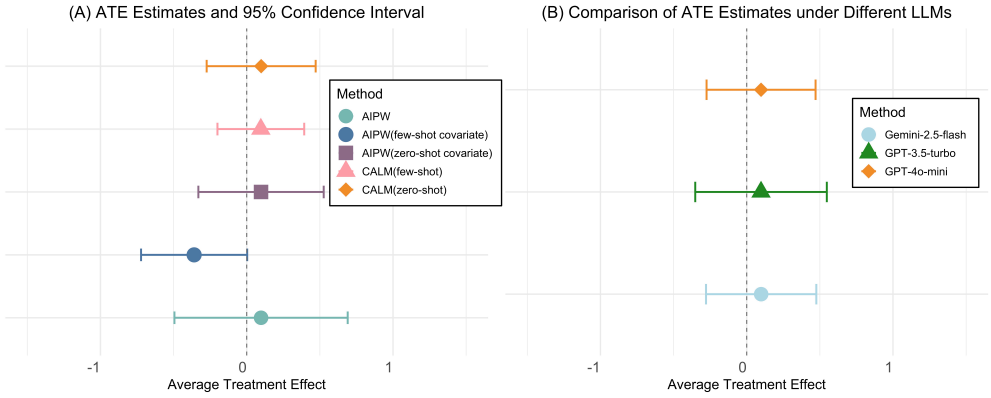


Figure 6: (A) Comparison of the ATE estimates and the associated 95% confidence intervals from the BRIGHTEN study under AIPW and CALM-based methods. The conditional outcome models are estimated using random forest all the methods in comparison. (B) Comparison of different LLMs for generating synthetic outcomes under CALM zero-shot learning method.

Furthermore, we investigate strata-level treatment effects, with results presented in Figure 7.

The figure shows that CALM with zero-shot or few-shot learning are able to detect a significant treatment effect among female Hispanic participants, which is not identified by the benchmark AIPW methods. This finding highlights how the efficiency gains achieved by CALM translate into improved power for subgroup analyses in the BRIGHTEN study, enabling the detection of significant treatment effect heterogeneity of the mobile app intervention. In this case, the enhanced efficiency also facilitates scientific discovery.

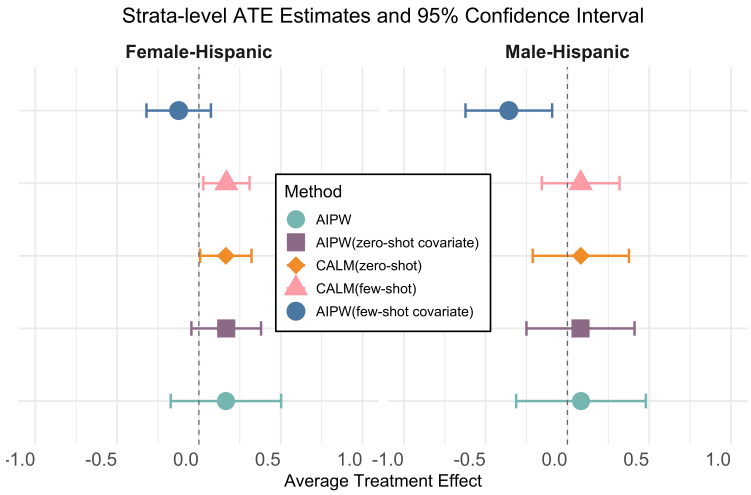


Figure 7: Comparison of the ATE estimates and the associated 95% confidence intervals from the BRIGHTEN study under AIPW and CALM-based methods in female Hispanic and male Hispanic strata.

Additionally, in the preceding simulation studies, we fix a single prompt engineering technique for generating LLM-based predictions. A natural question is whether these predictions are sensitive to prompt design, and whether alternative formulations might alter estimation performance within the CALM framework. To assess this, we conduct a robustness analysis using four prompt engineering strategies: (1) self-consistency, which generates multiple predictions under different sampling seeds and aggregates them to reduce variance; (2) role-based prompting, which assigns the LLM a specific role (e.g., "You are an experienced clinical researcher") to provide domain-specific context; (3) decomposition prompting, which breaks the task into sequential sub-questions to capture complex covariate dependencies; and (4) contrastive prompting, which presents systematically varied hypothetical cases to highlight differences in predicted outcomes. Examples of these prompt designs and the ATE estimates with associated 95% confidence intervals are shown in Figure 8.

Figure 8 shows that all four prompting strategies yield comparable point estimates and confidence intervals, regardless of the specific prompt design. Among them, role-based prompting produces slightly narrower intervals, indicating a modest efficiency gain, though the difference is small and not statistically significant relative to the other methods. Overall, the results suggest that CALM is robust to variations in prompt engineering, even when the structure and contextual framing differ substantially. This stability under different prompt engineering implies that CALM's performance does not critically depend on prompt format, enhancing its practical applicability in

settings where prompt design may vary.

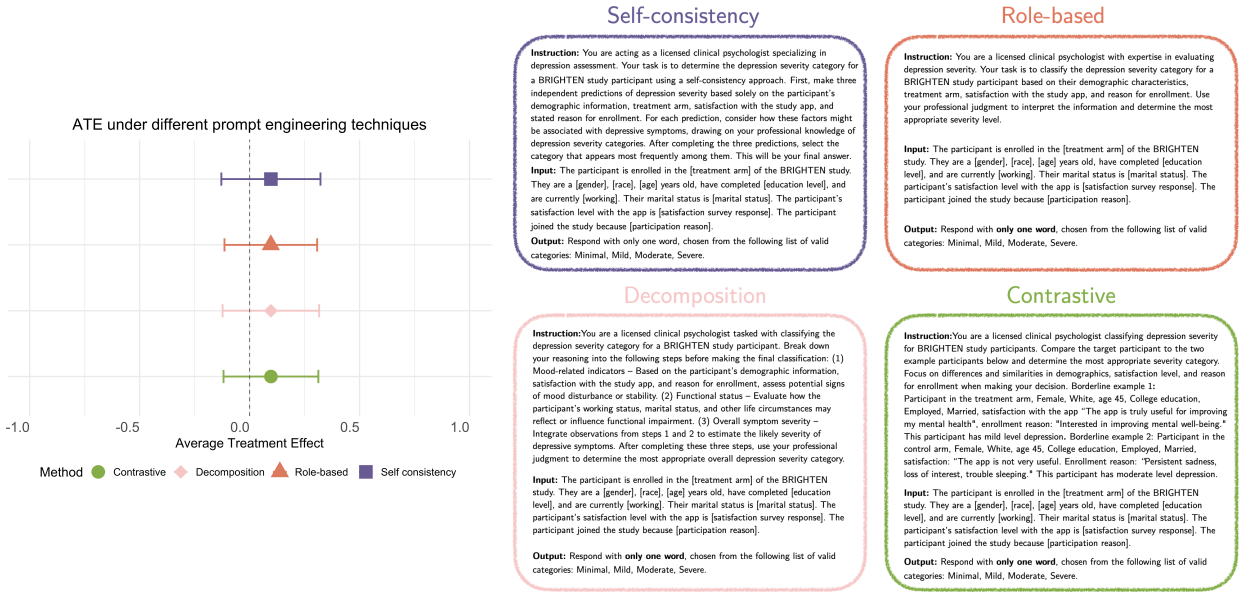


Figure 8: Comparison of ATE and the associated 95% confidence intervals under four different prompt engineering techniques: self-consistency, role-based, decomposition, and contrative methods.

7 Conclusion

In this manuscript, we introduced **CALM**, a statistically principled framework for leveraging LLM-generated predictions in randomized experiments. By residualizing and heterogeneously calibrating LLM predictions, CALM yields consistent estimators with valid inference even when zero- and few-shot predictions are biased, and it empirically improves efficiency relative to covariate-adjusted AIPW for mean potential outcomes; extensions to average and heterogeneous treatment effects follow directly. In simulations calibrated to the BRIGHTEN RCT, CALM reduces variance and increases power for detecting treatment-effect heterogeneity while remaining robust to prompt design.

Data availability

The BRIGHTEN study data are publicly accessible upon request via Synapse: `https://www.synapse.org/Synapse:syn10848316/wiki/548727` (Synapse Project:syn10848316).

Acknowledgments

Jingshen Wang acknowledges support from Neuromatch and the National Science Foundation (NSF DMS-2239047). Xinwei Ma and Jingshen Wang acknowledge support from the National Institutes of Health (NIH R01-AG089512).

References

[1] Abrevaya, J., Hsu, Y.-C., and Lieli, R. P. (2015). "Estimating conditional average treatment effects," *Journal of Business & Economic Statistics*, *33*(4), 485–505.

[2] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y. *et al.* (2022). "Flamingo: A visual language model for few-shot learning," In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 23716–23730.

[3] Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. (2023). "Prediction-powered inference," *Science*, *382*(6671), 669–674.

[4] Angelopoulos, A. N., Duchi, J. C., and Zrnic, T. (2024). "Ppi++: Efficient prediction-powered inference," arXiv preprint arXiv:2311.01453.

[5] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S. *et al.* (2021). "On the opportunities and risks of foundation models," arXiv preprint arXiv:2108.07258.

[6] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P. *et al.* (2020). "Language models are few-shot learners," In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 1877–1901.

[7] 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.

[8] Cattaneo, M. D. (2010). "Efficient semiparametric estimation of multi-valued treatment effects under ignorability," *Journal of Econometrics*, *155*(2), 138–154.

[9] Cattaneo, M. D., Jansson, M., and Ma, X. (2019). "Two-step estimation and inference with possibly many included covariates," *The Review of Economic Studies*, *86*(3), 1095–1122.

[10] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, *21*(1), C1–C68.

[11] Coravos, A., Goldsack, J. C., Karlin, D. R., Nebeker, C., Perakslis, E., and Zimmerman, N. (2020). "Modernizing and designing evaluation frameworks for connected sensor technologies in medicine," *npj Digital Medicine*, *3*(37), 1–10.

[12] De Bartolomeis, P., Abad, J., Wang, G., Donhauser, K., Duch, R. M., Yang, F., and Dahabreh, I. J. (2025). "Efficient randomized experiments using foundation models," arXiv preprint arXiv:2502.04262.

[13] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "BERT: Pre-training of deep bidirectional transformers for language understanding," In *Proceedings of NAACL-HLT 2019*, 4171–4186.

[14] Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B. *et al.* (2023). "PaLM-E: An embodied multimodal language model," In *Proceedings of the 40th International Conference on Machine Learning*, 8469–8488.

[15] Egami, N., Hinck, M., Stewart, B., and Wei, H. (2023). "Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models," In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 68589–68601.

[16] Farrell, M. H. (2015). "Robust inference on average treatment effects with possibly more covariates than observations," *Journal of Econometrics*, *189*(1), 1–23.

[17] Fisch, A., Maynez, J., Hofer, R., Dhingra, B., Globerson, A., and Cohen, W. W. (2024). "Stratified prediction-powered inference for effective hybrid evaluation of language models," In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 111489–111514.

[18] Freedman, D. A. (2008). "On regression adjustments to experimental data," *Advances in Applied Mathematics*, *40*(2), 180–193.

[19] Gao, T., Fisch, A., and Chen, D. (2021). "Making pre-trained language models better few-shot learners," In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 3816–3830.

[20] Gemini Team Google (2025). "Gemini: a family of highly capable multimodal models," arXiv preprint arXiv:2312.11805.

[21] Glasserman, P. and Lin, C. (2023). "Assessing look-ahead bias in stock return predictions generated by gpt sentiment analysis," arXiv preprint arXiv:2309.17322.

[22] Guterman, E. L., Kiekhofer, 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.

[23] 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.

[24] Ji, W., Lei, L., and Zrnic, T. (2025). "Predictions as surrogates: Revisiting surrogate outcomes in the age of AI," arXiv preprint arXiv:2501.09731.

[25] Kennedy, E. H. (2023). "Towards optimal doubly robust estimation of heterogeneous causal effects," *Electronic Journal of Statistics*, *17*(2), 3008–3049.

[26] Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F. *et al.* (2017). "Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review," *Journal of Biomedical Informatics*, *73*, 14–29.

[27] Kroenke, K., Spitzer, R. L., and Williams, J. B. (2001). "The PHQ-9: validity of a brief depression severity measure," *Journal of General Internal Medicine*, *16*(9), 606–613.

[28] Labrak, Y., Rouvier, M., and Dufour, R. (2024). "A zero-shot and few-shot study of instruction-finetuned large language models applied to clinical and biomedical tasks," In *LREC-COLING 2024*, 2049–2066.

[29] Li, J., Li, D., Savarese, S., and Hoi, S. (2023). "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," In *Proceedings of the 40th International Conference on Machine Learning*, 19730–19742.

[30] Li, Q. and Racine, J. S. (2007). *Nonparametric Econometrics: Theory and Practice*, Princeton, NJ: Princeton University Press.

[31] Lin, W. (2013). "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique," *The Annals of Applied Statistics*, *7*(1), 295–318.

[32] Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023). "Visual instruction tuning," In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 34892–34916.

[33] Lopez-Lira, A., Tang, Y., and Zhu, M. (2025). "The memorization problem: Can we trust LLMs' economic forecasts?," arXiv preprint arXiv:2504.14765.

[34] Ludwig, J., Mullainathan, S., and Rambachan, A. (2025). "Large language models: An applied econometric framework," arXiv preprint arXiv:2412.07031.

[35] Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., and Liu, T.-Y. (2022). "BioGPT: Generative pre-trained transformer for biomedical text generation and mining," *Briefings in Bioinformatics*, *23*(6), bbac409.

[36] Ma, X. and Wang, J. (2020). "Robust inference using inverse probability weighting," *Journal of the American Statistical Association*, *115*(532), 1851–1860.

[37] Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. (2022). "Rethinking the role of demonstrations: What makes in-context learning work?" In

Supplementary Materials of "Can language models boost the power of randomized experiments without statistical bias?"

Xinrui Ruan[1] Xinwei Ma[2] Yingfei Wang[3] Waverly Wei[4] Jingshen Wang[1]

[1]Division of Biostatistics, University of California, Berkeley
[2]Department of Economics, University of California, San Diego
[3]Michael G. Foster School of Business, University of Washington
[4]Department of Data Sciences and Operations, University of Southern California

Contents

1 Proofs **1**
1.1 Proof of Theorem 1 . . . 1
1.2 Proof of Lemma 1 . . . 2
1.3 Proof of Lemma 2 . . . 4
1.4 Proof of Corollary 1 . . . 5
1.5 Proof of Theorem 2 . . . 7
1.6 Proof of Lemma 6 . . . 9
1.7 Proof of Corollary 2 . . . 10
1.8 Proof of Corollary 3 . . . 12

2 Test of efficiency improvement **15**

3 Motivating real-world RCTs **16**

4 Synthetic superpopulation generation **16**

5 Additional simulation study details **18**

1 Proofs

1.1 Proof of Theorem 1

In this section, we provide the proof of Theorem 1 (asymptotic properties of CALM with zero-shot learning) from the main paper. Following standard arguments for analyzing estimators with

estimated nuisance functions, the proof is divided into two parts. In the first part, we establish the asymptotic properties of the *oracle* CALM estimator, which is defined using the true nuisance functions $\mu_t(\cdot)$, $\mu_t^\dagger(\cdot)$, and $\omega_t(\cdot)$ rather than their estimates:

$$\widehat{\mu}_{t,\texttt{CALM},\texttt{Oracle}} = \frac{1}{n}\sum_{i=1}^{n} \varphi_t(Y_i, T_i, X_i, Y_i^\dagger(t)),$$

where:

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

We first state the following lemma on the asymptotic behavior of $\widehat{\mu}_{t,\texttt{CALM},\texttt{Oracle}}$.

Lemma 1 (Asymptotic properties of the oracle CALM estimator)**.** *Under Assumptions 1-2, the oracle CALM estimator is a consistent and asymptotically normal estimator of μ_t. Specifically, we have $\widehat{\mu}_{t,\texttt{CALM},\texttt{Oracle}} \xrightarrow{p} \mu_t$, and:*

$$\sqrt{n}\left(\widehat{\mu}_{t,\texttt{CALM},\texttt{Oracle}} - \mu_t\right) \xrightarrow{d} \mathcal{N}\left(0, \mathsf{V}_{t,\texttt{CALM},\texttt{Oracle}}\right),$$

where the asymptotic variance is given by:

$$\mathsf{V}_{t,\texttt{CALM},\texttt{Oracle}} = \mathbb{E}\left[\left(\mu_t(X) - \mu_t\right)^2 + \frac{\mathbb{V}\mathrm{ar}(Y(t) \mid X)}{e_t(X)}\left(1 - (1 - e_t(X))\rho_t^2(X)\right)\right],$$

and $\rho_t(x) = \mathbb{C}orr(Y(t), Y^\dagger(t) \mid X = x)$ denotes the conditional correlation between the true potential outcome and the LLM-based prediction. For covariate values $x \in \mathcal{X}$ such that $\mathbb{V}\mathrm{ar}(Y^\dagger(t) \mid X = x) = 0$, we define $\rho_t(x) = 0$ by convention.

In the second part, we show that under Assumption 3, the estimation error arising from cross-fitted nuisance functions is asymptotically negligible relative to the oracle CALM estimator. This is formalized in the following lemma.

Lemma 2 (Asymptotic equivalence to oracle)**.** *Under Assumptions 1–3, the proposed CALM estimator $\widehat{\mu}_{t,\texttt{CALM}}$ (Section 2.1) is asymptotically equivalent to the oracle CALM estimator $\widehat{\mu}_{t,\texttt{CALM},\texttt{Oracle}}$, in the sense that*

$$\sqrt{n}\big(\widehat{\mu}_{t,\texttt{CALM}} - \widehat{\mu}_{t,\texttt{CALM},\texttt{Oracle}}\big) = o_p(1).$$

Lemma 2 implies that $\sqrt{n}\big(\widehat{\mu}_{t,\texttt{CALM}} - \mu_t\big)$ shares the same asymptotic distribution as $\sqrt{n}\big(\widehat{\mu}_{t,\texttt{CALM},\texttt{Oracle}} - \mu_t\big)$, thereby completing the proof of Theorem 1.

1.2 Proof of Lemma 1

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 $h(X, Z)$, we have

$$\begin{aligned}\mathbb{E}\left[\left(1-\frac{1\{T=t\}}{e_t(X)}\right)h(X,Z)\right] &= \mathbb{E}\left[\mathbb{E}\left[\left(1-\frac{1\{T=t\}}{e_t(X)}\right)h(X,Z)\middle|\, X,Z\right]\right] \\ &= \mathbb{E}\left[h(X,Z)\left(1-\frac{\mathbb{E}[1\{T=t\}\mid X,Z]}{e_t(X)}\right)\right] \\ &= 0,\end{aligned}$$

where the last equality uses Assumption 1 that $\mathbb{E}[1\{T=t\} \mid X, Z] = E[1\{T=t\} \mid X] = e_t(X)$. Recall from the main paper that the LLM-based zero-shot prediction $Y^\dagger(t)$ can be viewed as function of the query covariates (X, Z) and the treatment arm t, namely $f_\theta\big((X,Z),t\big)$, where θ denotes the pre-trained parameters of the LLM. Hence, $\mu_t(X)+\omega_t(X)(Y^\dagger(t)-\mu_t^\dagger(X))$ is a function with respect to (X, Z). The above argument then implies:

$$\mathbb{E}[\varphi_t(Y,T,X,Y^\dagger(t))] = \mathbb{E}\left[\frac{\mathbf{1}\{T=t\}Y}{e_t(X)}\right] = \mathbb{E}[Y(t)] = \mu_t.$$

Thus, $\varphi_t(Y, T, X, Y^\dagger(t))$ is an unbiased estimating function for μ_t. By Assumptions 2, the sequence $\{\varphi_t(Y_i, T_i, X_i, Y_i^\dagger(t))\}_{i=1}^n$ is i.i.d. with finite variance. Therefore, the standard Law of Large Numbers (LLN) and Central Limit Theorem (CLT) yield

$$\widehat{\mu}_{t,\texttt{CALM},\texttt{Oracle}} \xrightarrow{p} \mu_t, \quad \sqrt{n}\big(\widehat{\mu}_{t,\texttt{CALM},\texttt{Oracle}} - \mu_t\big) \xrightarrow{d} \mathcal{N}(0, \mathbb{V}\mathrm{ar}(\varphi_t))\,.$$

Asymptotic variance computation. We now show the detailed computation of the asymptotic variance $\mathbb{V}\mathrm{ar}(\varphi_t)$. By definition,

$$\begin{aligned}\mathbb{V}\mathrm{ar}(\varphi_t) =& \mathbb{E}\left[(\varphi_t(Y,T,X,Y^\dagger(t)) - \mu_t)^2\right] \\ =& \mathbb{E}\left\{\left[\mu_t(X) - \mu_t + \frac{\mathbf{1}\{T=t\}}{e_t(X)}(Y-\mu_t(X)) + \left(1-\frac{\mathbf{1}\{T=t\}}{e_t(X)}\right)\omega_t(X)\left(Y^\dagger(t)-\mu_t^\dagger(X)\right)\right]^2\right\} \\ =& \mathbb{E}\left[(\mu_t(X)-\mu_t)^2\right] + \underbrace{\mathbb{E}\Big[\Delta_t(X,T,Y,Y^\dagger(t))^2\Big]}_{\text{(I)}} + 2\underbrace{\mathbb{E}\Big[(\mu_t(X)-\mu_t)\cdot\Delta_t(X,T,Y,Y^\dagger(t))\Big]}_{\text{(II)}},\end{aligned}$$

where

$$\Delta_t(X,T,Y,Y^\dagger(t)) := \frac{\mathbf{1}\{T=t\}}{e_t(X)}(Y-\mu_t(X)) + \left(1-\frac{\mathbf{1}\{T=t\}}{e_t(X)}\right)\omega_t(X)\left(Y^\dagger(t)-\mu_t^\dagger(X)\right).$$

A direct variance–covariance expansion yields

$$\text{(I)} = \mathbb{E}\left[\frac{1}{e_t(X)}\mathbb{V}\mathrm{ar}(Y(t)\mid X) + \frac{1-e_t(X)}{e_t(X)}\Big(\omega_t^2(X)\mathbb{V}\mathrm{ar}(Y^\dagger(t)\mid X) - 2\,\omega_t(X)\mathbb{C}\mathrm{ov}(Y(t),Y^\dagger(t)\mid X)\Big)\right],$$

and

$$(\mathrm{II}) = \mathbb{E}\Big[(\mu_t(X) - \mu_t) \cdot \mathbb{E}[\Delta_t(X, T, Y, Y^\dagger(t)) \mid X]\Big] = 0,$$

since $\mathbb{E}[\Delta_t(X, T, Y, Y^\dagger(t)) \mid X] = 0$ by construction. Then, substituting (I) and (II) gives

$$\mathbb{V}\mathrm{ar}(\varphi_t) = \mathbb{E}\left[(\mu_t(X) - \mu_t)^2 + \mathbb{V}\mathrm{ar}(Y(t) \mid X) + \frac{1 - e_t(X)}{e_t(X)}\,\mathbb{V}\mathrm{ar}\big(Y(t) - \omega_t(X)Y^\dagger(t) \mid X\big)\right].$$

To minimize the asymptotic variance, recall from Step 3 of the main paper that the heterogeneous calibration function is chosen as

$$\omega_t(x) = \frac{\mathbb{C}\mathrm{ov}\left(Y, Y^\dagger(t) \mid T = t, X = x\right)}{\mathbb{V}\mathrm{ar}\left(Y^\dagger(t) \mid T = t, X = x\right)},$$

whenever $\mathbb{V}\mathrm{ar}(Y^\dagger(t) \mid X = x) > 0$. If instead $\mathbb{V}\mathrm{ar}(Y^\dagger(t) \mid X = x) = 0$, then $\mathbb{V}\mathrm{ar}\big(Y(t) - \omega_t(X)Y^\dagger(t) \mid X\big) = \mathbb{V}\mathrm{ar}\big(Y(t) \mid X\big)$ for any choice of $\omega_t(\cdot)$. Using the definition of $\rho_t(x)$, we obtain

$$\mathbb{V}\mathrm{ar}_{t,\texttt{CALM},\texttt{Oracle}} = \mathbb{V}\mathrm{ar}(\varphi_t) = \mathbb{E}\left[(\mu_t(X) - \mu_t)^2 + \frac{\mathbb{V}\mathrm{ar}(Y(t) \mid X)}{e_t(X)}\Big(1 - (1 - e_t(X))\rho_t^2(X)\Big)\right].$$

This completes the proof of Lemma 1. In the special case where $\omega_t(x) \equiv 0$ (i.e., no information is borrowed from the LLM-based prediction), the asymptotic variance degenerates to:

$$\mathbb{V}\mathrm{ar}_{\texttt{AIPW}} = \mathbb{E}\left[(\mu_t(X) - \mu_t)^2 + \frac{\mathbb{V}\mathrm{ar}(Y(t) \mid X)}{e_t(X)}\right] = \mathbb{V}\mathrm{ar}_{t,\texttt{CALM},\texttt{Oracle}}.$$

□

1.3 Proof of Lemma 2

Proof. By random sample splitting (Step 1) and the i.i.d. data collection procedure (Assumption 2), it suffices to show that

$$\sqrt{n}\left(\widehat{\mu}_{t,\texttt{CALM}}^{\mathcal{I}_1} - \widehat{\mu}_{t,\texttt{CALM},\texttt{Oracle}}^{\mathcal{I}_1}\right) = o_p(1).$$

We first decompose the difference as

$$\begin{aligned}\sqrt{n}\left(\widehat{\mu}_{t,\texttt{CALM}}^{\mathcal{I}_1} - \widehat{\mu}_{t,\texttt{CALM},\texttt{Oracle}}^{\mathcal{I}_1}\right) = \frac{\sqrt{n}}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(1 - \frac{\mathbf{1}\{T_i = t\}}{e_t\left(X_i\right)}\right)\Bigg\{ & \left(\widehat{\mu}_t^{I_2}\left(X_i\right) - \mu_t\left(X_i\right)\right) \\ & + \left(\widehat{\omega}_t^{I_2}\left(X_i\right) - \omega_t\left(X_i\right)\right)\left(Y_i^\dagger(t) - \mu_t^\dagger\left(X_i\right)\right) \\ & - \left(\widehat{\mu}_t^{\dagger I_2}\left(X_i\right) - \mu_t^\dagger\left(X_i\right)\right)\omega_t\left(X_i\right) \\ & - \left(\widehat{\mu}_t^{\dagger I_2}\left(X_i\right) - \mu_t^\dagger\left(X_i\right)\right)\left(\widehat{\omega}_t^{I_2}\left(X_i\right) - \omega_t\left(X_i\right)\right)\Bigg\}.\end{aligned}$$

We now bound the remainder terms. For the first term, conditional on $\mathcal{I}_2$ and the observed structured covariates $\{X_i\}$, its variance is

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{\sqrt{n}}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(1-\frac{\mathbf{1}\{T_i=t\}}{e_t\left(X_i\right)}\right)\left(\widehat{\mu}_t^{I_2}\left(X_i\right)-\mu_t\left(X_i\right)\right)\right)^2\middle|\mathcal{I}_2,\{X_i\}\right]\\
=&\mathbb{V}\mathrm{ar}\left[\frac{\sqrt{n}}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(1-\frac{\mathbf{1}\{T_i=t\}}{e_t\left(X_i\right)}\right)\left(\widehat{\mu}_t^{I_2}\left(X_i\right)-\mu_t\left(X_i\right)\right)\middle|I_2,\{X_i\}\right]\\
=&\frac{n}{|\mathcal{I}_1|^2}\sum_{i\in\mathcal{I}_1}\mathbb{E}\left[\left(\widehat{\mu}_t^{I_2}\left(X_i\right)-\mu_t\left(X_i\right)\right)^2\left(1-\frac{\mathbf{1}\{T_i=t\}}{e_t\left(X_i\right)}\right)^2\middle|I_2,\{X_i\}\right]\\
=&\frac{n}{|\mathcal{I}_1|^2}\sum_{i\in\mathcal{I}_1}\frac{1-e_t\left(X_i\right)}{e_t\left(X_i\right)}\left(\widehat{\mu}_t^{I_2}\left(X_i\right)-\mu_t\left(X_i\right)\right)^2\leqslant C\cdot\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\widehat{\mu}_t^{I_2}\left(X_i\right)-\mu_t\left(X_i\right)\right)^2=o_p\left(1\right),
\end{aligned}$$

where the inequality uses Assumption 1 (propensity scores bounded away from 0 and 1). Since $\widehat{\mu}_t^{\mathcal{I}_2}$ is a consistent cross-fitted estimator (Assumption 3), the bound is $o_p(1)$. By Chebyshev's inequality, the first remainder term is therefore $o_p(1)$ as well. Analogous arguments apply to the other remainder terms involving $\widehat{\omega}_t^{\mathcal{I}_2}$ and $\widehat{\mu}_t^{\dagger\mathcal{I}_2}$. For the final interaction term, the Cauchy–Schwarz inequality yields the same $o_p(1)$ rate. Putting everything together, we obtain $\sqrt{n}\left(\widehat{\mu}_{t,\texttt{CALM}}^{\mathcal{I}_1}-\widehat{\mu}_{t,\texttt{CALM,Oracle}}^{\mathcal{I}_1}\right)=o_p(1)$. □

1.4 Proof of Corollary 1

In this section, we present the proof of Corollary 1 from the main paper. To begin, we define a deterministic function $\omega_{t,\mathtt{R}}(x^{\mathtt{C}})$ that depends on the coarsened covariates $x^{\mathtt{C}}\in\{1,\ldots,M\}$. These coarsened covariates can be represented as the output of a deterministic mapping $C(X):\mathcal{X}\to\{1,\ldots,M\}$, where M is a fixed finite number that does not grow with the sample size n:

$$\omega_{t,\mathtt{R}}(x^{\mathtt{C}})=\frac{\mathbb{E}\Big[\lambda_t(X)\left(Y(t)-\widetilde{\mu}_t(X)\right)\left(Y^{\dagger}(t)-\widetilde{\mu}_t^{\dagger}(X)\right)\mid C(X)=x^{\mathtt{C}}\Big]}{\mathbb{E}\Big[\lambda_t(X)(Y^{\dagger}(t)-\widetilde{\mu}_t^{\dagger}(X))^2\mid C(X)=x^{\mathtt{C}}\Big]}.$$

By defining the function in this way, we implicitly assume that the denominator is bounded away from 0 for $x^{\mathtt{C}}\in\{1,\ldots,M\}$. Similar to the proof of Theorem 1, we divide the proof of Corollary 1 into two parts. In the first part, we define the *oracle* CALM estimator with robust heterogeneous calibration weights as:

$$\widehat{\mu}_{t,\mathtt{R},\mathtt{CALM},\mathtt{Oracle}}=\frac{1}{n}\sum_{i=1}^{n}\varphi_{t,\mathtt{R}}(Y_i,T_i,X_i,Y_i^{\dagger}(t)),$$

where:

$$\varphi_{t,\mathtt{R}}(Y_i,T_i,X_i,Y_i^{\dagger}(t))=\frac{\mathbf{1}\{T_i=t\}Y_i}{e_t(X_i)}+\left(1-\frac{\mathbf{1}\{T_i=t\}}{e_t\left(X_i\right)}\right)\left(\widetilde{\mu}_t\left(X_i\right)+\omega_{t,\mathtt{R}}\left(X_i^{\mathtt{C}}\right)\left(Y_i^{\dagger}(t)-\widetilde{\mu}_t^{\dagger}\left(X_i\right)\right)\right),$$

and the subscript R distinguishes this "robust" version of estimator from other parts. Analogous to Lemma 1, we now state the following results for $\widehat{\mu}_{t,\mathtt{R},\mathtt{CALM},\mathtt{Oracle}}$:

Lemma 3 (Asymptotic properties of the oracle CALM estimator with robust calibration)**.** *Under Assumptions 1-2, the oracle CALM estimator with robust heterogeneous calibration weights is a consistent and asymptotic normal estimator of μ_t. Specifically, we have $\widehat{\mu}_{t,\mathtt{R},\mathtt{CALM},\mathtt{Oracle}} \xrightarrow{p} \mu_t$, and $\sqrt{n}\big(\widehat{\mu}_{t,\mathtt{R},\mathtt{CALM}} - \mu_t\big) \xrightarrow{d} \mathcal{N}\Big(0, \widetilde{\mathsf{V}}_{t,\mathtt{CALM}}\Big)$, where:*

$$\widetilde{\mathsf{V}}_{t,\mathtt{CALM}} = \mathbb{E}\big[(\mu_t(X) - \mu_t)^2 + \mathbb{V}\mathrm{ar}(Y(t) \mid X) + \sigma_t^2(X^{\mathrm{C}})\{1 - \widetilde{\rho}_t^2(X^{\mathrm{C}})\}\big]\,,$$

and

$$\widetilde{\rho}_t(x^{\mathrm{C}}) = \frac{\mathbb{E}\Big[\lambda_t(X)\{Y(t) - \widetilde{\mu}_t(X)\}\{Y^\dagger(t) - \widetilde{\mu}_t^\dagger(X)\} \,\Big|\, C(X) = x^{\mathrm{C}}\Big]}{\sqrt{\sigma_t^2(x^{\mathrm{C}})\,\sigma_t^{\dagger 2}(x^{\mathrm{C}})}},$$

$\sigma_t^2(x^{\mathrm{C}})$ and $\sigma_t^{\dagger 2}(x^{\mathrm{C}})$ are the weighted conditional mean squared deviations defined in Section 4.1 of the main paper.

Proof. The consistency and asymptotic normality of $\widehat{\mu}_{t,\mathtt{R},\mathtt{CALM},\mathtt{Oracle}}$ follow directly from the proof of Lemma 1 (Section 1.2 in the Supplementary Materials), since $\omega_{t,\mathtt{R}}(X^{\mathtt{C}})$ is a measurable function of X, which can be written rigorously as $\omega_{t,\mathtt{R}}(C(X))$ for the coarsening map $C(\cdot)$. For the asymptotic variance, by an argument parallel to that in the proof of Lemma 1, we obtain

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(\varphi_{t,\mathtt{R}}) &= \mathbb{E}\Big[(\mu_t(X) - \mu_t)^2 + \mathbb{V}\mathrm{ar}(Y(t) \mid X) + \lambda_t(X)\Big((Y(t) - \widetilde{\mu}_t(X)) - \omega_{t,\mathtt{R}}(C(X)) \cdot (Y^\dagger(t) - \widetilde{\mu}_t^\dagger(X))\Big)^2\Big] \\
&= \mathbb{E}\Big[(\mu_t(X) - \mu_t)^2 + \mathbb{V}\mathrm{ar}(Y(t) \mid X)\Big] \\
&\quad + \mathbb{E}\Big[\mathbb{E}\Big[\lambda_t(X)\Big((Y(t) - \widetilde{\mu}_t(X)) - \omega_{t,\mathtt{R}}(C(X))(Y^\dagger(t) - \widetilde{\mu}_t^\dagger(X))\Big)^2 \Big| C(X) = X^{\mathtt{C}}\Big]\Big].
\end{aligned}$$

By the construction of $\omega_{t,\mathtt{R}}(X^{\mathtt{C}})$ and the definition of $\sigma_t^2(x^{\mathrm{C}})$ and $\sigma_t^{\dagger 2}(x^{\mathrm{C}})$, the inner conditional expectation reduces to $\sigma_t^2(X^{\mathtt{C}})\{1 - \widetilde{\rho}_t^2(X^{\mathtt{C}})\}$, so that

$$\widetilde{\mathsf{V}}_{t,\mathtt{CALM}} = \mathbb{V}\mathrm{ar}(\varphi_{t,\mathtt{R}}) = \mathbb{E}\big[(\mu_t(X) - \mu_t)^2 + \mathbb{V}\mathrm{ar}(Y(t) \mid X) + \sigma_t^2(X^{\mathrm{C}})\{1 - \widetilde{\rho}_t^2(X^{\mathrm{C}})\}\big]\,.$$

□

Next, we show that the difference between $\widehat{\mu}_{t,\mathtt{R},\mathtt{CALM},\mathtt{Oracle}}$ and $\widehat{\mu}_{t,\mathtt{R},\mathtt{CALM}}$ is asymptotically negligible.

Lemma 4. *Under Assumptions 1–2 and 4, we have $\sqrt{n}\big(\widehat{\mu}_{t,\mathtt{R},\mathtt{CALM}} - \widehat{\mu}_{t,\mathtt{R},\mathtt{CALM},\mathtt{Oracle}}\big) = o_p(1)$.*

Proof. As in Lemma 2, it suffices to show that the decomposition

$$\sqrt{n}\big(\widehat{\mu}_{t,\texttt{CALM}}^{\mathcal{I}_1} - \widehat{\mu}_{t,\texttt{CALM,Oracle}}^{\mathcal{I}_1}\big) = \frac{\sqrt{n}}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\Big(1-\frac{\mathbf{1}\{T_i=t\}}{e_t(X_i)}\Big)\Big\{\big(\widehat{\mu}_t^{I_2}(X_i)-\widetilde{\mu}_t(X_i)\big) + \big(\widehat{\omega}_{t,\mathtt{R}}^{I_2}(X_i^{\mathtt{C}})-\omega_{t,\mathtt{R}}(X_i^{\mathtt{C}})\big)\big(Y_i^{\dagger}(t)-\widetilde{\mu}_t^{\dagger}(X_i)\big) - \big(\widehat{\mu}_t^{\dagger I_2}(X_i)-\widetilde{\mu}_t^{\dagger}(X_i)\big)\,\omega_{t,\mathtt{R}}(X_i^{\mathtt{C}}) - \big(\widehat{\mu}_t^{\dagger I_2}(X_i)-\widetilde{\mu}_t^{\dagger}(X_i)\big)\big(\widehat{\omega}_{t,\mathtt{R}}^{I_2}(X_i^{\mathtt{C}})-\omega_{t,\mathtt{R}}(X_i^{\mathtt{C}})\big)\Big\} \tag{1}$$

consists of four remainder terms, each of which is $o_p(1)$. By the same argument as in Lemma 2, the first and third terms are $o_p(1)$ under Assumption 4. For the second term, write

$$R_2 = \frac{\sqrt{n}}{|\mathcal{I}_1|}\sum_{x^{\mathtt{C}}=1}^{M}\big(\widehat{\omega}_{t,\mathtt{R}}^{I_2}(x^{\mathtt{C}})-\omega_{t,\mathtt{R}}(x^{\mathtt{C}})\big)\sum_{i\in\mathcal{I}_1,\,X_i^{\mathtt{C}}=x^{\mathtt{C}}}\Big(1-\frac{\mathbf{1}\{T_i=t\}}{e_t(X_i)}\Big)(Y_i^{\dagger}(t)-\widetilde{\mu}_t^{\dagger}(X_i)).$$

Furthermore, we introduce the intermediate estimator $\widetilde{\omega}_{t,\mathtt{R}}^{\mathcal{I}_2}(x^{\mathtt{C}})$ with the limiting (possibly misspecified) nuisance functions $\widetilde{\mu}_t, \widetilde{\mu}_t^{\dagger}$:

$$\widetilde{\omega}_{t,\mathtt{R}}^{\mathcal{I}_2}(x^{\mathtt{C}}) = \frac{\sum_{i\in\mathcal{I}_1(t;x^{\mathtt{C}})}\lambda_t(X_i)(Y_i-\widetilde{\mu}_t(X_i))(Y_i^{\dagger}(t)-\widetilde{\mu}_t^{\dagger}(X_i))}{\sum_{i\in\mathcal{I}_1(t;x^{\mathtt{C}})}\lambda_t(X_i)(Y_i^{\dagger}(t)-\widetilde{\mu}_t^{\dagger}(X_i))^2}, \quad \mathcal{I}_1(t;x^{\mathtt{C}}) = \{i\in\mathcal{I}_1 : T_i=t, X_i^{\mathtt{C}}=x^{\mathtt{C}}\}.$$

By the LLN and Assumptions 1-2, $\widetilde{\omega}_{t,\mathtt{R}}^{\mathcal{I}_2}(x^{\mathtt{C}}) = \omega_{t,\mathtt{R}}(x^{\mathtt{C}}) + O_p(n^{-1/2})$. Moreover, under Assumption 4 the difference between $\widehat{\omega}_{t,\mathtt{R}}^{\mathcal{I}_2}(x^{\mathtt{C}})$ and $\widetilde{\omega}_{t,\mathtt{R}}^{\mathcal{I}_2}(x^{\mathtt{C}})$ is $o_p(1)$, since their numerators and denominators differ only by replacing $\widehat{\mu}_t, \widehat{\mu}_t^{\dagger}$ with their oracle counterparts, and we have assumed implicitly that the denominators are bounded away from 0. Hence $\widehat{\omega}_{t,\mathtt{R}}^{\mathcal{I}_2}(x^{\mathtt{C}}) = \omega_{t,\mathtt{R}}(x^{\mathtt{C}}) + o_p(1)$. It follows by Cauchy-Schwarz that $R_2 = o_p(1)$, and the fourth term in (1) is handled similarly.

Therefore, all four terms vanish at $o_p(1)$ rate, which proves the lemma. □

Analogous to the proof of Theorem 1, Lemmas 3 and 4 together yield the proof of Corollary 1.

1.5 Proof of Theorem 2

In this section, we provide the proof of Theorem 2 from the main paper. Before delving into the theoretical analysis, we revisit the definitions and notations introduced in Section 4.2 to clarify the setup. Recall that the few-shot prediction generated by the LLM can be viewed as a function of the demonstrative subset S, the query covariates (X, Z), and the query treatment t, denoted by $f_{\theta,\mathtt{FS}}((X,Z),t,S)$. Here, the subset $S = (O_1,\ldots,O_m)$ is an ordered collection of observations drawn i.i.d. from the unknown distribution P (Assumption 2). By design, the demonstrative subset S and the query covariates (X, Z) are sampled from disjoint data folds (Steps $1'$ and $2'$), and are therefore independent. To account for the randomness in the demonstrative set, we have defined the expected few-shot prediction as $\bar{f}_{\theta,\mathtt{FS}}((x,z),t) = \mathbb{E}_S[f_{\theta,\mathtt{FS}}((X,Z),t,S) \mid (X,Z)=(x,z)]$ as the

expected few-shot prediction after averaging over the randomness of the demonstrative subset S, and the corresponding functions related to the LLM-based prediction have been defined as:

$$\mu^{\dagger}_{t,\texttt{FS}}(x) = \mathbb{E}\left(\bar{f}_{\theta,\texttt{FS}}((X,Z),t) \mid X = x\right), \quad \omega^{\dagger}_{t,\texttt{FS}}(x) = \frac{\mathbb{C}\text{ov}\left(Y, \bar{f}_{\theta,\texttt{FS}}((X,Z),t) \mid X = x\right)}{\mathbb{V}\text{ar}\left(\bar{f}_{\theta,\texttt{FS}}((X,Z),t) \mid X = x\right)},$$

by replacing the zero-shot prediction $Y^{\dagger}(t)$ with $\bar{f}_{\theta,\texttt{FS}}((X,Z),t)$.

Following a similar two-part strategy as in the proof of Theorem 1, we first establish the asymptotic properties of the *oracle* version of the CALM estimator under the few-shot setting. This oracle estimator is defined using the expected few-shot prediction $\bar{f}_{\theta,\texttt{FS}}((X,Z),t)$, along with the true nuisance functions $\mu_t(\cdot)$, $\mu^{\dagger}_{t,\texttt{FS}}(\cdot)$, and $\omega^{\dagger}_{t,\texttt{FS}}(\cdot)$.

$$\widehat{\mu}_{t,\texttt{CALM,FS,Oracle}} = \frac{1}{n}\sum_{i=1}^{n} \varphi_{t,\texttt{FS}}(Y_i, T_i, X_i, \bar{f}_{\theta,\texttt{FS}}((X_i,Z_i),t)),$$

where:

$$\begin{aligned}\varphi_{t,\texttt{FS}}(Y_i, T_i, X_i, \bar{f}_{\theta,\texttt{FS}}((X_i,Z_i),t)) =& \frac{\mathbf{1}\{T_i = t\}Y_i}{e_t(X_i)} + \left(1 - \frac{\mathbf{1}\{T_i = t\}}{e_t\left(X_i\right)}\right) \\ & \times \left(\mu_t\left(X_i\right) + \omega_{t,\texttt{FS}}\left(X_i\right)\left(\bar{f}_{\theta,\texttt{FS}}((X_i,Z_i),t) - \mu^{\dagger}_{t,\texttt{FS}}\left(X_i\right)\right)\right).\end{aligned}$$

Analogous to Lemma 1, we now state a corresponding result for the oracle version of the CALM estimator under the few-shot setting.

Lemma 5 (Asymptotic properties of the oracle CALM estimator, under the few-shot setting)**.** *Under Assumptions 1-2, the oracle version of the CALM estimator under the few-shot setting is a consistent and asymptotically normal estimator of μ_t. Specifically, we have $\widehat{\mu}_{t,\texttt{CALM,FS,Oracle}} \xrightarrow{p} \mu_t$, and:*

$$\sqrt{n}\left(\widehat{\mu}_{t,\texttt{CALM,FS,Oracle}} - \mu_t\right) \xrightarrow{d} \mathcal{N}\left(0, \mathsf{V}_{t,\texttt{CALM,FS,Oracle}}\right),$$

where the asymptotic variance is given by:

$$\mathsf{V}_{t,\texttt{CALM,FS,Oracle}} = \mathbb{E}\left[(\mu_t(X) - \mu_t)^2 + \frac{\mathbb{V}\text{ar}(Y(t) \mid X)}{e_t(X)}\left(1 - (1 - e_t(X))\rho^2_{t,\texttt{FS}}(X)\right)\right],$$

and $\rho_{t,\texttt{FS}}(x) = \mathbb{C}\text{orr}(Y(t), \bar{f}_{\theta,\texttt{FS}}((X,Z),t)) \mid X = x)$.

Recall from the proof of Theorem 1 that the zero-shot prediction $Y^{\dagger}(t)$ can be written as $f_\theta((X,Z),t)$. Comparing Lemma 1 and Lemma 5, the only difference lies in replacing the zero-shot prediction $f_\theta((X,Z),t)$ with the expected few-shot prediction $\bar{f}_{\theta,\texttt{FS}}((X,Z),t)$. Therefore, the proof of Lemma 5 follows exactly the same steps as in Lemma 1 and is omitted for brevity.

In the second part of the proof, we show that, under the construction of the robust few-shot counterfactual predictions in Step $2'$ and under Assumption 5, the difference between the

constructed estimator and its oracle counterpart in the few-shot setting is asymptotically negligible. In contrast to the proof of Theorem 1, the proof of Theorem 2 must account not only for the estimation error from cross-fitted nuisance functions, but also for the approximation error arising from estimating the expected few-shot prediction $\bar{f}_{\theta,\texttt{FS}}((x,z),t)$ via the resampling-and-aggregation strategy. We state the following lemma:

Lemma 6 (Asymptotic equivalence to oracle)**.** *Under Assumptions 1–2 and 5, the proposed CALM estimator under the few-shot setting $\widehat{\mu}_{t,\texttt{CALM,FS}}$ (Section 2.2) is asymptotically equivalent to the oracle CALM estimator under the few-shot setting $\widehat{\mu}_{t,\texttt{CALM,FS,Oracle}}$, in the sense that*

$$\sqrt{n}\big(\widehat{\mu}_{t,\texttt{CALM,FS}} - \widehat{\mu}_{t,\texttt{CALM,FS,Oracle}}\big) = o_p(1).$$

Lemmas 3 and 4 together complete the proof of Theorem 2.

1.6 Proof of Lemma 6

Proof. By random sample splitting (Step 1′), it suffices to consider $(k_1, k_2, k_3) = (1, 2, 3)$, and show that:

$$\sqrt{n}(\widehat{\mu}^{\mathcal{I}_3}_{t,\texttt{CALM,FS}} - \widehat{\mu}^{\mathcal{I}_3}_{t,\texttt{CALM,FS,Oracle}}) = o_p(1).$$

We decompose the difference as:

$$\begin{aligned}
\sqrt{n}\left(\widehat{\mu}^{\mathcal{I}_3}_{t,\texttt{CALM}} - \widehat{\mu}^{\mathcal{I}_3}_{t,\texttt{CALM,Oracle}}\right) = \frac{\sqrt{n}}{|\mathcal{I}_3|}\sum_{i\in\mathcal{I}_3}\left(1 - \frac{\mathbf{1}\{T_i = t\}}{e_t(X_i)}\right)\Bigg\{ & \left(\widehat{\mu}^{\mathcal{I}_2}_t(X_i) - \mu_t(X_i)\right) \\
& + \left(\widehat{\omega}^{\mathcal{I}_2}_{t,\texttt{FS}}(X_i) - \omega_{t,\texttt{FS}}(X_i)\right)\left(\bar{f}_{\theta,\texttt{FS}}((X_i,Z_i),t) - \mu^{\dagger}_{t,\texttt{FS}}(X_i)\right) \\
& + \left(Y^{\dagger}_{i,\texttt{FS}}(t;\mathcal{I}_1) - \bar{f}_{\theta,\texttt{FS}}((X_i,Z_i),t)\right)\omega_{t,\texttt{FS}}(X_i) \\
& + \left(Y^{\dagger}_{i,\texttt{FS}}(t;\mathcal{I}_1) - \bar{f}_{\theta,\texttt{FS}}((X_i,Z_i),t)\right)\left(\widehat{\omega}^{\mathcal{I}_2}_{t,\texttt{FS}}(X_i) - \omega_{t,\texttt{FS}}(X_i)\right) \\
& - \left(\widehat{\mu}^{\dagger\mathcal{I}_2}_{t,\texttt{FS}}(X_i) - \mu^{\dagger}_{t,\texttt{FS}}(X_i)\right)\omega_{t,\texttt{FS}}(X_i) \\
& - \left(\widehat{\mu}^{\dagger\mathcal{I}_2}_{t,\texttt{FS}}(X_i) - \mu^{\dagger}_{t,\texttt{FS}}(X_i)\right)\left(\widehat{\omega}^{\mathcal{I}_2}_{t,\texttt{FS}}(X_i) - \omega_{t,\texttt{FS}}(X_i)\right)\Bigg\}.
\end{aligned} \tag{2}$$

To bound the remaining term, it suffices to show that for each $i \in \mathcal{I}_3$,

$$Y^{\dagger}_{i,\texttt{FS}}(t;\mathcal{I}_1) - \bar{f}_{\theta,\texttt{FS}}((X_i,Z_i),t) = o_p(1).$$

Recall that $Y^{\dagger}_{i,\texttt{FS}}(t;\mathcal{I}_1) = \frac{1}{B}\sum_{b=1}^{B} f_{\theta,\texttt{FS}}\big((X_i,Z_i),t,\mathcal{S}^{\dagger}_b(\mathcal{I}_1)\big)$ for $i \notin \mathcal{I}_1$. Thus, it suffices to show that as $B \to \infty$, for each fixed (x,z) in the support of (X,Z),

$$\frac{1}{B}\sum_{b=1}^{B} f_{\theta,\texttt{FS}}\big((x,z),t,\mathcal{S}^{\dagger}_b(\mathcal{I}_1)\big) - \bar{f}_{\theta,\texttt{FS}}((x,z),t) = o_p(1). \tag{3}$$

To simplify notation, we temporarily omit the fixed (x, z) and the treatment t, and focus solely on the randomness induced by demonstrative subset sampling. Let $\mathcal{S} = (O_1, \ldots, O_m)$ denote an ordered subset of size m, drawn without replacement from the data in $\mathcal{I}_1$. Define the function g such that $g(O_1, \ldots, O_m) := f_{\theta,\texttt{FS}}\big((x, z), t, (O_1, \ldots, O_m)\big)$, which is deterministic but potentially asymmetric (i.e., order-sensitive) in its arguments. Let each resampled subset $\mathcal{S}_b^\dagger(\mathcal{I}_1)$ be expressed as $(O_{1,b}, \ldots, O_{m,b})$. Then, Equation (3) can be rewritten as:

$$\frac{1}{B}\sum_{b=1}^{B} g(O_{1,b}, \ldots, O_{m,b}) - \mathbb{E}[g(O_1, \ldots, O_m)] = o_p(1).$$

To prove this, we invoke one-sample U-statistics theory [10, 15]. Define the U-statistic of order m with asymmetric kernel as:

$$U_{n_1}(g) := \frac{1}{\prod_{i=0}^{m-1}(n_1 - i)} \sum_{(\mathcal{I}_1,\ldots,i_m)\in J_{m,n_1}} g(O_{\mathcal{I}_1}, \ldots, O_{i_m}),$$

where $n_1 = |\mathcal{I}_1|$, and J_{m,n_1} is the set of m-tuples from $\mathcal{I}_1$ with distinct entries. We decompose the difference as follows:

$$\begin{aligned}
&\frac{1}{B}\sum_{b=1}^{B} g(O_{1,b}, \ldots, O_{m,b}) - \mathbb{E}[g(O_1, \ldots, O_m)] \\
&= \underbrace{\Big(\frac{1}{B}\sum_{b=1}^{B} g(O_{1,b}, \ldots, O_{m,b}) - U_{n_1}(g)\Big)}_{\text{(I)}} + \underbrace{\big(U_{n_1}(g) - \mathbb{E}[g(O_1, \ldots, O_m)]\big)}_{\text{(II)}} \\
&= o_p(1) + O_p(n_1^{-1/2}) = o_p(1),
\end{aligned}$$

where (I) is the re-sampling error that vanishes as $B \to \infty$, and the second term (II) can be bounded by employing standard U-statistic convergence rate $O_p(n_1^{-1/2})$ for fixed m [15]. This verifies Equation (3). To complete the proof of Lemma 6, we can simply apply Assumption 5 and follow analogous arguments as used in the proof of Lemma 2 to verify Equation (2). □

1.7 Proof of Corollary 2

We now provide the proof of Corollary 2 from the main paper. Analogous to the proof of Lemma 2 in Section 1.3, it can be shown that under the assumptions of Corollary 2, the effect of estimated nuisance functions $\widehat{\mu}_t^{\mathcal{I}_\ell}(\cdot)$, $\widehat{\mu}_{t'}^{\mathcal{I}_\ell}(\cdot)$, $\widehat{\mu}_t^{\dagger,\mathcal{I}_\ell}(\cdot)$, $\widehat{\mu}_{t'}^{\dagger,\mathcal{I}_\ell}(\cdot)$, and $\widehat{\boldsymbol{\omega}}_{\texttt{ATE}}^{\mathcal{I}_\ell}(\cdot)$ on the asymptotics of the CALM estimator for the ATE, $\widehat{\tau}_{t,t',\texttt{CALM}}$, is of higher order and negligible. Hence, to prove Corollary 2 it suffices to study the asymptotic properties of the oracle CALM estimator with true nuisance

functions $\mu_t(\cdot)$, $\mu_{t'}(\cdot)$, $\mu_t^\dagger(\cdot)$, $\mu_{t'}^\dagger(\cdot)$, and $\boldsymbol{\omega}_{\texttt{ATE}}(\cdot)$:

$$\widehat{\tau}_{t,t',\texttt{CALM},\texttt{oracle}} = \frac{1}{n}\sum_{i=1}^{n} \varphi_{t,\texttt{CALM},\texttt{ATE}}(Y_i, T_i, X_i, Y_i^\dagger(t)) - \varphi_{t',\texttt{CALM},\texttt{ATE}}(Y_i, T_i, X_i, Y_i^\dagger(t')),$$

where:

$$\varphi_{t,\texttt{CALM},\texttt{ATE}}(Y_i, T_i, X_i, Y_i^\dagger(t)) = \frac{\mathbf{1}\{T_i = t\}Y_i}{e_t(X_i)} + \left(1 - \frac{\mathbf{1}\{T_i = t\}}{e_t\left(X_i\right)}\right)\left(\mu_t\left(X_i\right) + \omega_{t,\texttt{ATE}}\left(X_i\right)\left(Y_i^\dagger(t) - \mu_t^\dagger\left(X_i\right)\right)\right).$$

The consistency and asymptotic normality of $\widehat{\tau}_{t,t',\texttt{CALM},\texttt{oracle}}$ follow directly from the proof of Lemma 1 (Section 1.2). The main challenge is the computation of its asymptotic variance, which is more complicated than in Theorem 1 because of the correlation between the influence functions for the two arms. Specifically, the asymptotic variance is:

$$\begin{aligned}
\mathsf{V}_{t,t',\texttt{CALM}} &= \mathbb{E}\Big[\big(\varphi_{t,\texttt{CALM},\texttt{ATE}} - \varphi_{t',\texttt{CALM},\texttt{ATE}} - \tau_{t,t'}\big)^2\Big] \\
&= \mathbb{E}\Bigg\{\bigg[\mu_t(X) + \mu_{t'}(X) - \tau_{t,t'} + \frac{\mathbf{1}\{T=t\}}{e_t(X)}\big(Y - \mu_t(X)\big) + \Big(1 - \frac{\mathbf{1}\{T=t\}}{e_t(X)}\Big)\omega_{t,\texttt{ATE}}(X)\big(Y^\dagger(t) - \mu_t^\dagger(X)\big) \\
&\quad - \frac{\mathbf{1}\{T=t'\}}{e_{t'}(X)}\big(Y - \mu_{t'}(X)\big) - \Big(1 - \frac{\mathbf{1}\{T=t'\}}{e_{t'}(X)}\Big)\omega_{t',\texttt{ATE}}(X)\big(Y^\dagger(t') - \mu_{t'}^\dagger(X)\big)\bigg]^2\Bigg\}.
\end{aligned}$$

This can be decomposed as

$$\mathsf{V}_{t,t',\texttt{CALM}} = \mathbb{E}\big[(\mu_t(X) - \mu_{t'}(X) - \tau_{t,t'})^2\big] + \underbrace{\mathbb{E}\big[\Delta_t(X,T,Y,Y^\dagger(t))^2\big]}_{\text{(I)}} + \underbrace{\mathbb{E}\big[\Delta_{t'}(X,T,Y,Y^\dagger(t'))^2\big]}_{\text{(II)}} - 2\underbrace{\mathbb{E}\big[\Delta_t\Delta_{t'}\big]}_{\text{(III)}},$$

where

$$\begin{aligned}
\Delta_t(X,T,Y,Y^\dagger(t)) &:= \frac{\mathbf{1}\{T=t\}}{e_t(X)}\big(Y - \mu_t(X)\big) + \Big(1 - \frac{\mathbf{1}\{T=t\}}{e_t(X)}\Big)\omega_{t,\texttt{ATE}}(X)\big(Y^\dagger(t) - \mu_t^\dagger(X)\big), \\
\Delta_{t'}(X,T,Y,Y^\dagger(t')) &:= \frac{\mathbf{1}\{T=t'\}}{e_{t'}(X)}\big(Y - \mu_{t'}(X)\big) + \Big(1 - \frac{\mathbf{1}\{T=t'\}}{e_{t'}(X)}\Big)\omega_{t',\texttt{ATE}}(X)\big(Y^\dagger(t') - \mu_{t'}^\dagger(X)\big).
\end{aligned}$$

After standard calculation, we have:

$$\text{(I)} = \mathbb{E}\Bigg[\frac{1}{e_t(X)}\,\mathbb{V}\text{ar}(Y(t)\mid X) + \frac{1-e_t(X)}{e_t(X)}\,\omega^2_{t,\texttt{ATE}}(X)\,\mathbb{V}\text{ar}(Y^\dagger(t)\mid X) \\ -\ 2\frac{1-e_t(X)}{e_t(X)}\,\omega_{t,\texttt{ATE}}(X)\,\mathbb{C}\text{ov}(Y(t),Y^\dagger(t)\mid X)\Bigg],$$

$$\text{(II)} = \mathbb{E}\Bigg[\frac{1}{e_{t'}(X)}\,\mathbb{V}\text{ar}(Y(t')\mid X) + \frac{1-e_{t'}(X)}{e_{t'}(X)}\,\omega^2_{t',\texttt{ATE}}(X)\,\mathbb{V}\text{ar}(Y^\dagger(t')\mid X) \\ -\ 2\frac{1-e_{t'}(X)}{e_{t'}(X)}\,\omega_{t',\texttt{ATE}}(X)\,\mathbb{C}\text{ov}(Y(t'),Y^\dagger(t')\mid X)\Bigg],$$

$$\text{(III)} = \mathbb{E}\Big[\omega_{t',\texttt{ATE}}(X)\,\mathbb{C}\text{ov}(Y^\dagger(t'),Y(t)\mid X) + \omega_{t,\texttt{ATE}}(X)\,\mathbb{C}\text{ov}(Y^\dagger(t),Y(t')\mid X) \\ -\ \omega_{t',\texttt{ATE}}(X)\,\omega_{t,\texttt{ATE}}(X)\,\mathbb{C}\text{ov}(Y^\dagger(t),Y^\dagger(t')\mid X)\Big].$$

Summing over the terms (I)–(III) and rearranging terms yield:

$$\mathsf{V}_{t,t',\texttt{CALM}} = \mathbb{E}\Bigg\{(\tau_{t,t'}(X)-\tau_{t,t'})^2 + \text{Var}(Y(t)-Y(t')\mid X) \\ +\mathbb{V}\text{ar}\left[\sqrt{\frac{1-e_t(X)}{e_t(X)}}\Big(Y(t)-\omega_{t,\texttt{ATE}}(X)Y^\dagger(t)\Big) + \sqrt{\frac{1-e_{t'}(X)}{e_{t'}(X)}}\Big(Y(t')-\omega_{t',\texttt{ATE}}(X)Y^\dagger(t')\Big)\,\middle|\, X\right]\Bigg\}.$$

To minimize $\mathsf{V}_{t,t',\texttt{CALM}}$ with respect to $\boldsymbol{\omega}_{\texttt{ATE}}(\cdot)$, define $V = \big(\sqrt{\frac{1-e_t(X)}{e_t(X)}}Y^\dagger(t), \sqrt{\frac{1-e_{t'}(X)}{e_{t'}(X)}}Y^\dagger(t')\big)'$, and $Z = \sqrt{\frac{1-e_t(X)}{e_t(X)}}Y(t) + \sqrt{\frac{1-e_{t'}(X)}{e_{t'}(X)}}Y(t')$, then we can rewrite $\mathsf{V}_{t,t',\texttt{CALM}}$ as:

$$\mathsf{V}_{t,t',\texttt{CALM}} = \mathbb{E}\Big\{(\tau_{t,t'}(X)-\tau_{t,t'})^2 + \text{Var}(Y(t)-Y(t')\mid X) + \mathbb{V}\text{ar}\big(Z-\boldsymbol{\omega}_{\texttt{ATE}}(X)'V\mid X\big)\Big\}.$$

Under the assumption that $\Sigma_V(x) \succ 0$ for all $x \in \mathcal{X}$, the minimizer $\boldsymbol{\omega}_{\texttt{ATE}}(X)$ is $\Sigma_V^{-1}(X)\mathbb{C}\text{ov}(V, Z \mid X)$, where $\Sigma_V(X)$ is the conditional variance of V given X, and $\mathbb{C}\text{ov}(V, Z \mid X)$ is the conditional covariance of V and Z given X. In this case, the asymptotic variance is given by:

$$\mathsf{V}_{t,t',\texttt{CALM}} = \mathbb{E}\Big[(\tau_{t,t'}(X)-\tau_{t,t'})^2 + \frac{\mathbb{V}\text{ar}(Y(t)\mid X)}{e_t(X)} + \frac{\mathbb{V}\text{ar}(Y(t')\mid X)}{e_{t'}(X)} - \mathbb{C}\text{ov}(V,Z\mid X)'\Sigma_V^{-1}(X)\mathbb{C}\text{ov}(V,Z\mid X)\Big].$$

1.8 Proof of Corollary 3

In this section, we provide the proof of Corollary 3 from the main paper. Similar to the proof of Corollary 2, we start by defining the oracle version of the CALM-based CATE estimator as

$$\widehat{\tau}_{t,t',\texttt{CALM},\texttt{Oracle}}(x) = \sum_{i=1}^n w_i(x;\mathbf{X})\left\{\varphi_{t,\texttt{CALM},\texttt{ATE}}(Y_i,T_i,X_i,Y_i^\dagger(t)) - \varphi_{t',\texttt{CALM},\texttt{ATE}}(Y_i,T_i,X_i,Y_i^\dagger(t'))\right\},$$

where the linear smoother weights $w_i(x;\mathbf{X})$ are kernel weights $\frac{K\left(\frac{X_i-x}{h}\right)}{\sum_{j=1}^n K\left(\frac{X_j-x}{h}\right)}$ with bandwidth h, and the influence functions $\varphi_{t,\texttt{CALM,ATE}}$ and $\varphi_{t',\texttt{CALM,ATE}}$ are defined in Section 1.7 of the Supplementary Materials.

Assumption S.1 (Regularity conditions for kernel regression estimator of CATE)**.**

1. ***(Smoothness)*** *The conditional effect $\tau_{t,t'}(x)$ has bounded continuous second derivatives in a neighborhood of x.*
2. ***(Kernel)*** *$K:\mathbb{R}^p\to\mathbb{R}$ is a bounded, symmetric kernel of order two, i.e.*
$$\int K(u)du=1,\quad \int uK(u)du=0,\quad \int uu^\top K(u)du=\mu_2(K)I_p,\quad 0<\mu_2(K)<\infty.$$
3. ***(Bandwidth)*** *The bandwidth h satisfies $h\to 0$, $nh^p\to\infty$, and $nh^{p+4}\to 0$.*
4. ***(Covariate distribution)*** *The density $f_X(x)$ is continuous and bounded away from 0.*
5. ***(Moments)*** $\mathbb{E}[(\varphi_{t,\texttt{CALM,ATE}}-\varphi_{t',\texttt{CALM,ATE}})^2\mid X=x]<\infty$.

Leveraging the standard theory of kernel regression estimators, we state the following result:

Lemma 7. *Under Assumptions 1–2 and S.1, for any fixed $x\in\mathcal{X}$, we have $\widehat{\tau}_{t,t',\texttt{CALM,Oracle}}(x)\xrightarrow{p}\tau_{t,t'}(x)$, and*
$$\sqrt{nh^p}\left(\widehat{\tau}_{t,t',\texttt{CALM,Oracle}}(x)-\tau_{t,t'}(x)\right)\xrightarrow{d}\mathcal{N}\left(0,\frac{\|K\|_2^2\,\mathsf{V}_{t,t',\texttt{CALM}}(x)}{f_X(x)}\right),$$
where $\|K\|_2^2=\int K(u)^2du$, and
$$\mathsf{V}_{t,t',\texttt{CALM}}(x)=\frac{\mathbb{V}\mathrm{ar}(Y(t)\mid X=x)}{e_t(x)}+\frac{\mathbb{V}\mathrm{ar}(Y(t')\mid X=x)}{e_{t'}(x)}-\mathbb{C}\mathrm{ov}(V,Z\mid X=x)^\top\Sigma_V^{-1}(x)\mathbb{C}\mathrm{ov}(V,Z\mid X=x).$$

Proof. We provide only a short proof of this lemma, as most arguments follow directly from the standard theory of kernel regression estimators [8, 17]. Under Assumptions 1-2 and S.1, a second-order Taylor expansion of $\tau_{t,t'}(\cdot)$ around x combined with the symmetry of the kernel shows that
$$\mathbb{E}[\widehat{\tau}_{t,t',\texttt{CALM,Oracle}}(x)]-\tau_{t,t'}(x)=\tfrac{1}{2}h^2\mu_2(K)\,\Delta\tau_{t,t'}(x)+o(h^2),$$
where $\Delta\tau_{t,t'}(x)=\sum_{j=1}^p\partial^2\tau_{t,t'}(x)/\partial x_j^2$ denotes the Laplacian of $\tau_{t,t'}(\cdot)$ at x.

Next, define the conditional variance of the difference of influence functions as $\sigma^2_{t,t',\texttt{CALM}}(x)=\mathbb{V}\mathrm{ar}(\varphi_{t,\texttt{CALM,ATE}}-\varphi_{t',\texttt{CALM,ATE}}\mid X=x)$. Standard kernel variance expansion then gives
$$\mathbb{V}\mathrm{ar}\big(\widehat{\tau}_{t,t',\texttt{CALM,Oracle}}(x)\big)=\frac{1}{nh^p}\,\frac{\|K\|_2^2}{f_X(x)}\,\sigma^2_{t,t',\texttt{CALM}}(x)+o\left(\frac{1}{nh^p}\right).$$

Combining the bias and variance expansions, we obtain the CLT:

$$\sqrt{nh^p}\Big(\widehat{\tau}_{t,t',\texttt{CALM},\texttt{Oracle}}(x) - \tau_{t,t'}(x) - \tfrac{1}{2}h^2\mu_2(K)\Delta\tau_{t,t'}(x)\Big) \xrightarrow{d} \mathcal{N}\left(0,\ \frac{\|K\|_2^2}{f_X(x)}\,\sigma^2_{t,t',\texttt{CALM}}(x)\right).$$

Finally, adapting the asymptotic variance computation from the proof of Corollary 2, we have

$$\begin{aligned}\sigma^2_{t,t',\texttt{CALM}}(x) &= \mathbb{E}\big[(\varphi_{t,\texttt{CALM},\texttt{ATE}} - \varphi_{t',\texttt{CALM},\texttt{ATE}} - \tau_{t,t'}(x))^2 \mid X = x\big] \\ &= \frac{\mathbb{V}\mathrm{ar}(Y(t) \mid X = x)}{e_t(x)} + \frac{\mathbb{V}\mathrm{ar}(Y(t') \mid X = x)}{e_{t'}(x)} - \mathbb{C}\mathrm{ov}(V, Z \mid X = x)^\top \Sigma_V^{-1}(x)\mathbb{C}\mathrm{ov}(V, Z \mid X = x).\end{aligned}$$

Since the bias is $O(h^2)$ and the variance is $O((nh^p)^{-1})$, both vanish under the bandwidth conditions in Assumption S.1(3), ensuring consistency. To perform valid inference, one may employ undersmoothing so that $\sqrt{nh^p}\,h^2 \to 0$, which removes the asymptotic bias. This completes the proof of Lemma 7.

□

Next, we show that the asymptotic difference between $\widehat{\tau}_{t,t',\texttt{CALM},\texttt{Oracle}}(x)$ and the proposed CALM-based CATE estimator in Section 3.2 of the main paper $\widehat{\tau}_{t,t',\texttt{CALM}}(x)$ is negligible in the kernel regression regime.

Lemma 8. *Suppose the assumptions of Corollary 2 hold. Under Assumption S.1, we have:*

$$\sqrt{nh^p}\big(\widehat{\tau}_{t,t',\texttt{CALM}}(x) - \widehat{\tau}_{t,t',\texttt{CALM},\texttt{Oracle}}(x)\big) = o_p(1).$$

Proof. By analogy to the proof of Lemma 2 in Section 1.3, and by symmetry between arms t and t', it suffices to show that the remainder term

$$R_3 = \sqrt{nh^p}\sum_{i\in\mathcal{I}_1} w_i(x;\mathbf{X})\left(1 - \frac{\mathbf{1}\{T_i = t\}}{e_t(X_i)}\right)\left(\widehat{\mu}_t^{I_2}(X_i) - \mu_t(X_i)\right)$$

satisfies $R_3 = o_p(1)$, where the kernel weights are $w_i(x;\mathbf{X}) = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^n K\left(\frac{X_j - x}{h}\right)}$. Conditionally on $\mathcal{I}_2$ and the observed structured covariates $\{X_i\}$, the variance of R_3 is:

$$\begin{aligned}&\mathbb{E}\left[\left(\sqrt{nh^p}\sum_{i\in\mathcal{I}_1} w_i(x;\mathbf{X})\Big(1 - \frac{\mathbf{1}\{T_i = t\}}{e_t(X_i)}\Big)\big(\widehat{\mu}_t^{\mathcal{I}_2}(X_i) - \mu_t(X_i)\big)\right)^2 \,\Big|\, \mathcal{I}_2, \{X_i\}\right] \\ &= nh^p\sum_{i\in\mathcal{I}_1} w_i(x;\mathbf{X})^2\,\frac{1 - e_t(X_i)}{e_t(X_i)}\big(\widehat{\mu}_t^{\mathcal{I}_2}(X_i) - \mu_t(X_i)\big)^2 \\ &\le C\cdot nh^p\sum_{i\in\mathcal{I}_1} w_i(x;\mathbf{X})^2\big(\widehat{\mu}_t^{\mathcal{I}_2}(X_i) - \mu_t(X_i)\big)^2\end{aligned}$$

$$= C \cdot nh^p \cdot \frac{\sum_{i\in\mathcal{I}_1} K\Big(\frac{X_i-x}{h}\Big)^2 \big(\widehat{\mu}_t^{\mathcal{I}_2}(X_i) - \mu_t(X_i)\big)^2}{\Big(\sum_{j=1}^n K\Big(\frac{X_j-x}{h}\Big)\Big)^2}$$
$$= C \cdot \frac{|\mathcal{I}_1|}{n} \cdot \frac{\frac{1}{|\mathcal{I}_1|h^p}\sum_{i\in\mathcal{I}_1} K\Big(\frac{X_i-x}{h}\Big)^2 \big(\widehat{\mu}_t^{\mathcal{I}_2}(X_i) - \mu_t(X_i)\big)^2}{\Big(\frac{1}{nh^p}\sum_{j=1}^n K\Big(\frac{X_j-x}{h}\Big)\Big)^2}.$$

For the denominator, by the LLN and Assumption S.1,

$$\frac{1}{nh^p}\sum_{j=1}^n K\left(\frac{X_j-x}{h}\right) \xrightarrow{p} \frac{1}{h^p}\int K\left(\frac{u-x}{h}\right) f_X(u)du = f_X(x) > 0,$$

where the limit follows from a change of variables and $\int K(u)\,du = 1$. For the numerator, since K is bounded with $\int K(u)^2 du < \infty$, by the dominated convergence theorem and Assumption 3 ($\|\widehat{\mu}_t^{\mathcal{I}_2} - \mu_t\|_{L^2} = o_p(1)$), we obtain

$$\frac{1}{|\mathcal{I}_1|h^p}\sum_{i\in\mathcal{I}_1} K\Big(\frac{X_i-x}{h}\Big)^2 \big(\widehat{\mu}_t^{\mathcal{I}_2}(X_i) - \mu_t(X_i)\big)^2 \xrightarrow{p} f_X(x)\int K(z)^2 dz \cdot \mathbb{E}\big[(\widehat{\mu}_t^{\mathcal{I}_2}(X) - \mu_t(X))^2\big] = 0.$$

By Slutsky's theorem, the variance of R_3 is therefore $o_p(1)$. By Chebyshev's inequality, $R_3 = o_p(1)$. Analogous arguments apply to the other remainder terms as in Lemma 2, completing the proof. □

Combining Lemmas 7 and 8 establishes Corollary 3.

2 Test of efficiency improvement

In this section, we provide additional details of the implementation of the efficiency improvement test described in Remark 1 of the main text, which assesses whether the conditional covariance $\gamma_t(x)$ is uniformly zero, without delving into the full theoretical details. Let $n_t = \#\{i : T_i = t\}$ denote the number of subjects assigned to treatment t, and let $f_{X,t}(x)$ denote the conditional density of X given $T = t$. Since $\widehat{\gamma}_t(x)$ is estimated using a kernel estimator, standard arguments [8, 17] yield the expansion:

$$\sqrt{n_t h^p}\,\big(\widehat{\gamma}_t(x) - \gamma_t(x)\big) = \frac{1}{\sqrt{n_t h^p}}\sum_{i:T_i=t} \psi_t(x; X_i, Y_i, Y_i^\dagger(t)) + \mathrm{Bias}_t(x) + R_t(x),$$

where $\mathrm{Bias}_t(x)$ is the kernel smoothing bias, $R_t(x)$ is a negligible remainder term, and:

$$\psi_t(x; X_i, Y_i, Y_i^\dagger(t)) = \frac{1}{f_{X,t}(x)} K\Big(\frac{X_i-x}{h}\Big)\big\{Y_i Y_i^\dagger(t) - \mu_t(X_i)Y_i^\dagger(t) - \mu_t^\dagger(X_i)Y_i + \mu_t(X_i)\mu_t^\dagger(X_i) - \gamma_t(X_i)\big\}.$$

With a properly chosen bandwidth h and regularity conditions, both $\mathrm{Bias}_t(x)$ and $R_t(x)$ vanish uniformly in x at sufficiently fast rates (see, for example, [2, 3, 5]). Under $\mathcal{H}_0$ where $\gamma_t(x) \equiv 0$, we

replace ψ_t by its plug-in estimate:

$$\widehat{\psi}_t(x; X_i, Y_i, Y_i^\dagger(t)) = \frac{1}{\widehat{f}_{X,t}(x)} K\Big(\tfrac{X_i - x}{h}\Big) \big\{ Y_i Y_i^\dagger(t) - \widehat{\mu}_t(X_i) Y_i^\dagger(t) - \widehat{\mu}_t^\dagger(X_i) Y_i + \widehat{\mu}_t(X_i)\widehat{\mu}_t^\dagger(X_i) \big\},$$

where $\widehat{f}_{X,t}(x) = \frac{1}{n_t h^p} \sum_{i:T_i=t} K((X_i - x)/h)$ and $\widehat{\mu}_t, \widehat{\mu}_t^\dagger$ are kernel estimators of $\mu_t, \mu_t^\dagger$. The covariance kernel is then estimated by

$$\widehat{\Sigma}_t(x, x') = \frac{1}{n_t h^p} \sum_{i:T_i=t} \widehat{\psi}_t(x; X_i, Y_i, Y_i^\dagger(t))\, \widehat{\psi}_t(x'; X_i, Y_i, Y_i^\dagger(t)), \qquad \widehat{\sigma}_t^2(x) = \widehat{\Sigma}_t(x, x),$$

and on a fine grid $\{x_1, \ldots, x_M\} \subset \mathcal{X}$ (with M possibly increasing with n), we form the correlation matrix

$$\widehat{R}_t = (\widehat{\rho}_t(x_j, x_k))_{j,k=1}^M, \qquad \widehat{\rho}_t(x_j, x_k) = \frac{\widehat{\Sigma}_t(x_j, x_k)}{\widehat{\sigma}_t(x_j)\widehat{\sigma}_t(x_k)}.$$

We then simulate Gaussian vectors $G \sim N(0, \widehat{R}_t)$ to approximate the distribution of $\sup_x |G(x)|$. The test statistic

$$T = \sup_{x \in \{x_1, \ldots, x_M\}} \left| \frac{\sqrt{n_t h^p}\, \widehat{\gamma}_t(x)}{\widehat{\sigma}_t(x)} \right|$$

leads to rejection of $\mathcal{H}_0$ whenever T_{n_t} exceeds the simulated $(1 - \alpha)$ quantile. As an alternative, the distribution may also be approximated using the multiplier bootstrap [4].

3 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.

4 Synthetic superpopulation generation

First, we preprocess the covariates in the original BRIGHTEN study dataset to prepare for the downstream synthetic data generation. Structured numeric covariates are median-imputed and standardized, while categorical covariates are one-hot encoded. Unstructured covariates, such as survey responses, are first normalized so that missing entries are mapped to the string "No response." We then embed each unstructured covariate separately using the pretrained sentence transformer `all-MiniLM-L6-v2` [16], applying padding, truncation, mean pooling with attention masks, and l_2 normalization to obtain dense text embeddings. We adopt `all-MiniLM-L6-v2` because it is a compact, contrastively trained transformer model that produces high-quality sentence embeddings with well-preserved semantic structure. Finally, embeddings from the two unstructured covariate columns are then combined with the processed structured covariates to form the final co-

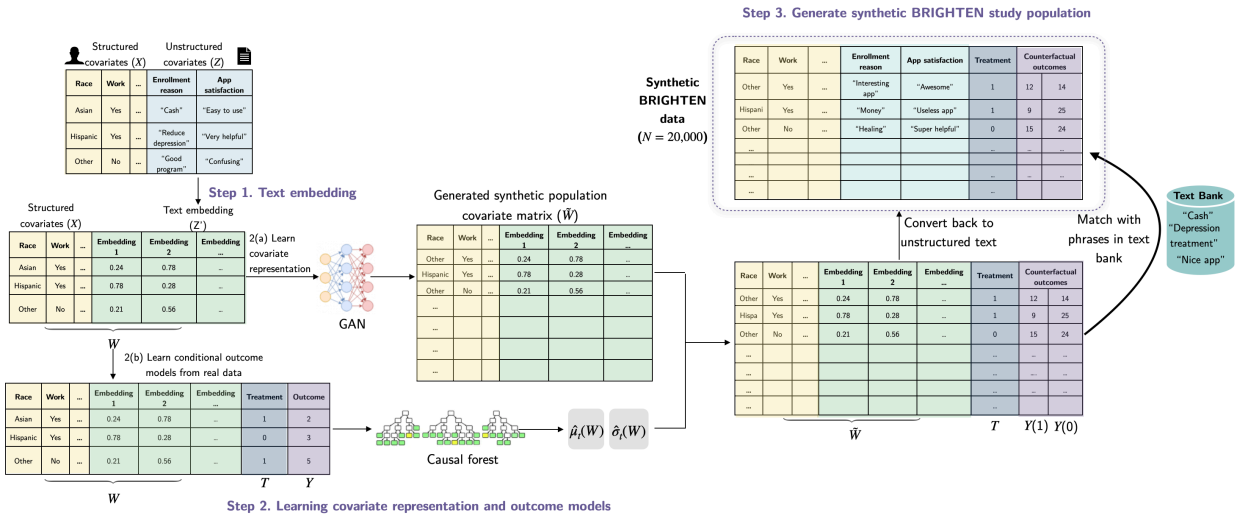


Figure 1: Illustration of the synthetic BRIGHTEN data generation.

variate representation, denoted as $W_i = (X_i, Z_i')$, where Z_i'' represents the embedding vectors.

Second, we learn both a generative model for the covariates and predictive models for the outcomes. We first model the joint covariate distribution $p(W)$ using a conditional Generative Adversarial Network (CTGAN) [18]. From the BRIGHTEN dataset, we construct a CTGAN training table that includes both the structured covariates and the two text-embedding blocks obtained from preprocessing. To improve sample efficiency and training stability, each 384-dimensional text block is compressed via principal component analysis (PCA) to 32 components.The structured covariates and PCA-compressed embeddings are then concatenated to form a unified training table, whose schema is validated with `Metadata` before fitting a `CTGANSynthesizer` [9]. This yields a generator G that models $\widehat{p}(W)$, which we use to draw synthetic covariates for downstream simulations. Next, using the embedded BRIGHTEN dataset $\{W_i, T_i, Y_i\}_{i=1}^n$, we fit causal forests to learn the conditional mean and conditional variance, $\mu_t(W) := \mathbb{E}[Y|W, T = t]$, $\sigma_t^2(W) := \mathbb{E}[(Y - \mu_t(W)^2)]$, $t \in \{0, 1\}$. using the `R` package `grf` [14]. We denote the fitted models as $\widehat{\mu}_t(W)$ and $\widehat{\sigma}_t^2(W)$. Finally, because the original outcome is the PHQ-9 score (0-27), we categorize it into five clinically standard levels of depression severity—minimal, mild, moderate, high, and severe—to simplify prediction tasks for LLMs. LLMs are more effective when predicting discrete, semantically meaningful categories rather than continuous numeric scores. After classifying participants into these categories, we recode the categories numerically from 1 to 5 to represent increasing severity.

Third, we generate the synthetic population of sample size $N = 20,000$. Using the generator learned in the second step, we first generate the synthetic covariate matrix, denoted as $\widetilde{W}_j = G(\nu_j)$, where $\nu_j \sim N(0, I_r)$, $j = 1, \ldots, N$. We then generate the treatment variable as $\widetilde{T}_j \sim \text{Bernoulli}(0.5)$, $j = 1, \ldots, N$, and the potential outcomes as $\widetilde{Y}_j(t) = \widehat{\mu}_t(\widetilde{W}_j) + \widehat{\sigma}_t(\widetilde{W}_j)\varepsilon_j$, $\varepsilon_j \sim N(0, 1)$, $t \in \{0, 1\}$. The observed synthetic outcomes are generated as $\widetilde{Y}_j = \widetilde{T}_j \cdot \widetilde{Y}_j(1) + (1 - \widetilde{T}_j) \cdot \widetilde{Y}_j(0)$, $j = 1, \ldots, N$. The true ATE based on the synthetic population is computed as $\tau := \frac{1}{N} \sum_{j=1}^N \widetilde{Y}_j(1) - \widetilde{Y}_j(0)$.

Lastly, because the generated covariates $\widetilde{W}_j$ consist of $(\widetilde{X}_j, \widetilde{Z}_j')$, where $\widetilde{Z}_j'$ is represented in the

numerical embedding space, we recover $\widetilde{Z}'_j$ back into unstructured text. To recover $\widetilde{Z}'_j$, we first construct a phrase bank for each free-text field (impression of the mobile app, reasons for enrolling) using responses from the original BRIGHTEN dataset after normalizing the raw texts. The surviving unique strings form the seed set $\mathcal{S}_k$ for unstructured covariate column k, where $k \in \{1, 2\}$. We then paraphrase each seed using pretrained LLMs–Pre-training with Extracted Gap-sentences for Abstractive Summarization(PEGASUS) and Text-to-Text Transfer Transformer(T5)– to increase lexical and semantic diversity. PEGASUS is a transformer model specialized for text summarization and paraphrase generation, trained with a gap-sentence generation objective that makes it highly effective for producing semantically faithful variations of input text [19]. T5 (Text-to-Text Transfer Transformer) is a general-purpose encoder-decoder transformer that frames all NLP tasks in a unified text-to-text format, enabling effective paraphrase generation among other tasks [12]. Candidate paraphrases are filtered by enforcing length constraints and require cosine similarity in the PCA-compressed embedding space to fall within a semantic band-pass window $[0.65, 0.98]$, relative to the original seed. This procedure yields the final phrase bank $\mathcal{B}_k$ for each unstructured covariate column, with size fixed at 50,000. All phrases in $\mathcal{B}_k$ are then re-embedded using the same encoder applied in the first step and projected through the saved PCA transformation. For each row of $\widetilde{Z}_i$, we perform top-1 nearest neighbor retrieval to identify the closest matching string, thereby recovering natural-language representations $\widetilde{Z}_i$ corresponding to the embedding vectors $\widetilde{Z}'_i$.

5 Additional simulation study details

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 2. Second, we provide a table of the coverage probabilities (Table 2) for all the methods discussed in the simulation section in the main manuscript. Third, we provide example prompts under zero-shot training, few-shot training, 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 2. Figure 2 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.

In what follows, we provide an example of the zero-shot learning prompt in Figure 3, the few-shot learning prompt in Figure 4, and example prompts under different prompt engineering techniques in Figure 5.

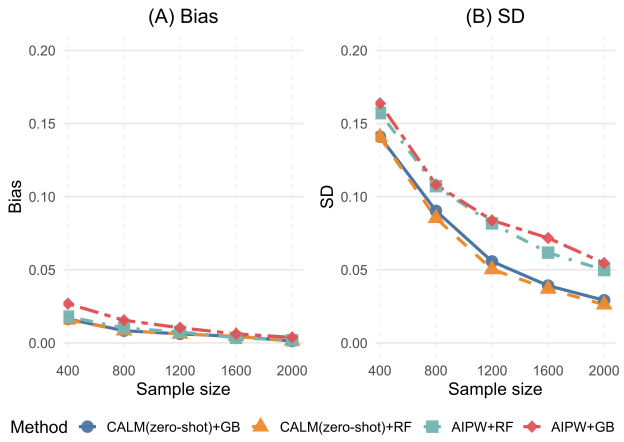


Figure 2: Absolute bias and standard deviation of ATE estimates for CALM (zero-shot) and AIPW-based methods across 300 Monte Carlo simulations. "CALM(zero-shot)+GB" and "CALM(zero-shot)+RF" refer to using gradient boosting and random forest for estimating the conditional mean model, respectively."AIPW+GB" and "AIPW+RF" refer to the standard AIPW estimator where the conditional mean model is estimated using gradient boosting and random forest, respectively.

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 [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 ["Help towards being happier"].

Figure 3: Example zero-shot learning prompt. The input example is synthetic and does not reflect real participant information from the study.

Trial Name	Brief Summary	Patient-Centered Outcomes	Effect Modifiers	Unstructured Pre-treatment Data
A. Precision Clinical Trial Recruitment to Promote Cancer Health Equity [7]	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 [13]	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 [11]	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 [6]	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.

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(0.02)
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 outcomes as covariates)	0.96(0.01)	0.94(0.01)
AIPW(few-shot outcomes as covariates, m=6)	0.92(0.01)	0.93(0.01)
AIPW(few-shot outcomes as covariates, m=10)	0.90(0.02)	0.92(0.02)
AIPW(few-shot outcomes as covariates, m=14)	0.89(0.02)	0.92(0.01)

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 [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 ["Help towards being happier"]. The participant has [Mild] level of depression.

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], [17] 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 ["Help me understand how i cope better"]. The participant has [Moderate] level of depression.

Figure 4: 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 "The app is truly useful for improving my mental health", 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: "The app is not very useful. 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 ["Help towards being happier"].

Figure 5: 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] Cattaneo, M. D., Chandak, R., Jansson, M., and Ma, X. (2024a). "Boundary adaptive local polynomial conditional density estimators," *Bernoulli*, *30*(4), 3193–3223.

[3] Cattaneo, M. D., Jansson, M., and Ma, X. (2024b). "Local regression distribution estimators," *Journal of Econometrics*, *240*(2), 105074.

[4] Chernozhukov, V., Chetverikov, D., and Kato, K. (2013a). "Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors," *The Annals of Statistics*, *41*(6), 2786–2819.

[5] Chernozhukov, V., Lee, S., and Rosen, A. M. (2013b). "Intersection bounds: Estimation and inference," *Econometrica*, *81*(2), 667–737.

[6] Guterman, E. L., Kiekhofer, 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.

[7] 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.

[8] Li, Q. and Racine, J. S. (2007). *Nonparametric Econometrics: Theory and Practice*, Princeton, NJ: Princeton University Press.

[9] Patki, N., Wedge, R., and Veeramachaneni, K. (2016). "The synthetic data vault," In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 399–410.

[10] Politis, D. N. and Romano, J. P. (1994). "Large sample confidence regions based on subsamples under minimal assumptions," *The Annals of Statistics*, *22*(4), 2031–2050.

[11] 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.

[12] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M. *et al.* (2020). "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, *21*(140), 1–67.

[13] 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.

[14] Tibshirani, J., Athey, S., Wager, S., Friedberg, R., Hadad, V., Miner, L. *et al.* (2024). *GRF: Generalized Random Forests*, R package version 2.3.2.

[15] van der Vaart, A. W. (2000). *Asymptotic Statistics*, Cambridge, UK: Cambridge University Press.

[16] Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. (2020). "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 5776–5788.

[17] Wasserman, L. (2006). *All of Nonparametric Statistics*, New York, NY: Springer.

[18] Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. (2019). "Modeling tabular data using conditional GAN," In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 7335–7345.

[19] Zhang, J., Zhao, Y., Saleh, M., and Liu, P. (2020). "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," In *Proceedings of the 37th International Conference on Machine Learning*, 11328–11339.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 11048–11064.

[38] OpenAI *et al.* (2023). "GPT-4 technical report," arXiv preprint arXiv:2303.08774.

[39] Patki, N., Wedge, R., and Veeramachaneni, K. (2016). "The synthetic data vault," In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 399–410.

[40] Poulet, P.-E., Tran, M., du Montcel, S. T., Dubois, B., Durrleman, S., and Jedynak, B. (2025). "Prediction-powered inference for clinical trials," medRxiv.

[41] 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.

[42] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S. *et al.* (2021). "Learning transferable visual models from natural language supervision," In *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763.

[43] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. *et al.* (2019). "Language models are unsupervised multitask learners," OpenAI blog.

[44] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M. *et al.* (2020). "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, *21*(140), 1–67.

[45] 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.

[46] Schick, T. and Schütze, H. (2021). "Exploiting cloze questions for few shot text classification and natural language inference," arXiv preprint arXiv:2001.07676.

[47] Song, Y., Wang, T., Cai, P., Mondal, S. K., and Sahoo, J. P. (2023). "A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities," *ACM Computing Surveys*, *55*(13s), 1–40.

[48] Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E. *et al.* (2022). "Galactica: A large language model for science," arXiv preprint arXiv:2211.09085.

[49] Tibshirani, J., Athey, S., Wager, S., Friedberg, R., Hadad, V., Miner, L. *et al.* (2024). *GRF: Generalized Random Forests*, R package version 2.3.2.

[50] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T. *et al.* (2023). "Llama: Open and efficient foundation language models," arXiv preprint arXiv:2302.13971.

[51] van der Vaart, A. W. (2000). *Asymptotic Statistics*, Cambridge, UK: Cambridge University Press.

[52] Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. (2020). "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 5776–5788.

[53] Wasserman, L. (2006). *All of Nonparametric Statistics*, New York, NY: Springer.

[54] Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B. *et al.* (2022). "Finetuned language models are zero-shot learners," arXiv preprint arXiv:2109.01652.

[55] Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. (2019). "Modeling tabular data using conditional GAN," In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 7335–7345.

[56] Zhang, B., Li, J., Hortaçsu, A., Ye, X., Chernozhukov, V., Ni, A., and Huang, E. (2025). "Agentic economic modeling," arXiv preprint arXiv:2510.25743.

[57] Zhang, J., Zhao, Y., Saleh, M., and Liu, P. (2020). "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," In *Proceedings of the 37th International Conference on Machine Learning*, 11328–11339.

[58] Zrnic, T. and Candès, E. J. (2024). "Cross-prediction-powered inference," *Proceedings of the National Academy of Sciences*, *121*(15), e2322083121.