MITIGATING HIDDEN CONFOUNDING BY PROGRESSIVE CONFOUNDER IMPUTATION VIA LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hidden confounding remains a central challenge in estimating treatment effects from observational data, as unobserved confounders can lead to biased causal estimates. Although recent work has explored the use of large language models (LLMs) for causal inference, existing approaches still rely on the unconfoundedness assumption and therefore cannot address hidden confounding. In this paper, we make the first attempt to mitigate hidden confounding using LLMs. We propose ProCI (Progressive Confounder Imputation), a framework that leverages two key strengths of LLMs: i) their ability to interpret rich textual information in causal benchmarks, which is often ignored by prior methods, and ii) their internalized world knowledge, which enables the generation of plausible latent confounders and supports counterfactual reasoning. ProCI alternates between imputing a plausible hidden confounder based on the semantics of observed variables, and validating the imputation by empirically assessing whether the new confounder can help restore conditional independence between treatment and outcome. To enhance robustness, ProCI adopts a distributional reasoning strategy rather than direct value imputation, preventing collapsed outputs. Extensive experiments demonstrate that ProCI uncovers meaningful hidden confounders and substantially improves treatment effect estimation across diverse datasets, base models, and LLM backbones.

1 INTRODUCTION

Estimating treatment effects from observational data is a central problem in causal inference, with widespread applications in decision making or high-stakes domains, such as healthcare (Prosperi et al., 2020; Grootendorst, 2007), social sciences (Gangl, 2010), and economics (Varian, 2016; Sun et al., 2024). Different from Randomized Controlled Trials (RCTs), in observational data, the non-random treatment assignments between treatment and control groups often lead to confounding bias (Shalit et al., 2017; Wang et al., 2023).

To estimate treatment effects in an unbiased manner, recent advances have focused on representation-based methods that aim to balance latent distributions between groups (Shalit et al., 2017; Wang et al., 2023; Assaad et al., 2021; Yao et al., 2018). More recently, with the emergence of large language models (LLMs) demonstrating strong capabilities in understanding real-world relationships, a growing body of work has begun to explore their potential for treatment effect estimation (Ma, 2025; Hüyük et al., 2025; Jin et al., 2023; Imai & Nakamura, 2026). For example, Jin et al. (2023) prompts LLMs using chain-of-thought reasoning to generate natural language-based causal conclusions, while Imai & Nakamura (2026) leverages LLMs to encode high-dimensional text treatments for mitigating confounding bias.

Despite these advances, existing LLM-based methods still mostly rely on the **unconfoundedness assumption**, which assumes that no unobserved variables can cause confounding bias. However, hidden confounding is common in real-world observational data (Chen et al., 2024a; Li et al., 2023). For example, in healthcare, treatment decisions may be influenced not only by observed clinical features such as age, but also by unobserved factors like socioeconomic status. Traditional methods for tackling hidden confounders can be summarized into three categories. i) *Sensitivity analysis* aims to quantify the impact of hidden confounding by deriving bounds on treatment effects (Robins et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

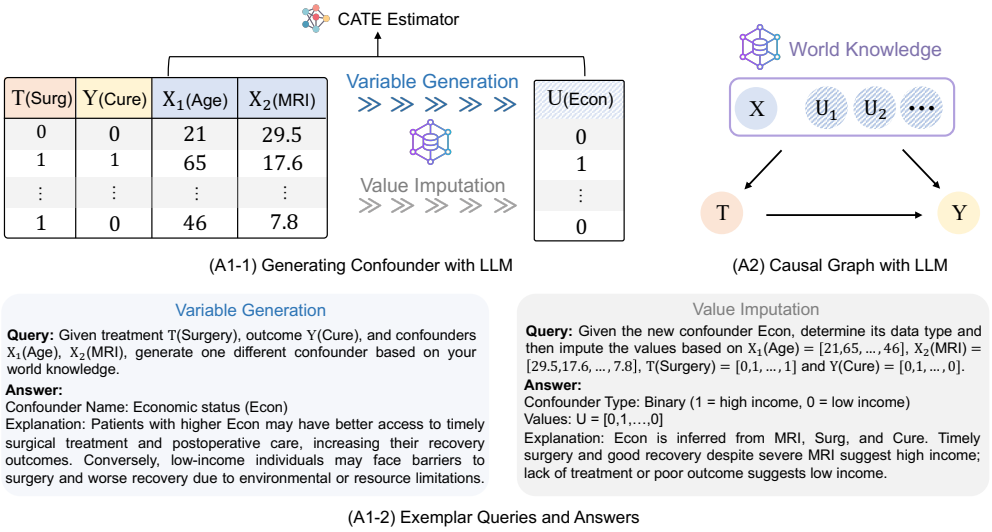


Figure 1: Motivating Illustration. **(A1-1)** The LLM takes structured data to generate a missing confounder and impute its individual-level values for CATE estimation. **(A1-2)** Exemplar queries and responses show how LLMs leverage semantic and world knowledge in variable generation and value imputation. **(A2)** From a causal perspective, LLMs embed latent confounders within their world knowledge, enabling counterfactual reasoning without requiring explicit access to them.

2000; Rosenbaum & Rubin, 1983), but typically relies on fixed and untestable assumptions about the confounding mechanism (Franks et al., 2018; Veitch & Zaveri, 2020). ii) *Auxiliary variable methods*, including instrumental variables and front-door adjustment, use external information or intermediate pathways to recover unbiased estimates (Li et al., 2023; Fulcher et al., 2020; Shah et al., 2023), but depend on strong structural assumptions that are hard to be verified (Imbens, 2014; Wu et al., 2022; Bellemare et al., 2024). iii) *RCT integration methods* combine randomized and observational data to correct for hidden confounding bias (Kallus et al., 2018; Hatt et al., 2022; Wu & Yang, 2022), but the high cost and limited availability of RCTs often restrict their practical utility. Taken together, these limitations show that hidden confounding remains an unsolved challenge, as current methods rely on conditions that rarely hold in real-world observational settings.

To fill this gap, we make the first attempt to leverage LLMs to alleviate hidden confounding in treatment effect estimation. Compared to traditional methods, LLMs offers two key advantages. First, LLMs have strong semantic reasoning capabilities that allow them to deal with the rich textual descriptions in the dataset. Second, LLMs, pretrained with large-scale corpora, have internalized broad causal and associative world knowledge. To leverage these two advantages in practice, we propose **ProCI** (Progressive Confounder Imputation), which leverages LLMs to iteratively uncover and validate for hidden confounders. ProCI alternates between two phases: (1) *confounder imputation*, where the LLM is encouraged to generate a plausible missing confounder based on the semantics of the observed variables, and to impute its concrete values; and (2) *unconfoundedness validation*, which builds on the LLM's counterfactual reasoning ability—by prompting the LLM to impute missing potential outcomes, we empirically test whether the generated confounders restore conditional independence between treatment and outcomes. Rather than directly predicting values, which we find empirically leads to collapsed or inconsistent outputs, ProCI adopts a distributional reasoning strategy. The above progressive process continues until the generated confounders pass the independence test, indicating that key sources of hidden confounding have been mitigated.

2 Preliminaries

2.1 Problem Setup

We consider the standard setup in binary treatment effect estimation. Let $T \in \{0, 1\}$ denote a binary treatment assignment, where $T = 1$ indicates receiving the treatment and $T = 0$ denotes the control group. Let $Y \in \mathbb{R}$ be the observed outcome, and $X \in \mathcal{X}$ be the observed covariates.

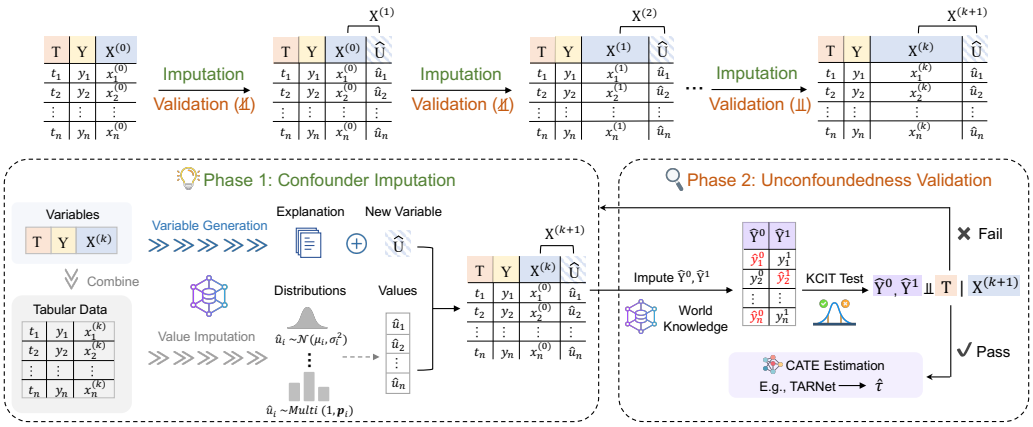


Figure 2: Overview of the ProCI framework. ProCI alternates between two phases: (1) **Confounder Imputation**, where the LLM generates a new confounder and imputes its instance-level values based on current variables. (2) **Unconfoundedness Validation**, where counterfactual outcomes are imputed and a KCIT is applied to assess whether unconfoundedness holds. The process repeats until the test passes, after which any standard estimator can be used for final CATE estimation.

The observational dataset consists of i.i.d. samples $\{(x_i, t_i, y_i)\}_{i=1}^{n} \sim \mathcal{D}$. Under the potential outcomes framework (Rubin, 2005), each unit is associated with two potential outcomes: Y^1 and Y^0, representing the outcomes the unit would receive under treatment and control, respectively. In practice, only one factual outcome Y is observed for each unit, depending on the assigned treatment. The individual-level treatment effect is defined as:

$$\tau(x) = \mathbb{E}[Y^1 - Y^0 \mid X = x], \tag{1}$$

which is commonly referred to as the *Conditional Average Treatment Effect* (CATE) (Abrevaya et al., 2015). It quantifies the expected difference in potential outcomes for a given covariate profile x, and serves as the primary estimand of interest in individual-level treatment effect estimation.

2.2 Identifiability

CATE estimation from observed data requires several identifiability assumptions (Pearl, 2009), including consistency, positivity, SUTVA, and unconfoundedness. A detailed introduction to these assumptions can be found in Appendix A. This paper focuses on the unconfoundedness assumption, which we formally define as follows:

Assumption 1 (Unconfoundedness)**.** The treatment assignment is independent of the potential outcomes, conditional on the observed covariates, i.e., $(Y^1, Y^0) \perp\!\!\!\perp T \mid X$.

When these assumptions hold, the CATE function in Eq. (1) becomes identifiable as follows:

$$\tau(x) = \mathbb{E}[Y \mid T = 1, X = x] - \mathbb{E}[Y \mid T = 0, X = x],$$

which can be directly estimated from a finite observational dataset $\mathcal{D}$ by learning an estimator $\hat{\tau}(x)$, *e.g.*, using CFRNet or TARNet (Shalit et al., 2017), that approximates the true CATE function $\tau(x)$.

Hidden Confounding Challenge. However, in practice, the assumption of unconfoundedness can be easily violated due to limited context (Jesson et al., 2021), where observed covariates X do not capture all relevant confounders that jointly affect treatment and outcome. In particular, treatment assignment may depend on hidden variables U that also influence the outcome Y. In such cases, the conditional independence assumption $(Y^1, Y^0) \perp\!\!\!\perp T \mid X$ is no longer valid, and is instead replaced by $(Y^1, Y^0) \not\perp\!\!\!\perp T \mid X$, but possibly $(Y^1, Y^0) \perp\!\!\!\perp T \mid (X, U)$ if U were additionally observed. This hidden confounding results in a systematic estimation bias: the learned estimator $\hat{\tau}(x)$, which assumes unconfoundedness given X, is biased with respect to the true CATE $\tau(x)$.

3 THE PROPOSED PROCI FRAMEWORK

3.1 MOTIVATION

Motivated by this challenge, we leverage LLMs to mitigate hidden confounding, building on two key advantages.

Advantage 1: Semantic Utilization Beyond Tabular Data. Unlike traditional causal methods that operate solely on structured covariates, LLMs can leverage semantic information associated with (X, T, Y) to generate plausible latent confounders. As illustrated in Figure 1 (A1-1) and (A1-2), the LLM generates candidate confounders that are not explicitly recorded but are semantically and causally related to both treatment and outcome.

Then, we make the following modeling assumption.

Assumption 2. Let $\phi_{\text{LLM}}(X, T, Y)$ denote the latent representation implicitly used by the LLM. We assume that, conditioned on $(X, \phi_{\text{LLM}}(X, T, Y))$, the potential outcomes are approximately independent of the treatment:

$$(Y^0, Y^1) \perp\!\!\!\perp T \mid X, \phi_{\text{LLM}}(X, T, Y).$$

Advantage 2: Implicit Counterfactual Reasoning via World Knowledge. Under Assumption 2, conditioning on the information accessible to the LLM approximates conditioning on $(X, U_1, U_2, \dots)$, as illustrated in Figure 1 (A2). This can enable LLM-based counterfactual outcome imputation.

However, Assumption 2 may fail when hidden confounders are weakly connected to the observed semantic context. This motivates the need for an explicit mechanism to empirically assess whether the inferred confounding information is sufficient, which leads to the design of the ProCI framework.

3.2 OVERVIEW OF THE PROCI FRAMEWORK

We propose ProCI, a framework for mitigating hidden confounding in by leveraging the above LLM capabilities. As shown in Figure 2, ProCI alternates between two phases: *confounder imputation* and *unconfoundedness validation*.

Phase 1: Confounder Imputation. Given the observed dataset $\{X^{(0)}, T, Y\}$, where $X^{(0)}$ contains covariates, ProCI guides the LLM to generate a plausible confounder $\hat{U}$ based on the semantic relationships among $X^{(k)}$, T, and Y, along with a textual explanation. It then imputes unit-level values for $\hat{U}$ by determining its distribution type and parameters via LLM-guided reasoning. The imputed confounder is added to the dataset to form $X^{(k+1)}$, where k denotes the iteration step.

Phase 2: Unconfoundedness Validation. To assess whether the generated confounder $\hat{U}$ captures sufficient hidden bias, ProCI performs an imputation-based empirical unconfoundedness test. Using the LLM's world knowledge, ProCI first imputes counterfactual outcomes $\hat{Y}^0$ and $\hat{Y}^1$, and then applies a kernel-based conditional independence test (KCIT) to check if $(\hat{Y}^0, \hat{Y}^1) \perp\!\!\!\perp T \mid X^{(k+1)}$. If the test fails, ProCI returns to Phase 1 to generate an additional confounder; if it passes, the framework proceeds to CATE estimation using a standard estimator like TARNet.

3.3 CONFOUNDER IMPUTATION VIA LLMS

In this section, we propose to prompt LLMs to generate hidden confounders that are semantically grounded and causally relevant. Imputing a confounder involves two key steps: (1) generating a variable with a meaningful textual description, and (2) imputing its concrete values for each individual in the dataset. We describe these steps as follows.

Variable Generation. Given a dataset $\mathcal{D}_{\text{obs}}$ consisting of observed variables $\{X, T, Y\}$, where X denotes covariates, T the treatment, and Y the outcome, we first design a prompt function $\mathcal{P}_{\text{var}}$ to convert the textual descriptions of these variables into a natural language query. We then use this prompt to query the LLM for a candidate confounder variable $\hat{U}$ that is semantically associated with both T and Y, while being orthogonal to X:

$$\hat{U}, \hat{U}_{\text{exp}} = \text{LLM}(\mathcal{P}_{\text{var}}(X, T, Y)), \tag{2}$$

where $\hat{U}$ is the name or description of the generated confounder, and $\hat{U}_{\text{exp}}$ is the accompanying natural language explanation provided by the LLM. This explanation enhances the interpretability of the generation process and can serve as supporting evidence for the semantic plausibility of $\hat{U}$.

Value Imputation. After generating the confounder variable $\hat{U}$, we next impute its values for each individual to obtain an augmented dataset $\hat{\mathcal{D}}_{\text{obs}} = \{x_i, t_i, y_i, \hat{u}_i\}_{i=1}^{N}$. However, direct value imputation of $\hat{u}_i$ for each individual tends to produce collapsed outputs where many individuals receive similar or identical values. To address this issue, we decompose value imputation into distribution identification and parameter inference as below.

Distribution Identification. Intuitively, by leveraging its embedded world knowledge, the LLM can reliably identify the distribution type of a variable, for example, inferring that height follows a normal distribution, while birth month is approximately uniform in a community. Therefore, for the generated variable $\hat{U}$, we can confidently ask the LLM to identify a reasonable distribution type $\mathcal{F} \in \mathbb{F} = \{\mathcal{F}_z\}_{z=1}^{Z}$:

$$\mathcal{F} \sim \text{LLM}(\mathcal{P}_{\text{dist}}(X, T, Y, \hat{U})) \;\big|\; \mathcal{F} \in \mathbb{F}. \tag{3}$$

In practice, the candidate distribution types $\mathbb{F}$ include, but are not limited to, Gaussian and categorical distributions.

Parameter Inference. Once the distribution family $\mathcal{F}$ is selected, we request the LLM to infer individual-specific distribution parameters θ_i based on each individual's observed features. For example, $\theta_i = (\mu_i, \sigma_i)$ if $\mathcal{F}$ is a normal distribution. We then sample the imputed value $\hat{u}_i$ from the corresponding personalized distribution:

$$\theta_i = \text{LLM}(\mathcal{P}_{\text{param}}(x_i, t_i, y_i)), \quad \hat{u}_i \sim \mathcal{F}(\theta_i). \tag{4}$$

3.4 Imputation-based Unconfoundedness Validation

While LLMs can help to impute the confounder, it is still challenging to test the validity of the generated confounder. Intuitively, a crucial criterion is the unconfoundedness assumption (Assumption 1). However, in observational data, we can only observe one of two potential outcomes for each individual, making empirical testing of the unconfoundedness assumption seemingly infeasible.

To bridge this gap, we use LLMs to empirically impute the missing outcomes Y^0 and Y^1 as follows:

$$\hat{y}_i^{1-t_i} = \text{LLM}(\mathcal{P}_{\text{out}}(x_i, \hat{u}_i, t_i, y_i)). \tag{5}$$

Then, we construct the variables $\hat{Y}^0$ and $\hat{Y}^1$, where for each individual i, $\hat{y}_i^0 = (1 - t_i) \cdot y_i + t_i \cdot \hat{y}_i^{1-t_i}$, $\hat{y}_i^1 = t_i \cdot y_i + (1 - t_i) \cdot \hat{y}_i^{1-t_i}$.

Rationality Analysis. LLMs embed extensive world knowledge that potentially covers relevant hidden confounders, and can therefore be viewed as approximately unbiased counterfactual estimators. Conditioned on the full context accessible to the LLM, including both structured and unstructured information, the potential outcomes become independent of the treatment. Formally, we posit that the LLM implicitly conditions on a latent variable U^* such that $(Y^0, Y^1) \perp\!\!\!\perp T \mid X, U^*$, where U^* denotes a sufficient set of hidden confounders encoded in the LLM's knowledge.

After obtaining the imputed outcomes $\hat{Y}^0$ and $\hat{Y}^1$, we can statistically test the unconfoundedness assumption using a non-parametric independence test. Specifically, we adopt the Kernel-based Conditional Independence Test (KCIT), which is well-suited for high-dimensional conditioning sets (Pogodin et al., 2024). KCIT evaluates whether the treatment T is conditionally independent of the imputed outcomes $(\hat{Y}^0, \hat{Y}^1)$ given the observed and generated confounders X and $\hat{U}$:

$$\mathbb{I}\left(\texttt{KCIT}\big((\hat{Y}^0, \hat{Y}^1), T \mid X, \hat{U}\big) > \alpha\right) = 1, \tag{6}$$

where $\texttt{KCIT}(\cdot)$ returns a p-value, and α is a pre-defined significance level. The indicator function $\mathbb{I}(\cdot)$ outputs 1 if the null hypothesis of conditional independence is not rejected, and 0 otherwise.

Theoretically, we establish the asymptotic validity of using imputed counterfactual outcomes and LLM-generated confounders in conditional independence test as follows.

Theorem 1. *Under standard regularity conditions on the kernel function and the class of imputation distributions, the KCIT applied to imputed variables satisfies:*

$$\begin{aligned} &\texttt{KCIT}\big((\hat{Y}^0, \hat{Y}^1), T \mid X, \hat{U}\big) \\ &= \texttt{KCIT}\big((Y^0, Y^1), T \mid X, U\big) + o_p(1), \end{aligned}$$

where $o_p(1)$ denotes a term that converges to zero in probability as the sample size increases. Please kindly refer to Appendix C for a detailed proof of this theorem.

Here, Theorem 1 implies that applying the conditional independence test to LLM-imputed variables is asymptotically equivalent to applying it to the true, but unobserved, variables. This means that passing the KCIT with imputed counterfactual outcomes provides the same statistical evidence for unconfoundedness as in the ideal setting. Therefore, LLM-based imputation can be safely used as a surrogate for hidden confounders in our validation procedure.

3.5 Progressive Confounder Imputation Procedure

Up to now, we have discussed imputing a single confounder and testing its validity. However, in real-world scenarios, hidden confounding often arises from multiple factors, such as demographic and socioeconomic in healthcare (Prosperi et al., 2020; Grootendorst, 2007), many of which are difficult to observe. A naive approach of generating multiple confounders simultaneously may lead to redundancy or semantic overlap. To address this, we propose the complete ProCI framework, which incrementally generates confounders step by step, ensuring both diversity and sufficiency.

Formally, let $X^{(0)} = X$ denote the initial observed covariates. At each iteration k, we query the LLM using the current context $(X^{(k-1)}, T, Y)$ to produce a new confounder $\hat{U}$ and update X as:

$$\hat{U}, \hat{U}_{\text{exp}} = \text{LLM}(\mathcal{P}_{\text{var}}(X^{(k-1)}, T, Y)),\ X^{(k)} = [X^{(k-1)}, \hat{U}].$$

After each generation step, we apply the empirical unconfoundedness test described in Eq. (6) to assess whether the updated set $X^{(k)}$ sufficiently captures all relevant confounding information. The process terminates at the smallest iteration k such that $X^{(k)}$ passes the test:

$$\min \left\{ k \,\middle|\, \texttt{KCIT}\big((\hat{Y}^0, \hat{Y}^1), T \mid X^{(k)}\big) > \alpha \right\}. \tag{7}$$

To ensure termination, we impose a maximum iteration limit $K_{\max}$ in practice.

4 Experiments

In this section, we conduct experiments to evaluate the effectiveness of the ProCI framework, guided by the following research questions: **[RQ1]** Can imputed confounders from different LLMs improve treatment effect estimation? **[RQ2]** Can ProCI benefit from domain-adaptive LLM routing? **[RQ3]** Do the imputed confounders capture additional confounding information beyond observed covariates? **[RQ4]** Is ProCI robust to varying levels of hidden confounding? **[RQ5]** What is the contribution of each ProCI component?

4.1 Experimental Setup

Datasets. We evaluate on two commonly-used causal benchmarks: Twins (Almond et al., 2005) and Jobs (LaLonde, 1986). **Twins** contains 8,244 twin pairs from U.S. birth records, with treatment indicating the heavier twin and outcome measuring one-year mortality. The dataset includes 50 demographic and birth-related covariates. Selection bias is introduced following (Louizos et al., 2017). **Jobs** studies the effect of a job training program on employment status. It includes 297 treated individuals, 425 randomized controls, and 2,490 observational controls, with 7 covariates describing demographic and economic features. Twins is evaluated using ϵ_{PEHE} and ϵ_{ATE} (Hill, 2011), while Jobs uses ϵ_{ATT} and policy risk ($\mathcal{R}_{pol}$) (Shalit et al., 2017) for evaluation. Please refer to Appendix D for more experimental details.

Base Models. Since ProCI is model-agnostic, we evaluate it with: (i) meta-learners (**S-Learner** (Künzel et al., 2019)); (ii) matching-based methods (**PSM** (Caliendo & Kopeinig, 2008));

Table 1: Overall comparison of treatment effect estimation performance on the Jobs and Twins datasets. For each base model, the best results across methods are highlighted.

Datasets	Jobs				Twins			
Test Types	In-sample		Out-sample		In-sample		Out-sample	
Methods	ϵ_{ATT}	$\mathcal{R}_{pol}$	ϵ_{ATT}	$\mathcal{R}_{pol}$	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}
S-Learner	$0.0491_{\pm 0.0011}$	$0.2289_{\pm 0.0008}$	$0.0876_{\pm 0.0018}$	$0.1678_{\pm 0.0002}$	$0.0131_{\pm 0.0020}$	$0.2527_{\pm 0.0012}$	$0.0037_{\pm 0.0024}$	$0.2819_{\pm 0.0001}$
+ProCI-4o	$0.0793_{\pm 0.0043}$	$0.2288_{\pm 0.0001}$	$0.0833_{\pm 0.0001}$	$0.1667_{\pm 0.0014}$	$0.0057_{\pm 0.0000}$	$0.2524_{\pm 0.0001}$	$0.0077_{\pm 0.0000}$	$0.2867_{\pm 0.0003}$
+ProCI-R1	$0.0216_{\pm 0.0002}$	$0.2324_{\pm 0.0000}$	$0.0712_{\pm 0.0001}$	$0.1745_{\pm 0.0002}$	$0.0066_{\pm 0.0008}$	$0.2596_{\pm 0.0001}$	$0.0028_{\pm 0.0009}$	$0.2817_{\pm 0.0004}$
PSM	$0.6197_{\pm 0.0000}$	$0.2707_{\pm 0.0000}$	$0.1259_{\pm 0.0015}$	$0.2192_{\pm 0.0013}$	$0.0457_{\pm 0.0006}$	$0.3399_{\pm 0.0007}$	$0.0840_{\pm 0.0000}$	$0.4027_{\pm 0.0001}$
+ProCI-4o	$0.6149_{\pm 0.0001}$	$0.2691_{\pm 0.0017}$	$0.1125_{\pm 0.0032}$	$0.2219_{\pm 0.0002}$	$0.0454_{\pm 0.0051}$	$0.3396_{\pm 0.0004}$	$0.0825_{\pm 0.0021}$	$0.4002_{\pm 0.0011}$
+ProCI-R1	$0.6245_{\pm 0.0000}$	$0.2633_{\pm 0.0000}$	$0.1292_{\pm 0.0041}$	$0.2163_{\pm 0.0131}$	$0.0454_{\pm 0.0040}$	$0.3399_{\pm 0.0027}$	$0.0849_{\pm 0.0009}$	$0.4033_{\pm 0.0125}$
TARNet	$0.0191_{\pm 0.0002}$	$0.2177_{\pm 0.0001}$	$0.1466_{\pm 0.0026}$	$0.2201_{\pm 0.0002}$	$0.0233_{\pm 0.0044}$	$0.2917_{\pm 0.0001}$	$0.0310_{\pm 0.0005}$	$0.3237_{\pm 0.0001}$
+ProCI-4o	$0.0424_{\pm 0.0023}$	$0.2167_{\pm 0.0022}$	$0.3223_{\pm 0.0098}$	$0.2150_{\pm 0.0202}$	$0.0144_{\pm 0.0218}$	$0.2729_{\pm 0.0017}$	$0.0185_{\pm 0.0001}$	$0.3107_{\pm 0.0206}$
+ProCI-R1	$0.0131_{\pm 0.0001}$	$0.2266_{\pm 0.0000}$	$0.0656_{\pm 0.0024}$	$0.2111_{\pm 0.0004}$	$0.0159_{\pm 0.0001}$	$0.2844_{\pm 0.0009}$	$0.0243_{\pm 0.0001}$	$0.3172_{\pm 0.0001}$
CFR-Wass	$0.0355_{\pm 0.0006}$	$0.2150_{\pm 0.0001}$	$0.1487_{\pm 0.0028}$	$0.2191_{\pm 0.0004}$	$0.0189_{\pm 0.0000}$	$0.2818_{\pm 0.0000}$	$0.0186_{\pm 0.0002}$	$0.3138_{\pm 0.0000}$
+ProCI-4o	$0.0300_{\pm 0.0009}$	$0.2085_{\pm 0.0001}$	$0.0402_{\pm 0.0003}$	$0.2151_{\pm 0.0005}$	$0.0094_{\pm 0.0001}$	$0.2729_{\pm 0.0001}$	$0.0178_{\pm 0.0001}$	$0.3110_{\pm 0.0001}$
+ProCI-R1	$0.0303_{\pm 0.0010}$	$0.2264_{\pm 0.0042}$	$0.1141_{\pm 0.0067}$	$0.2088_{\pm 0.0004}$	$0.0215_{\pm 0.0002}$	$0.2796_{\pm 0.0061}$	$0.0282_{\pm 0.0003}$	$0.3119_{\pm 0.0032}$
ESCFR	$0.0543_{\pm 0.0012}$	$0.2184_{\pm 0.0001}$	$0.2245_{\pm 0.0390}$	$0.2274_{\pm 0.0002}$	$0.0199_{\pm 0.0001}$	$0.2715_{\pm 0.0001}$	$0.0207_{\pm 0.0003}$	$0.3059_{\pm 0.0007}$
+ProCI-4o	$0.0369_{\pm 0.0004}$	$0.2174_{\pm 0.0001}$	$0.3225_{\pm 0.0622}$	$0.1891_{\pm 0.0002}$	$0.0191_{\pm 0.0001}$	$0.2700_{\pm 0.0032}$	$0.0297_{\pm 0.0029}$	$0.3045_{\pm 0.0001}$
+ProCI-R1	$0.0130_{\pm 0.0001}$	$0.2219_{\pm 0.0000}$	$0.1294_{\pm 0.0072}$	$0.2101_{\pm 0.0008}$	$0.0087_{\pm 0.0001}$	$0.2684_{\pm 0.0002}$	$0.0132_{\pm 0.0001}$	$0.3038_{\pm 0.0000}$

and (iii) representation learning models (**TARNet**, **CFR-Wass** (Shalit et al., 2017), and **ESCFR** (Wang et al., 2023)). For confounder generation, we use two advanced LLMs: **GPT-4o** (Hurst et al., 2024) and **DeepSeek-R1** (Guo et al., 2025).

4.2 [RQ1] Overall CATE Comparison

To answer RQ1, we apply our proposed ProCI framework using two LLMs, GPT-4o and DeepSeek-R1, denoted as ProCI-4o and ProCI-R1 respectively. We evaluate both variants across a diverse set of base CATE estimators. Each experiment is repeated five times to ensure stability.

Results. From Table 1, we can see: i) ProCI-augmented estimators consistently outperform base models in both in- and out-of-sample settings, indicating the effectiveness of imputed confounders. ii) All base models benefit from ProCI, with representation-based methods gaining the most, likely due to improved variable balancing. iii) ProCI-4o and ProCI-R1 achieve comparable overall performance across most settings. The qualitative analysis in Appendix E.4 further reveals that the two models exhibit complementary strengths, with each performing better in different types of scenarios, as illustrated by representative cases. Furthermore, the results from two open-source models, **LLaMA 3-8B** and **Qwen2.5-7B**, are provided in Appendix E.1. In addition, we conduct a targeted sanity-check experiment in Appendix E.3, where true hidden confounders are known by construction, to evaluate how well the generated confounders align with the removed covariates. At last, we further examine the effect of using different models for confounder generation and outcome imputation by decoupling these two stages, with results reported in Appendix E.5.

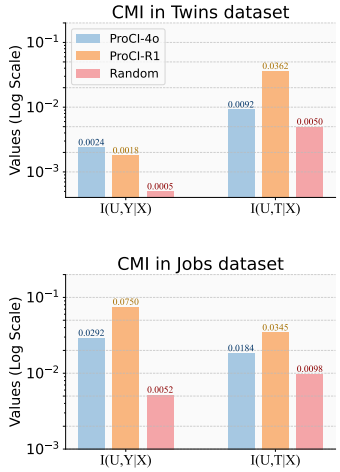
4.3 [RQ2] Domain-Adaptive LLM Routing

In this section, we evaluate domain-adaptive LLM routing for ProCI on the Twins dataset, which involves healthcare-related causal settings. Following the same protocol as previous experiments, ProCI is applied as a plug-in to four base estimators: S-Learner, PSM, TARNet, and CFR-Wass. We compare three configurations: **ProCI-General**, which uses GPT-4o; **ProCI-Wrong**, which intentionally applies a mismatched domain-specific LLM; and **ProCI-Routed**, which selects among domain-specific LLMs based on variable semantics, with fallback to GPT-4o. The domain-specific model pool includes **BioMistral-7B** (medical) (Labrak et al., 2024), **LawLLM-7B** (law) (Yue et al., 2024), and **DISC-FinLLM** (financial) (Chen et al., 2023). All other components of ProCI are kept identical across settings.

Results. The results are summarized in Table 2. Across all base estimators, ProCI-Routed achieves the best or comparable performance relative to ProCI-General, while ProCI-Wrong consistently degrades performance. For the Twins dataset, the routing mechanism primarily selects the BioMistral-

Table 2: Effect of Domain-Adaptive LLM Routing on ProCI Performance on the Twins Dataset. The best results are highlighted in bold, and the second-best results are underlined.

Base Model	S-Learner				PSM			
Test Types	In-sample		Out-sample		In-sample		Out-sample	
Methods	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}
Original	$0.0131_{\pm 0.0020}$	$0.2527_{\pm 0.0012}$	$\mathbf{0.0037}_{\pm 0.0024}$	$\mathbf{0.2819}_{\pm 0.0001}$	$\underline{0.0457}_{\pm 0.0006}$	$0.3399_{\pm 0.0007}$	$\underline{0.0840}_{\pm 0.0000}$	$0.4027_{\pm 0.0001}$
ProCI-General	$\underline{0.0057}_{\pm 0.0000}$	$\underline{0.2524}_{\pm 0.0001}$	$0.0077_{\pm 0.0000}$	$0.2867_{\pm 0.0003}$	$\mathbf{0.0454}_{\pm 0.0051}$	$\mathbf{0.3396}_{\pm 0.0004}$	$\mathbf{0.0825}_{\pm 0.0021}$	$\mathbf{0.4002}_{\pm 0.0011}$
ProCI-Wrong	$0.0072_{\pm 0.0031}$	$0.2549_{\pm 0.0002}$	$0.0066_{\pm 0.0011}$	$\underline{0.2845}_{\pm 0.0007}$	$0.0459_{\pm 0.0003}$	$0.3398_{\pm 0.0001}$	$0.0845_{\pm 0.0017}$	$0.4015_{\pm 0.0018}$
ProCI-Routed	$\mathbf{0.0051}_{\pm 0.0012}$	$\mathbf{0.2519}_{\pm 0.0001}$	$\underline{0.0064}_{\pm 0.0008}$	$0.2850_{\pm 0.0006}$	$0.0459_{\pm 0.0013}$	$\underline{0.3397}_{\pm 0.0002}$	$\underline{0.0840}_{\pm 0.0020}$	$\underline{0.4014}_{\pm 0.0009}$
Base Model	TARNet				CFR-Wass			
Test Types	In-sample		Out-sample		In-sample		Out-sample	
Methods	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}
Original	$0.0233_{\pm 0.0044}$	$0.2917_{\pm 0.0001}$	$0.0310_{\pm 0.0005}$	$0.3237_{\pm 0.0001}$	$0.0189_{\pm 0.0000}$	$0.2818_{\pm 0.0000}$	$0.0186_{\pm 0.0002}$	$0.3138_{\pm 0.0000}$
ProCI-General	$\underline{0.0144}_{\pm 0.0218}$	$\underline{0.2729}_{\pm 0.0017}$	$\underline{0.0185}_{\pm 0.0001}$	$\underline{0.3107}_{\pm 0.0206}$	$\underline{0.0094}_{\pm 0.0001}$	$\mathbf{0.2729}_{\pm 0.0001}$	$\underline{0.0178}_{\pm 0.0001}$	$\underline{0.3110}_{\pm 0.0001}$
ProCI-Wrong	$0.0192_{\pm 0.0025}$	$0.2821_{\pm 0.0095}$	$0.0296_{\pm 0.0009}$	$0.3282_{\pm 0.0043}$	$0.0096_{\pm 0.0005}$	$0.2815_{\pm 0.0005}$	$0.0193_{\pm 0.0017}$	$0.3143_{\pm 0.0024}$
ProCI-Routed	$\mathbf{0.0132}_{\pm 0.0018}$	$\mathbf{0.2619}_{\pm 0.0032}$	$\mathbf{0.0170}_{\pm 0.0003}$	$\mathbf{0.3105}_{\pm 0.0023}$	$\mathbf{0.0078}_{\pm 0.0009}$	$\underline{0.2772}_{\pm 0.0006}$	$\mathbf{0.0174}_{\pm 0.0011}$	$\mathbf{0.3106}_{\pm 0.0003}$



Confounder 1: Access to Childcare

Explanation: Individuals without reliable childcare may struggle to attend training projects regularly or maintain full-time employment, reducing their chances of program enrollment (treatment) and job placement (outcome).

Confounder 2: Criminal Record

Explanation: Individuals with a criminal record may be less likely to participate in job training programs due to eligibility restrictions (treatment). Simultaneously, a criminal record can reduce employment opportunities due to employer discrimination (outcome). This confounder is distinct from existing confounders like race or education.

Confounder 3: Employment History

Explanation: Individuals with strong work histories are more likely to join training to improve skills while those with weaker histories may perceive limited job prospects after training (treatment), affecting their chances of securing employment (outcome).

Figure 3: **(Left)** CMI measures dependence between LLM-generated confounders and treatment/outcome, compared to random confounders. **(Right)** Confounders generated in Jobs demonstrate their semantic relevance and impact on treatment/outcome.

7B, reflecting the healthcare-related semantics of the variables. This explains why ProCI-Routed often matches or slightly outperforms the general-purpose configuration. In contrast, forcing ProCI to use an incorrect domain-specific LLM (law or financial) leads to substantial performance drops across all base learners.

4.4 [RQ3] Information in Imputed Confounders

To answer RQ3, we evaluate whether the confounders imputed by ProCI contain additional information beyond observed covariates. We use conditional mutual information (CMI) to measure the dependency of imputed confounders U with treatment T and outcome Y given observed variables X, i.e., $I(U, T|X)$ and $I(U, Y|X)$. Higher CMI values indicate more confounding information.

Results. Figure 3 (left) compares imputed confounders from ProCI variants using GPT-4o and DeepSeek-R1 against random confounders generation baseline. The notably higher CMI values confirm that ProCI's confounders carry significant additional information about treatment and outcome beyond observed covariates. In addition, Figure 3 (right) shows three example confounders generated by DeepSeek-R1 on Jobs. These confounders have plausible links to treatment and outcomes and are distinct from observed covariates, highlighting ProCI's ability to uncover meaningful latent confounders omitted by traditional data. We also investigate the impact of LLM's temperature on confounder generation quality in Appendix E.2.

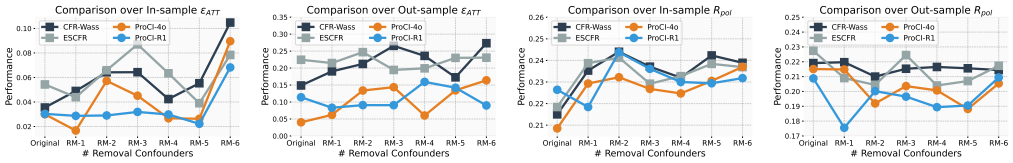


Figure 4: Robustness of ProCI in CATE estimation with varying hidden confounder removal.

4.5 [RQ4] Robustness of ProCI to Hidden Confounding

In this section, we test the robustness of CATE estimators under hidden confounders. Since hidden confounders are unobservable, we simulate their effects by progressively removing $\{0, 1, 2, 3, 4, 5, 6\}$ confounders from the dataset. This allows us to mimic the impact of latent confounders in a controlled manner. We experiment on Jobs and select CFR-Wass as base model.

Results. As shown in Figure 4, when the number of removed confounders increases, the performance of vanilla CATE estimators, such as CFR-Wass and ESCFR, degrades significantly across all metrics. This is expected, since these models are vulnerable to hidden confounder bias, leading to distorted CATE estimations. In contrast, our proposed ProCI framework maintains stable CATE estimation performance despite the progressive removal of confounders from the observed dataset. This robustness is primarily attributed to ProCI's progressive confounder generation capability, which introduces novel and informative confounders to effectively counteract the sparsity induced by the removal of observed confounders.

4.6 [RQ5] Ablation Studies of ProCI Components

In this experiment, we conduct ablation studies to evaluate ProCI components. In specific, we tailor three variants: i) ProCI w/o DR, which removes distributional reasoning; ii) ProCI w/o PI, which eliminates the progressive manner in generating confounders; and iii) ProCI w/o UT, which omits the unconfoundedness test. Experiments are conducted on Jobs using GPT-4o as the LLM backbone.

Table 3: Results of ablation studies.

	In-sample		Out-sample	
Methods	ϵ_{ATT}	$\mathcal{R}_{pol}$	ϵ_{ATT}	$\mathcal{R}_{pol}$
w/o DR	0.0327	0.2178	0.1599	0.2229
w/o PI	0.0308	0.2223	0.1020	0.2374
w/o UT	0.0315	0.2218	0.0890	0.2270
ProCI	**0.0300**	**0.2085**	**0.0402**	**0.2151**

Results. Table 3 shows that removing any component from ProCI reduces performance when using CFR-Wass as the base model. Specifically, removing distributional reasoning (w/o DR) leads to less robust estimations due to the corrupted values often generated by LLMs. Removing the progressive imputation process (w/o PI), where all confounders are generated at once, can introduce semantic overlap. Finally, omitting the unconfoundedness test (w/o UT) can include irrelevant confounders, further degrading the performance of ProCI. Together, these results suggest that robust value imputation, progressive variable generation, and empirical validation play complementary roles in improving treatment effect estimation.

5 Conclusion

In this work, we present ProCI, a principled framework that leverages LLMs to mitigate hidden confounding in treatment effect estimation by integrating data with semantic knowledge. Rather than assuming unconfoundedness a priori, ProCI adopts an iterative procedure that imputes candidate confounders and empirically validates their sufficiency. By augmenting observed data with inferred confounders, ProCI can be integrated with different treatment effect estimators, enabling semantic knowledge to complement their existing estimation procedures. Extensive experiments across diverse datasets, base estimators, and LLMs demonstrate consistent improvements in treatment effect estimation and robustness to varying confounding strengths. Beyond empirical performance, ProCI is able to generate semantically meaningful and causally plausible confounders, offering interpretable insights into potential sources of hidden bias. Overall, this work highlights the potential of using LLMs to complement, rather than replace, principled statistical causal inference.

AI USE STATEMENT

Generative AI assistance was used to inspect the existing references, organize the experimental protocol, help to implement and polish the manuscript. We have not used generative AI tools for the core intellectual contributions of this work, including the theoretical derivations and algorithm design. We have carefully reviewed and edited all AI-generated content to ensure accuracy and clarity.

REPRODUCIBILITY STATEMENT

To support reproducibility, we describe the proposed method and experimental setup in the main text. Further details needed to reproduce the experiments, including dataset descriptions, preprocessing procedures, model configurations, hyperparameter selection, training and evaluation protocols, and computational resources, are provided in the experimental sections and the appendix. We have also provided the code and data used in our experiments in the supplementary material.

REFERENCES

Sara Abdali, Anjali Parikh, Steve Lim, and Emre Kiciman. Extracting self-consistent causal insights from users feedback with llms and in-context learning. *arXiv preprint arXiv:2312.06820*, 2023.

Jason Abrevaya, Yu-Chin Hsu, and Robert P Lieli. Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33(4):485–505, 2015.

Douglas Almond, Kenneth Y Chay, and David S Lee. The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083, 2005.

Arash A Amini and Zahra S Razaee. Concentration of kernel matrices with application to kernel spectral clustering. *The Annals of Statistics*, 49(1):531–556, 2021.

Serge Assaad, Shuxi Zeng, Chenyang Tao, Shounak Datta, Nikhil Mehta, Ricardo Henao, Fan Li, and Lawrence Carin. Counterfactual representation learning with balancing weights. In *International Conference on Artificial Intelligence and Statistics*, pp. 1972–1980. PMLR, 2021.

Marc F. Bellemare, Jeffrey R. Bloem, and Noah Wexler. The paper of how: Estimating treatment effects using the front-door criterion*. *Oxford Bulletin of Economics and Statistics*, 2024.

Marco Caliendo and Sabine Kopeinig. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, 22(1):31–72, 2008.

Li Chen, Chunlin Li, Xiaotong Shen, and Wei Pan. Discovery and inference of a causal network with hidden confounding. *Journal of the American Statistical Association*, 119(548):2572–2584, 2024a.

Meiqi Chen, Yixin Cao, Yan Zhang, and Chaochao Lu. Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective. *arXiv preprint arXiv:2403.18346*, 2024b.

Sirui Chen, Bo Peng, Meiqi Chen, Ruiqi Wang, Mengying Xu, Xingyu Zeng, Rui Zhao, Shengjie Zhao, Yu Qiao, and Chaochao Lu. Causal evaluation of language models. *arXiv preprint arXiv:2405.00622*, 2024c.

Wei Chen, Qiushi Wang, Zefei Long, Xianyin Zhang, Zhongtian Lu, Bingxuan Li, Siyuan Wang, Jiarong Xu, Xiang Bai, Xuanjing Huang, et al. Disc-finllm: A chinese financial large language model based on multiple experts fine-tuning. *arXiv preprint arXiv:2310.15205*, 2023.

Nikita Dhawan, Leonardo Cotta, Karen Ullrich, Rahul G Krishnan, and Chris J Maddison. End-to-end causal effect estimation from unstructured natural language data. *Advances in Neural Information Processing Systems*, 37:77165–77199, 2024.

Alexander M. Franks, Alexander D'Amour, and Avi Feller. Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*, 115:1730 – 1746, 2018.

Isabel R Fulcher, Ilya Shpitser, Stella Marealle, and Eric J Tchetgen Tchetgen. Robust inference on population indirect causal effects: the generalized front door criterion. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):199–214, 2020.

Markus Gangl. Causal inference in sociological research. *Annual review of sociology*, 36(1):21–47, 2010.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Paul Grootendorst. A review of instrumental variables estimation of treatment effects in the applied health sciences. *Health Services and Outcomes Research Methodology*, 7:159–179, 2007.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Tobias Hatt, Jeroen Berrevoets, Alicia Curth, Stefan Feuerriegel, and Mihaela van der Schaar. Combining observational and randomized data for estimating heterogeneous treatment effects. *ArXiv*, abs/2202.12891, 2022.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Alihan Hüyük, Xinnuo Xu, Jacqueline Maasch, Aditya Nori, and Javier Hernandez. Reasoning elicitation in language models via counterfactual feedback. In *International Conference on Learning Representations*, volume 2025, pp. 76804–76832, 2025.

Kosuke Imai and Kentaro Nakamura. Causal inference with generative artificial intelligence: Application to texts as treatments. *Journal of the American Statistical Association*, pp. 1–27, 2026.

Guido Imbens. Instrumental variables: An econometrician's perspective. *Political Methods: Quantitative Methods eJournal*, 2014.

Andrew Jesson, Sören Mindermann, Yarin Gal, and Uri Shalit. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. In *International Conference on Machine Learning*, pp. 4829–4838. PMLR, 2021.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. Cladder: Assessing causal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:31038–31065, 2023.

Nathan Kallus, Aahlad Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. *ArXiv*, abs/1810.11646, 2018.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the association for computational linguistics: acl 2024*, pp. 5848–5864, 2024.

Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pp. 604–620, 1986.

Po-Han Lee, Yu-Cheng Lin, Chan-Tung Ku, Chan Hsu, Pei-Cing Huang, Ping-Hsun Wu, and Yihuang Kang. Llm-based agents for automated confounder discovery and subgroup analysis in causal inference. In *2025 IEEE International Conference on Information Reuse and Integration and Data Science (IRI)*, pp. 265–270. IEEE, 2025.

Haoxuan Li, Kunhan Wu, Chunyuan Zheng, Yanghao Xiao, Hao Wang, Zhi Geng, Fuli Feng, Xiangnan He, and Peng Wu. Removing hidden confounding in recommendation: a unified multi-task learning approach. *Advances in Neural Information Processing Systems*, 36:54614–54626, 2023.

Victoria Lin, Louis-Philippe Morency, and Eli Ben-Michael. Text-transport: Toward learning causal effects of natural language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1288–1304, 2023.

Victoria Lin, Eli Ben-Michael, and Louis-Philippe Morency. Optimizing language models for human preferences is a causal inference problem. *arXiv preprint arXiv:2402.14979*, 2024.

Chenxi Liu, Yongqiang Chen, Tongliang Liu, Mingming Gong, James Cheng, Bo Han, and Kun Zhang. Discovery of the hidden world with large language models. *Advances in Neural Information Processing Systems*, 37:102307–102365, 2024.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

Jing Ma. Causal inference with large language model: A survey. *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5901–5913, 2025.

Martina Miliani, Serena Auriemma, Alessandro Bondielli, Emmanuele Chersoni, Lucia Passaro, Irene Sucameli, and Alessandro Lenci. Explica: Evaluating explicit causal reasoning in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 17335–17355, 2025.

Nick Pawlowski, James Vaughan, Joel Jennings, and Cheng Zhang. Answering Causal Questions with Augmented LLMs. In *ICML 2023 Workshops: DeployableGenerativeAI*, 2023.

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

Roman Pogodin, Antonin Schrab, Yazhe Li, Danica J Sutherland, and Arthur Gretton. Practical kernel tests of conditional independence. *arXiv preprint arXiv:2402.13196*, 2024.

Mattia Prosperi, Yi Guo, Matt Sperrin, James S Koopman, Jae S Min, Xing He, Shannan Rich, Mo Wang, Iain E Buchan, and Jiang Bian. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, 2(7):369–375, 2020.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL `https://arxiv.org/abs/2412.15115`.

Richard W. Robins, Avshalom Caspi, and Terrie E. Moffitt. Two personalities, one relationship: both partners' personality traits shape the quality of their relationship. *Journal of personality and social psychology*, 79 2:251–9, 2000.

Paul R. Rosenbaum and Donald B. Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the royal statistical society series b-methodological*, 45:212–218, 1983.

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American statistical Association*, 100(469):322–331, 2005.

Abhin Shah, Karthikeyan Shanmugam, and Murat Kocaoglu. Front-door adjustment beyond markov equivalence with limited graph knowledge. *Advances in Neural Information Processing Systems*, 36:43800–43825, 2023.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pp. 3076–3085. PMLR, 2017.

Ivaxi Sheth, Bahare Fatemi, and Mario Fritz. Causalgraph2llm: Evaluating llms for causal queries. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2076–2098, 2025.

Zexu Sun, Hao Yang, Dugang Liu, Yunpeng Weng, Xing Tang, and Xiuqiang He. End-to-end cost-effective incentive recommendation under budget constraint with uplift modeling. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pp. 560–569, 2024.

Juanhe TJ Tan. Causal abstraction for chain-of-thought reasoning in arithmetic word problems. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 155–168, 2023.

Hal R Varian. Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences*, 113(27):7310–7315, 2016.

Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth Balasubramanian, and Amit Sharma. Causal order: The key to leveraging imperfect experts in causal inference. In *International Conference on Learning Representations*, volume 2025, pp. 95330–95371, 2025.

Victor Veitch and Anisha Zaveri. Sense and sensitivity analysis: Simple post-hoc analysis of bias due to unobserved confounding. *Advances in neural information processing systems*, 33:10999–11009, 2020.

Hao Wang, Jiajun Fan, Zhichao Chen, Haoxuan Li, Weiming Liu, Tianqiao Liu, Quanyu Dai, Yichao Wang, Zhenhua Dong, and Ruiming Tang. Optimal transport for treatment effect estimation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Anpeng Wu, Kun Kuang, B. Li, and Fei Wu. Instrumental variable regression with confounder balancing. In *International Conference on Machine Learning*, 2022.

Anpeng Wu, Kun Kuang, Ruoxuan Xiong, Bo Li, and Fei Wu. Stable estimation of heterogeneous treatment effects. In *International Conference on Machine Learning*, pp. 37496–37510. PMLR, 2023.

Lili Wu and Shu Yang. Integrative r-learner of heterogeneous treatment effects combining experimental and observational studies. In *Conference on Causal Learning and Reasoning*, pp. 904–926. PMLR, 2022.

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. *Advances in neural information processing systems*, 31, 2018.

Shengbin Yue, Shujun Liu, Yuxuan Zhou, Chenchen Shen, Siyuan Wang, Yao Xiao, Bingxuan Li, Yun Song, Xiaoyu Shen, Wei Chen, et al. Lawllm: Intelligent legal system with legal reasoning and verifiable retrieval. In *International Conference on Database Systems for Advanced Applications*, pp. 304–321. Springer, 2024.

Shitian Zhao, Zhuowan Li, Yadong Lu, Alan Yuille, and Yan Wang. Causal-cog: A causal-effect look at context generation for boosting multi-modal language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13342–13351, 2024.

A In-Depth Causal Inference Preliminaries

This section provides additional background on causal inference, which is particularly intended to assist readers who may be less familiar with the foundational concepts or technical aspects of treatment effect estimation from observational data.

We begin by introducing key definitions and assumptions underlying causal inference from observational data. For an individual characterized by covariates x, there are two potential outcomes: Y^1 if the individual receives treatment and Y^0 if assigned to control. The Conditional Average Treatment Effect (CATE) captures the expected difference in outcomes between these two scenarios:

Definition 1. The CATE for individuals with covariates x is defined as

$$\tau(x) = \mathbb{E}[Y^1 - Y^0 \mid X = x], \tag{8}$$

where X denotes the covariate variable, and Y^1, Y^0 are the potential outcomes under treatment and control, respectively.

Estimating CATE from observational data poses two primary challenges:

1. *Missing counterfactuals*: For each individual, we only observe the outcome corresponding to the assigned treatment. The unobserved counterfactual remains inaccessible.
2. *Selection bias*: Treatment assignment may depend on covariates related to the outcome, leading to systematic differences between treated and control groups.

To address these challenges, Pearl & Mackenzie (2018) proposed a two-stage framework. The first stage, *identification*, aims to express causal quantities, such as $\tau(x)$, in terms of observed data using assumptions and adjustment formulas. Identification is not always guaranteed, and depends on the following assumptions:

Assumption 3 (Ignorability / Unconfoundedness)**.** The treatment assignment is independent of the potential outcomes given the covariates, i.e., $(Y^1, Y^0) \perp\!\!\!\perp T \mid X = x$.

Assumption 4 (Positivity)**.** For any value x, both treatment and control must have non-zero probability, i.e., $0 < P(T = t \mid X = x) < 1$ for all $t \in \{0, 1\}$.

Assumption 5 (SUTVA)**.** The potential outcomes for any individual are unaffected by others' treatment assignments, and each treatment corresponds to a single well-defined outcome.

Assumption 6 (Consistency)**.** The observed outcome equals the potential outcome under the treatment actually received.

Once identification is established, the second stage, *estimation*, transforms the causal estimand into a statistical estimand that can be computed from the data:

$$\begin{aligned}\mathbb{E}[Y^1 - Y^0 \mid X = x] &= \mathbb{E}[Y^1 \mid X = x] - \mathbb{E}[Y^0 \mid X = x] \\ &\overset{(1)}{=} \mathbb{E}[Y^1 \mid X = x, T = 1] - \mathbb{E}[Y^0 \mid X = x, T = 0] \\ &\overset{(2)}{=} \mathbb{E}[Y \mid X = x, T = 1] - \mathbb{E}[Y \mid X = x, T = 0],\end{aligned} \tag{9}$$

where step (1) uses Assumption 3, and step (2) additionally relies on Assumptions 4, 5, and 6.

In practice, numerous approaches have been developed to estimate the last quantity. Classical methods include matching strategies (Caliendo & Kopeinig, 2008), which pair treated and control units with similar covariates, and meta-learners (Künzel et al., 2019), which adapt supervised learning techniques to causal inference tasks. More recent advances leverage deep learning, notably representation learning methods (Assaad et al., 2021; Wu et al., 2023), which aim to mitigate selection bias by learning balanced latent spaces. A prominent example is Counterfactual Regression (CFR) (Shalit et al., 2017), which introduces regularization terms based on distributional distances—such as the Wasserstein distance or the maximum mean discrepancy—to align representations between treatment groups.

Despite their success, these approaches heavily rely on the unconfoundedness assumption, which is often violated in real-world datasets where hidden confounding factors may influence both treatment

and outcome. While some recent methods attempt to account for hidden confounding, they typically depend on strong structural assumptions, additional proxies, or access to experimental (RCT) data—which may be costly or unavailable in practice.

In this paper, we take a new direction by exploring the potential of large language models (LLMs) to assist in imputing latent confounders from observational text and tabular data. Specifically, we propose a framework ProCI that leverages LLMs' implicit knowledge and generative ability to infer proxy variables that encode hidden causal information—thus helping to relax the unconfoundedness assumption and improve the robustness of causal estimates in the presence of hidden confounding.

B Related Works

In recent years, LLMs have been widely explored for causal inference tasks (Liu et al., 2024; Zhao et al., 2024; Chen et al., 2024b; Lin et al., 2024; Wei et al., 2022; Tan, 2023; Chen et al., 2024c; Lin et al., 2023; Vashishtha et al., 2025; Miliani et al., 2025; Sheth et al., 2025; Lee et al., 2025), especially in estimating treatment effects (Abdali et al., 2023; Jin et al., 2023; Pawlowski et al., 2023; Dhawan et al., 2024; Imai & Nakamura, 2026). These methods typically focus on leveraging causal structures from text or guiding LLMs to simulate interventional queries via carefully crafted prompts. For example, recent works use self-consistency (Abdali et al., 2023), tool augmentation (Pawlowski et al., 2023), and chain-of-thought prompting (Jin et al., 2023) to improve causal variable identification and treatment effect estimation. Dhawan et al. (2024) further automates treatment effect estimation by combining LLM outputs with traditional treatment effect estimators.

However, most prior LLM-based works rely solely on observed data and textual information, but ignore or fail to consider the existence of hidden confounders, which may lead to biased estimation of treatment effect. Traditional approaches for mitigating hidden confounding generally fall into three categories. Sensitivity analysis methods (Rosenbaum & Rubin, 1983; Robins et al., 2000) assess the potential effect of hidden confounders by deriving bounds on treatment effect estimates, though they rely on fixed, unverifiable assumptions (Franks et al., 2018; Veitch & Zaveri, 2020). Auxiliary variable techniques, including instrumental variables and front-door adjustments (Li et al., 2023; Fulcher et al., 2020; Shah et al., 2023), exploit external or intermediate variables to achieve unbiased estimates, but these depend on unverifiable structural assumptions (Imbens, 2014; Wu et al., 2022; Bellemare et al., 2024). Another line of work relies on RCT data (Kallus et al., 2018; Hatt et al., 2022; Wu & Yang, 2022) to correct for hidden biases, but their applicability is often constrained by the high costs of RCT data. While theoretically grounded, these methods for hidden confounding rely on strong auxiliary variables, which are often unavailable or hard to validate in real-world datasets. In contrast, our ProCI framework leverages LLMs to generate plausible hidden confounders from observational data and contextual priors, enabling knowledge-enhanced causal inference without requiring predefined auxiliary variables.

C Proof of Theorem 1

In Theorem 1, we aim to employ the kernel-based conditional independence test (KCIT) to assess whether the potential outcomes are conditionally independent of the treatment variable, given both observed and latent covariates. Specifically, we test whether:

$$\boldsymbol{Y} = (Y^0, Y^1) \perp\!\!\!\perp T \mid X, U$$

Hypotheses.

- **Null Hypothesis** (H_0): $\boldsymbol{Y} \perp\!\!\!\perp T \mid X, U$ — the potential outcomes are conditionally independent of treatment given covariates.
- **Alternative Hypothesis** (H_1): $\boldsymbol{Y} \not\perp\!\!\!\perp T \mid X, U$ — there exists residual dependence between treatment and outcomes after conditioning on covariates.

Test Statistic. KCIT estimates the squared Hilbert-Schmidt norm of the partial cross-covariance operator between $\boldsymbol{Y} = (Y^0, Y^1)$ and T given (X, U). Let n be the sample size. For two random vectors $X, Y \in \mathcal{X} \times \mathcal{Y}$, define the cross-variance operator as $\langle g, \Sigma_{YX} f \rangle = \mathbb{E}_{XY}(f(X)g(Y)) -$

$\mathbb{E}_X f(X)\mathbb{E}_Y g(Y)$ for $f \in \mathcal{H}_X$ and $g \in \mathcal{H}_Y$, the RKHS of $\mathcal{X}$ and $\mathcal{Y}$ respectively. The cross-variance operator is estimated via $\hat{\Sigma}_{YX} = \frac{1}{n}Tr(\tilde{K}_X\tilde{K}_Y)$, with $\tilde{K}_X = HK_XH$, $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ and the (i,j)-th entry of K_X is $k(x_i, x_j)$. The KCIT operator is estimated via:

$$\hat{\Sigma}_{\boldsymbol{Y}T|(X,U)} = \hat{\Sigma}_{\boldsymbol{Y}T} - \hat{\Sigma}_{\boldsymbol{Y}Z}(\hat{\Sigma}_{ZZ} + \gamma I)^{-1}\hat{\Sigma}_{ZT} \tag{10}$$

with $Z = (X, U)$, and $\gamma > 0$ is a regularization parameter. Finally, the KCIT test statistics is constructed by

$$\texttt{KCIT}(\boldsymbol{Y}, T \mid (X,U)) = \frac{1}{n}Tr(\hat{\Sigma}_{\boldsymbol{Y}T|(X,U)}). \tag{11}$$

Lemma 1 (Theorem 1 in (Amini & Razaee, 2021))**.** *Let $X_i \in \mathrm{LC}(\mu_i, \Sigma_i, \omega), i = 1, \ldots, n$, be a collection of independent random vectors from the LC distribution class defined in Amini & Razaee (2021), and let $K = K(X)$ be the kernel matrix for an L-Lipschitz kernel function $k(x_1, x_2)$, i.e. $|k(x_1, x_2) - k(y_1, y_2)| \le L(\|x_1 - y_1\| + \|x_2 - y_2\|)$. Then, for some universal constant $c > 0$, with probability at least $1 - \exp(-ct^2)$,*

$$\|K - \mathbb{E}K\| \le 2L\omega\sigma_\infty(Cn + \sqrt{n}t),$$

where $\sigma_\infty^2 := \max_i \|\Sigma_i\|$ and $C = c^{-1/2}$.

Lemma 2. *Let $\hat{\boldsymbol{Y}}_i, \hat{Z}_i$ be samples from LC classes $\mathrm{LC}(\mu_{\boldsymbol{Y}}, \Sigma_{Yi}, \omega)$ and $\mathrm{LC}(\mu_Z, \Sigma_{Zi}, \omega)$ respectively[1], such that $\|\mathbb{E}\tilde{K}_{\hat{\boldsymbol{Y}}} - \tilde{K}_{\boldsymbol{Y}}\| = o_p(1)$ and $\|\mathbb{E}\tilde{K}_{\hat{Z}} - \tilde{K}_Z\| = o_p(1)$. Then, for Gaussian RBF such that $\sigma \to \infty$ as $n \to \infty$, we have*

$$\begin{aligned}
\|\hat{\Sigma}_{\hat{\boldsymbol{Y}}T} - \hat{\Sigma}_{\boldsymbol{Y}T}\| &= o_p(1) \\
\|\hat{\Sigma}_{\hat{\boldsymbol{Y}}\hat{Z}} - \hat{\Sigma}_{\boldsymbol{Y}Z}\| &= o_p(1) \\
\|\hat{\Sigma}_{\hat{Z}T} - \hat{\Sigma}_{ZT}\| &= o_p(1) \\
\|\hat{\Sigma}_{\hat{Z}\hat{Z}} - \hat{\Sigma}_{ZZ}\| &= o_p(1)
\end{aligned} \tag{12}$$

with the last equation implies $\|\hat{\Sigma}^{-1}_{\hat{Z}\hat{Z}} - \hat{\Sigma}^{-1}_{ZZ}\| = o_p(1)$.

Proof of Lemma 2. From Lemma 1, with probability at least $1 - \exp(-ct^2)$,

$$\|K_{\hat{\boldsymbol{Y}}} - \mathbb{E}K_{\hat{\boldsymbol{Y}}}\| \le 2L\omega\sigma_\infty(Cn + \sqrt{n}t).$$

Therefore, since $L = o(\sigma^{-1}) = o(1)$ as n tends to infinity, $\|\tilde{K}_{\hat{\boldsymbol{Y}}} - \mathbb{E}\tilde{K}_{\hat{\boldsymbol{Y}}}\| \le \|H\|^2\|K_{\hat{\boldsymbol{Y}}} - \mathbb{E}K_{\hat{\boldsymbol{Y}}}\| = o_p(n)$. From above we have

$$\begin{aligned}
\|\hat{\Sigma}_{\hat{\boldsymbol{Y}}T} - \hat{\Sigma}_{\boldsymbol{Y}T}\| &= \frac{1}{n}Tr(\tilde{K}_X(\tilde{K}_{\hat{\boldsymbol{Y}}} - \tilde{K}_{\boldsymbol{Y}})) \\
&\le \frac{1}{n}\|\tilde{K}_X\| \cdot \|\tilde{K}_{\hat{\boldsymbol{Y}}} - \tilde{K}_{\boldsymbol{Y}}\| \\
&\le \frac{1}{n}\|\tilde{K}_X\| \cdot \big(\|\tilde{K}_{\hat{\boldsymbol{Y}}} - \mathbb{E}\tilde{K}_{\hat{\boldsymbol{Y}}}\| + \|\mathbb{E}\tilde{K}_{\hat{\boldsymbol{Y}}} - \tilde{K}_{\boldsymbol{Y}}\|\big) \\
&= o_p(1),
\end{aligned}$$

which proves the first equation in Eq. (12). The second equation comes from the fact that

$$\begin{aligned}
\|\hat{\Sigma}_{\hat{\boldsymbol{Y}}\hat{Z}} - \hat{\Sigma}_{\boldsymbol{Y}Z}\| &= \frac{1}{n}Tr(\tilde{K}_{\hat{\boldsymbol{Y}}}(\tilde{K}_{\hat{Z}} - \tilde{K}_Z) + \tilde{K}_Z(\tilde{K}_{\hat{\boldsymbol{Y}}} - \tilde{K}_{\boldsymbol{Y}})) \\
&\le \frac{1}{n}\Big(\|\tilde{K}_{\hat{\boldsymbol{Y}}}\| \cdot \|\tilde{K}_{\hat{Z}} - \tilde{K}_Z\| + \|\tilde{K}_Z\| \cdot \|\tilde{K}_{\hat{\boldsymbol{Y}}} - \tilde{K}_{\boldsymbol{Y}}\|\Big) \\
&= o_p(1).
\end{aligned}$$

with the last equation resulting from $\|\tilde{K}_{\hat{\boldsymbol{Y}}} - \tilde{K}_{\boldsymbol{Y}}\| = o_p(n)$ and $\|\tilde{K}_{\hat{Z}} - \tilde{K}_Z\| = o_p(n)$. The third equation in Eq. (12) comes from the same deduction as for the first equation, and the last equation comes from the same deduction for the second equation.

Finally, the conclusion on the inverse matrix is straightforward observing that

$$(\hat{\Sigma}_{\hat{Z}\hat{Z}} + \mathcal{E})^{-1} = \hat{\Sigma}^{-1}_{\hat{Z}\hat{Z}} - \hat{\Sigma}^{-1}_{\hat{Z}\hat{Z}}\mathcal{E}\hat{\Sigma}^{-1}_{\hat{Z}\hat{Z}} + O_p(\|\mathcal{E}\|) = \hat{\Sigma}^{-1}_{ZZ} + o_p(1)$$

with $\|\mathcal{E}\| = o_p(1)$. □

[1]Here $\hat{Z} = (X, \hat{U})$, where $\hat{U}$ is the estimation of latent confounders.

Table 4: Hyperparameter search space used in all experiments.

Hyperparameter	Search Range	Description
lr	$\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$	learning rate
bs	$\{16, 32, 64, 128\}$	batch size
λ	$\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$	loss balancing coefficient
d_ϕ	$\{16, 32, 64\}$	hidden dimension in encoder network ϕ
d_h	$\{16, 32, 64\}$	hidden dimension in outcome heads h_0 and h_1

Proof of Theorem 1. Based on Lemma 2 and Eq. (10), we have

$$\begin{aligned}\hat{\Sigma}_{\hat{\boldsymbol{Y}}T|\hat{Z}} &= \hat{\Sigma}_{\hat{\boldsymbol{Y}}T} - \hat{\Sigma}_{\hat{\boldsymbol{Y}}\hat{Z}}(\hat{\Sigma}_{\hat{Z}\hat{Z}} + \gamma I)^{-1}\hat{\Sigma}_{\hat{Z}T} \\ &= \hat{\Sigma}_{\boldsymbol{Y}T} - \hat{\Sigma}_{\boldsymbol{Y}Z}(\hat{\Sigma}_{ZZ} + \gamma I)^{-1}\hat{\Sigma}_{ZT} + o_p(1) \\ &= \hat{\Sigma}_{\boldsymbol{Y}T|Z} + o_p(1).\end{aligned}$$

Therefore, based on Eq. (11), we have

$$\begin{aligned}&|\texttt{KCIT}\big((Y^0, Y^1), T \mid X, U\big) - \texttt{KCIT}\big((\hat{Y}^0, \hat{Y}^1), T \mid X, \hat{U}\big)| \\ =&\frac{1}{n}|Tr(\hat{\Sigma}_{\hat{\boldsymbol{Y}}T|\hat{Z}} - \hat{\Sigma}_{\boldsymbol{Y}T|Z})| \\ \leq&\|\hat{\Sigma}_{\hat{\boldsymbol{Y}}T|\hat{Z}} - \hat{\Sigma}_{\boldsymbol{Y}T|Z}\| \\ =&o_p(1),\end{aligned}$$

which proves the theorem.

D Further Experimental Details

D.1 Dataset

Twins. The Twins dataset is derived from all recorded twin births in the United States between 1989 and 1991 (Almond et al., 2005). We focus on twin pairs where both individuals weighed less than 2000 grams at birth. Each instance contains 50 pre-treatment covariates related to parental characteristics, pregnancy conditions, and birth outcomes. The treatment assignment is defined such that $T = 1$ corresponds to the heavier twin and $T = 0$ to the lighter one. The outcome variable Y is one-year mortality.

After removing records with missing values, the resulting dataset comprises 8,244 samples. Since data for both twins in each pair is available, we observe outcomes under both treatment assignments ($T = 0$ and $T = 1$). To emulate an observational setting, we simulate unobserved counterfactuals by selectively masking one twin per pair. When this selection is randomized, the data mimics a randomized controlled trial. In specific, we model confounding via a proxy variable, where we assign treatment based on a single feature—GESTAT10—which represents gestational age in 10 categories. Formally, treatment is drawn as: $T_i \mid X_i, Z_i \sim \text{Bern}\left(\sigma\left(W_o^\top X_i + W_h(Z_i/10 - 0.1)\right)\right)$, where $W_o \sim \mathcal{N}(0, 0.1 \cdot I)$ and $W_h \sim \mathcal{N}(5, 0.1)$. Here, X_i denotes the 49 non-GESTAT10 features and Z_i is the GESTAT10 value for unit i.

Jobs. This dataset combines the Lalonde randomized experiment (297 treated and 425 control units) with an observational sample from the PSID (2,490 control units) (LaLonde, 1986). Each record includes 7 covariates such as age, education level, ethnicity, and prior earnings. The outcome reflects post-intervention employment status. By merging the experimental and observational subsets, we can introduce the selection bias between treated and control groups, making this dataset useful for evaluating robustness to such bias.

D.2 Training and Evaluation Protocols

Training Protocol. All models are optimized using grid search based on validation performance. The learning rate and batch size are tuned over predefined discrete sets: $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ for learning rates and $\{16, 32, 64, 128\}$ for batch sizes. For methods involving balancing

losses (CFR-Wass, CFR-MMD, and ESCFR), the regularization weight λ is selected from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ to control the trade-off between outcome prediction and representation alignment. Table 4 summarizes the full hyperparameter configuration space used during training. All models, including the baselines and our proposed method, are tuned under the same conditions to ensure fair comparison.

Training is performed for a maximum of 200 epochs, with early stopping applied based on validation loss. Specifically, we stop the training process if no improvement is observed within 30 consecutive epochs, which helps prevent overfitting—particularly relevant for datasets like Twins, where ground-truth outcomes are fully known. We use grid search for hyperparameter tuning on validation sets, and adopt consistent settings across shared base models. LLM temperature is fixed at 0.7. For both datasets, we split the data into training, validation, and test sets (63:27:10 for Twins, 56:24:20 for Jobs). While in ProCI, the progressive procedure is defined based on the KCIT passing condition, we additionally impose a maximum iteration limit $K_{\max} = 10$ to ensure bounded computation. The algorithm terminates when either the KCIT criterion is satisfied or the maximum number of iterations is reached. All experiments are implemented using PyTorch 1.10 and trained with the Adam optimizer. Hardware used includes an NVIDIA A40 GPU and an Intel(R) Xeon(R) Gold 5318Y CPU at 2.10GHz.

Evaluation Protocol. For the Twins dataset, where the distributions of potential outcomes are available, we evaluate model performance using two metrics: the Precision in Estimation of Heterogeneous Effect (ϵ_{PEHE}) and the Average Treatment Effect error (ϵ_{ATE}) (Hill, 2011).

The PEHE is defined as:

$$\epsilon_{PEHE} = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{E}_{(y_i^0, y_i^1)\sim\mathcal{P}_{\mathbf{Y}|\mathbf{x}_i}}\left(y_i^1 - y_i^0\right) - \left(\hat{y}_i^1 - \hat{y}_i^0\right)\right)^2,$$

where $\hat{y}_i^0$ and $\hat{y}_i^1$ denote the estimated outcomes under control and treatment, respectively, and y_i^0 and y_i^1 represent the corresponding true outcomes.

For the ATE error, we compute it as:

$$\epsilon_{ATE} = \left|\frac{1}{N}\sum_{i=1}^{N}\left(y_i^1 - y_i^0\right) - \frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i^1 - \hat{y}_i^0\right)\right|.$$

Lower values of ϵ_{PEHE} and ϵ_{ATE} indicate better estimation performance.

For the Jobs dataset, where ground-truth ITE is not available, we use two metrics: policy risk $\mathcal{R}_{pol}$ (Shalit et al., 2017) and the error in estimating the Average Treatment effect on the Treated (ϵ_{ATT}).

Policy risk is defined as:

$$\mathcal{R}_{pol} = 1 - \left(\mathbb{E}[Y^1 \mid \pi(x) = 1]\cdot\mathbb{P}(\pi(x) = 1) + \mathbb{E}[Y^0 \mid \pi(x) = 0]\cdot\mathbb{P}(\pi(x) = 0)\right),$$

where $\pi(x) = 1$ if $\hat{y}^1 - \hat{y}^0 > 0$, and $\pi(x) = 0$ otherwise.

We estimate this metric using only units from the randomized component of the dataset:

$$\begin{aligned}\mathcal{R}_{pol} = 1 - \Bigg(&\frac{1}{|A^1 \cap T^1 \cap E|}\sum_{\mathbf{x}_i \in A^1\cap T^1\cap E} y_i^1 \cdot \frac{|A^1 \cap E|}{|E|} \\ &+ \frac{1}{|A^0 \cap T^0 \cap E|}\sum_{\mathbf{x}_i \in A^0\cap T^0\cap E} y_i^0 \cdot \frac{|A^0 \cap E|}{|E|}\Bigg)\end{aligned} \tag{13}$$

with E denoting the randomized experiment set, $A^1 = \{\mathbf{x}_i : \hat{y}_i^1 - \hat{y}_i^0 > 0\}$, $A^0 = \{\mathbf{x}_i : \hat{y}_i^1 - \hat{y}_i^0 < 0\}$, $T^1 = \{\mathbf{x}_i : t_i = 1\}$, and $T^0 = \{\mathbf{x}_i : t_i = 0\}$. A lower value of $\mathcal{R}_{pol}$ indicates that the CATE estimation method provides better support for the decision-making strategy.

We also report ϵ_{ATT} as:

$$\epsilon_{ATT} = \left|\frac{1}{N_1}\sum_{i:t_i=1}\left(y_i^1 - y_i^0\right) - \frac{1}{N_1}\sum_{i:t_i=1}\left(\hat{y}_i^1 - \hat{y}_i^0\right)\right|,$$

where N_1 is the number of treated units in the randomized group. A lower ϵ_{ATT} indicates more accurate treatment effect estimation for the treated population.

D.3 Base Models

Since our proposed ProCI framework is model-agnostic and only augments the original dataset with new confounders, it can be flexibly combined with a variety of existing CATE estimation methods. In our experiments, we treat several well-established and state-of-the-art CATE estimators as base models to assess how their performance improves when equipped with the confounders generated by ProCI.

We consider representative methods from three major categories: meta-learning, matching-based, and representation-based approaches.

i) **Meta-learners**: These methods differ in how they handle treatment information. The **S-Learner** (Künzel et al., 2019) uses a single model that includes treatment as a feature.

ii) **Matching-based methods**: We include propensity score matching (**PSM**) (Caliendo & Kopeinig, 2008), followed by regression on the matched samples. Propensity scores in PSM are estimated using logistic regression, consistent with the implementation in (Caliendo & Kopeinig, 2008).

iii) **Representation-based methods**: This includes **TARNet** (Shalit et al., 2017), which employs a shared feature representation with separate heads for predicting potential outcomes; **CFR-Wass** (Shalit et al., 2017), which adds distributional regularization using the Wasserstein metric; and **ESCFR** (Wang et al., 2023), which leverages unbalanced optimal transport to achieve mini-batch-level representation balance and robustness to outliers.

D.4 Time Complexity.

Let n denote the number of samples and $K_{\max}$ the maximum number of progressive iterations. The overall complexity of the proposed ProCI framework consists of three components executed within each iteration.

(1) LLM-based parameter inference. At each iteration, the LLM is queried over the dataset (in mini-batches) to estimate $\{p_i\}_{i=1}^n$, resulting in a cost of $\mathcal{O}(n \cdot C_{\text{LLM}})$ per iteration, and $\mathcal{O}(K_{\max} \cdot n \cdot C_{\text{LLM}})$ overall.

(2) Conditional independence testing (KCIT). The KCIT procedure is executed at each iteration to evaluate the stopping criterion. In its standard kernel-based form, KCIT incurs a computational cost of $\mathcal{O}(n^3)$, leading to a total cost of $\mathcal{O}(K_{\max} \cdot n^3)$.

(3) CATE model training. The downstream model is trained once after the progressive procedure, with complexity $\mathcal{C}_{\text{CATE}}(n, d)$.

Overall, the time complexity is:

$$\mathcal{O}\big(K_{\max}(n \cdot C_{\text{LLM}} + n^3) + \mathcal{C}_{\text{CATE}}(n, d)\big).$$

Practical considerations. In practice, we employ mini-batch LLM inference and approximate KCIT (e.g., via subsampling or low-rank kernel approximations), reducing the effective cost significantly. Moreover, we set a small iteration limit (e.g., $K_{\max} = 10$), ensuring tractable execution.

E Additional Experimental Results

E.1 Effectiveness of ProCI with Open-Source Language Models

Experimental Setup. To further investigate the generalizability of ProCI under different LLM architectures, we introduce two additional models: LLaMA 3–8B and Qwen2.5–7B, denoted as **ProCI-La** and **ProCI-Qw**, respectively. These models are selected for their competitive reasoning capabilities and open accessibility. ProCI-La is based on Meta's LLaMA 3 series (Grattafiori et al., 2024), a dense decoder-only transformer optimized for instruction-following. ProCI-Qw leverages Alibaba's Qwen2.5 family (Qwen et al., 2025), which has shown strong performance in multi-lingual and causal reasoning tasks.

Table 5: Overall performance comparison of treatment effect estimation between base models and their enhanced versions with ProCI-La (LLaMA 3-8B) and ProCI-Qw (Qwen2.5-7B). Best-performing results across all methods are highlighted.

Datasets	Jobs				Twins			
Test Types	In-sample		Out-sample		In-sample		Out-sample	
Methods	ϵ_{ATT}	$\mathcal{R}_{pol}$	ϵ_{ATT}	$\mathcal{R}_{pol}$	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}
S-Learner	$0.0491_{\pm 0.0011}$	$0.2289_{\pm 0.0008}$	$0.0876_{\pm 0.0018}$	$0.1678_{\pm 0.0002}$	$0.0131_{\pm 0.0020}$	$0.2527_{\pm 0.0012}$	$0.0037_{\pm 0.0024}$	$0.2819_{\pm 0.0001}$
+ProCI-La	$0.0227_{\pm 0.0004}$	$0.2253_{\pm 0.0013}$	$0.1009_{\pm 0.0001}$	$0.1648_{\pm 0.0002}$	$0.0117_{\pm 0.0001}$	$0.2536_{\pm 0.0000}$	$0.0035_{\pm 0.0029}$	$0.2833_{\pm 0.0032}$
+ProCI-Qw	$0.0320_{\pm 0.0007}$	$0.2301_{\pm 0.0000}$	$0.0832_{\pm 0.0001}$	$0.1697_{\pm 0.0010}$	$0.0068_{\pm 0.0013}$	$0.2518_{\pm 0.0027}$	$0.0050_{\pm 0.0004}$	$0.2846_{\pm 0.0012}$
PSM	$0.6197_{\pm 0.0000}$	$0.2707_{\pm 0.0000}$	$0.1259_{\pm 0.0015}$	$0.2192_{\pm 0.0013}$	$0.0457_{\pm 0.0006}$	$0.3399_{\pm 0.0007}$	$0.0840_{\pm 0.0000}$	$0.4027_{\pm 0.0001}$
+ProCI-La	$0.6185_{\pm 0.0008}$	$0.2716_{\pm 0.0003}$	$0.1126_{\pm 0.0002}$	$0.2181_{\pm 0.0001}$	$0.0460_{\pm 0.0012}$	$0.3394_{\pm 0.0004}$	$0.0845_{\pm 0.0000}$	$0.4026_{\pm 0.0000}$
+ProCI-Qw	$0.6173_{\pm 0.0000}$	$0.2691_{\pm 0.0001}$	$0.0992_{\pm 0.0003}$	$0.2191_{\pm 0.0035}$	$0.0455_{\pm 0.0037}$	$0.3401_{\pm 0.0000}$	$0.0818_{\pm 0.0001}$	$0.4017_{\pm 0.0000}$
TARNet	$0.0191_{\pm 0.0002}$	$0.2177_{\pm 0.0001}$	$0.1466_{\pm 0.0026}$	$0.2201_{\pm 0.0002}$	$0.0233_{\pm 0.0044}$	$0.2917_{\pm 0.0001}$	$0.0310_{\pm 0.0005}$	$0.3237_{\pm 0.0001}$
+ProCI-La	$0.0163_{\pm 0.0021}$	$0.2059_{\pm 0.0027}$	$0.1179_{\pm 0.0032}$	$0.2139_{\pm 0.0011}$	$0.0190_{\pm 0.0021}$	$0.2755_{\pm 0.0013}$	$0.0270_{\pm 0.0019}$	$0.3043_{\pm 0.0012}$
+ProCI-Qw	$0.0129_{\pm 0.0001}$	$0.2005_{\pm 0.0000}$	$0.0761_{\pm 0.0010}$	$0.2201_{\pm 0.0032}$	$0.0101_{\pm 0.0000}$	$0.2775_{\pm 0.0014}$	$0.0092_{\pm 0.0001}$	$0.3123_{\pm 0.0002}$
CFR-Wass	$0.0355_{\pm 0.0006}$	$0.2150_{\pm 0.0001}$	$0.1487_{\pm 0.0028}$	$0.2191_{\pm 0.0004}$	$0.0189_{\pm 0.0000}$	$0.2818_{\pm 0.0000}$	$0.0186_{\pm 0.0002}$	$0.3138_{\pm 0.0000}$
+ProCI-La	$0.0312_{\pm 0.0003}$	$0.2023_{\pm 0.0001}$	$0.1373_{\pm 0.0073}$	$0.1986_{\pm 0.0004}$	$0.0132_{\pm 0.0001}$	$0.2762_{\pm 0.0001}$	$0.0150_{\pm 0.0001}$	$0.3073_{\pm 0.0001}$
+ProCI-Qw	$0.0295_{\pm 0.0004}$	$0.2029_{\pm 0.0001}$	$0.1389_{\pm 0.0068}$	$0.2140_{\pm 0.0002}$	$0.0137_{\pm 0.0000}$	$0.2724_{\pm 0.0000}$	$0.0110_{\pm 0.0001}$	$0.3041_{\pm 0.0002}$
ESCFR	$0.0543_{\pm 0.0012}$	$0.2184_{\pm 0.0001}$	$0.2245_{\pm 0.0390}$	$0.2274_{\pm 0.0002}$	$0.0199_{\pm 0.0001}$	$0.2715_{\pm 0.0001}$	$0.0207_{\pm 0.0003}$	$0.3059_{\pm 0.0007}$
+ProCI-La	$0.0362_{\pm 0.0005}$	$0.2074_{\pm 0.0000}$	$0.1218_{\pm 0.0038}$	$0.2055_{\pm 0.0015}$	$0.0160_{\pm 0.0001}$	$0.2714_{\pm 0.0002}$	$0.0177_{\pm 0.0001}$	$0.3030_{\pm 0.0001}$
+ProCI-Qw	$0.0317_{\pm 0.0008}$	$0.2154_{\pm 0.0011}$	$0.2114_{\pm 0.0430}$	$0.2239_{\pm 0.0033}$	$0.0131_{\pm 0.0001}$	$0.2699_{\pm 0.0021}$	$0.0062_{\pm 0.0000}$	$0.2989_{\pm 0.0001}$

We apply the same ProCI framework using LLaMA 3–8B and Qwen2.5–7B as the confounder generators. All settings (e.g., temperature = 0.7, prompt structure, distribution identification and imputation, progressive confounder generation) remain consistent with the original experiments to ensure fair comparisons. The resulting augmented datasets are then passed into the same downstream CATE estimators as in the original setup.

Results. As shown in Table 5, both ProCI-La and ProCI-Qw significantly improve CATE estimation performance over their corresponding base models, confirming the effectiveness of using LLM-generated hidden confounders even beyond proprietary GPT models. Notably, the improvements are consistent across different types of base estimators, with the following observations:

- **ProCI-La, based on LLaMA 3–8B, achieves strong performance gains on both in-sample and out-of-sample evaluations.** Its improvements are particularly evident when paired with representation-based base models such as TARNet and CFR-Wass, indicating that LLaMA 3's semantic reasoning helps uncover latent variables that enhance feature balancing in the learned representations.
- **ProCI-Qw, leveraging Qwen2.5–7B, also brings stable improvements over base models.** While its performance slightly lags behind ProCI-La in some cases, it still consistently enhances treatment effect estimation, especially under models sensitive to unobserved confounding. This suggests Qwen2.5 can effectively contribute causal priors despite its smaller scale.
- **Across both models, we observe that representation learning-based estimators benefit the most from ProCI augmentation.** These models are designed to learn balanced representations of treated and control groups, and the inclusion of high-quality confounders improves this balancing, thereby reducing estimation bias and variance more effectively.

These findings demonstrate that ProCI is model-agnostic and remains effective when paired with open-source, instruction-tuned LLMs. This extends its practical applicability and offers a cost-efficient, scalable solution for treatment effect estimation under hidden confounding in real-world settings.

E.2 Impact of Temperature on Confounder Quality

Experimental Setup. To assess how the temperature coefficient λ affects the quality of generated confounders, we conduct a controlled experiment using two open-source LLMs: **LLaMA 3–8B** (**ProCI-La**) and **Qwen2.5–7B** (**ProCI-Qw**). We vary λ in the range $\{0.1, 0.3, 0.5, 0.7, 0.9\}$, and for

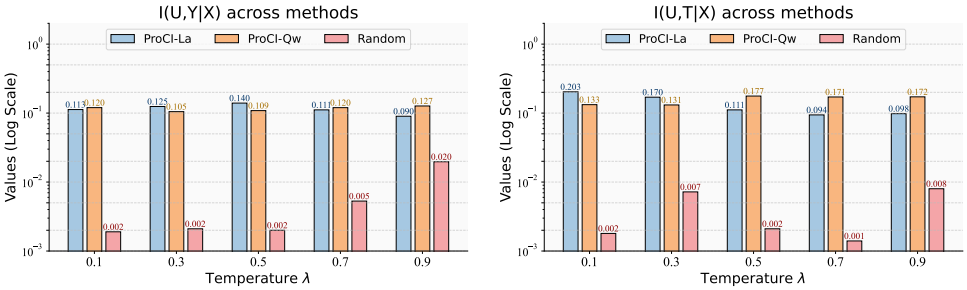


Figure 5: **Conditional mutual information (CMI) values across different temperatures for confounders generated by ProCI-La (LLaMA 3–8B), ProCI-Qw (Qwen2.5–7B), and a Random baseline.** The left plot shows $I(U, Y \mid X)$, and the right plot shows $I(U, T \mid X)$, with all values on a log scale. ProCI methods consistently outperform the Random baseline, with peak informativeness generally occurring at moderate temperatures (0.5–0.7).

each value, use the respective LLM to generate confounders following the standard ProCI framework. To evaluate the informativeness of the generated variables, we compute their conditional mutual information (CMI) with the treatment and outcome, conditioned on observed covariates: $I(U, Y \mid X)$ and $I(U, T \mid X)$. We also include a **Random** baseline that generates synthetic variables from uniform or Gaussian noise independent of the data. This helps isolate the contribution of semantically meaningful generation from LLMs.

Results. Figure 5 shows the CMI values $I(U, Y \mid X)$ and $I(U, T \mid X)$ across different temperature values. Each x-axis tick corresponds to a specific temperature coefficient λ, and we compare three methods: Random, ProCI-La, and ProCI-Qw.

- **ProCI-La and ProCI-Qw significantly outperform the Random baseline at all temperatures**, with CMI values often one to two orders of magnitude higher, validating that LLM-generated confounders encode semantically relevant information about treatment and outcome.
- **The impact of temperature varies by method and metric.** For $I(U, Y \mid X)$, ProCI-La achieves the highest value at $\lambda = 0.5$, while ProCI-Qw peaks at $\lambda = 0.9$. For $I(U, T \mid X)$, ProCI-La performs best at the lowest temperature ($\lambda = 0.1$), whereas ProCI-Qw shows more stable performance across mid-to-high temperatures.
- **ProCI-La generally outperforms ProCI-Qw in capturing outcome-relevant information**, especially at lower temperatures. In contrast, ProCI-Qw sometimes surpasses ProCI-La on treatment-related informativeness ($I(U, T \mid X)$), suggesting complementary strengths between the LLMs.

These findings reinforce the benefit of using temperature tuning to control the diversity and informativeness of generated confounders. Moderate values ($\lambda \in [0.5, 0.7]$) typically offer the best trade-off, with performance degrading slightly at the extremes.

E.3 Semantic Comparison Between Removed Covariate and Imputed Confounder

Experimental Setup. We conduct a targeted sanity-check experiment to evaluate whether ProCI can recover known hidden confounders. Specifically, we remove one observed covariate at a time from the Jobs dataset and treat it as a hidden confounder. ProCI is then applied to generate a new confounder based on the remaining variables. We compare the generated confounder with the removed covariate, categorizing their relationship as fully matched, strongly related, or weakly related based on semantic similarity.

Results. The results from Table 6 show that ProCI can recover confounders that are semantically aligned with the removed covariates in most cases. When an exact recovery is not achieved, the generated confounder typically corresponds to a semantically meaningful proxy rather than an unrelated variable. For example, when the covariate *education years* is removed, ProCI often generates

Table 6: Semantic comparison between removed covariates and imputed confounders. The number of ✓symbols indicates the level of semantic matching: three checkmarks for a full match, two for a strong relation, and one for a weaker relation.

Covariate	Scope	Confounder Generated by ProCI		
		Name	Scope	Matching
Age	$[17, 55]$	Age	$[20, 50]$	✓✓✓
Education years	$[3, 16]$	Education background	{Bachelor, Master, ...}	✓✓
Married	$\{0, 1\}$	Marital status	{Single, Married, Divorced}	✓✓
Hispanic	$\{0, 1\}$	Living country	{USA, UK, Italy, ...}	✓
Previous earning	$[0, 37431]$	Average Income per year	$[0, 120000]$	✓✓

ProCI-4o (GPT-4o): Latent Confounder = `EmploymentHistory`

One-step rationale.

GPT-4o identifies `Employment History` as a latent confounder, arguing that prior work experience influences both participation in job training and post-training employment outcomes. Individuals with stable employment histories are more likely to enroll in training programs and to secure employment afterward, while those with limited employment experience may both seek training and face worse employment prospects.

ProCI-R1 (DeepSeek-R1): Latent Confounder = `TransportationAccess`

Multi-step causal rationale.

Step 1: Latent factor identification. `Transportation Access` captures whether an individual has reliable means to commute to training sites.

Step 2: Effect on treatment assignment ($U \to T$). Limited transportation reduces the feasibility of attending training sessions, decreasing the likelihood of program participation.

Step 3: Effect on outcome ($U \to Y$). Transportation constraints independently limit access to job opportunities, negatively affecting employment outcomes.

Step 4: Confounding mechanism. By simultaneously influencing both participation and employment, `Transportation Access` induces a confounding pathway not explained by observed covariates.

Figure 6: Qualitative comparison of causal rationales in a reasoning-intensive scenario. DeepSeek-R1 produces an explicit multi-step causal explanation, while GPT-4o provides a concise summary-style rationale.

confounders such as *education background*, which captures similar underlying concepts despite differences in value ranges. These findings indicate that ProCI tends to recover relevant confounding information or reasonable proxies, rather than producing arbitrary or spurious confounders.

E.4 Qualitative Comparison of Causal Rationales

Experimental Setup. We conduct a qualitative comparison between **ProCI-4o** and **ProCI-R1** on the Jobs dataset to analyze how different LLM reasoning styles affect the generation and explanation of latent confounders. The ProCI framework is kept identical across settings, and only the underlying LLM used for confounder generation and explanation is varied between GPT-4o and DeepSeek-R1. We select representative cases with different levels of complexity, including scenarios that require multi-step causal decomposition as well as scenarios where confounding relationships are largely direct.

Results. Figures 6 and 7 illustrate two contrasting qualitative cases. We note that different LLMs may generate different latent confounders for the same dataset, reflecting their distinct inductive biases and reasoning styles.

In particular, in reasoning-intensive scenarios (Figure 6), DeepSeek-R1 identifies a mechanism-driven latent confounder and produces a structured, multi-step causal rationale that explicitly decomposes intermediate pathways, making the confounding mechanism more transparent. In contrast, GPT-4o

ProCI-4o (GPT-4o): Latent Confounder = `EducationLevel`

Concise rationale.

GPT-4o identifies `Education Level` as a latent confounder, noting that individuals with higher education are more likely to participate in job training programs and have better employment outcomes. This confounding relationship is largely direct and does not require explicit intermediate mechanisms.

ProCI-R1 (DeepSeek-R1): Latent Confounder = `JobSearchMotivation`

Structured rationale.

DeepSeek-R1 identifies `Job Search Motivation` as a latent confounder and decomposes its effect into multiple intermediate mechanisms, including **individual expectations, persistence in job search,** and **responsiveness to training opportunities**. While plausible, this structured decomposition introduces additional complexity for a confounding relationship that hard to be accurately identified.

Figure 7: Qualitative comparison of causal rationales in a low-complexity scenario. GPT-4o generates a direct and concise confounder explanation, while DeepSeek-R1 introduces additional intermediate reasoning steps.

Table 7: Performance comparison on the Twins dataset when decoupling confounder generation and outcome imputation in ProCI.

Base Model	**GPT-4o**				**DeepSeek-R1**			
Test Types	In-sample		Out-sample		In-sample		Out-sample	
Methods	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}
LLM	–	–	–	–	–	–	–	–
+LLM	$0.0191_{\pm 0.0001}$	$0.2700_{\pm 0.0032}$	$\mathbf{0.0297}_{\pm 0.0029}$	$0.3045_{\pm 0.0001}$	$\mathbf{0.0087}_{\pm 0.0001}$	$0.2684_{\pm 0.0002}$	$\mathbf{0.0132}_{\pm 0.0001}$	$\mathbf{0.3038}_{\pm 0.0000}$
+S-Learner	$0.0186_{\pm 0.0003}$	$0.2711_{\pm 0.0003}$	$0.0311_{\pm 0.0014}$	$0.3041_{\pm 0.0001}$	$0.0091_{\pm 0.0003}$	$\mathbf{0.2659}_{\pm 0.0004}$	$0.0148_{\pm 0.0010}$	$0.3051_{\pm 0.0002}$
+CFR-Wass	$\mathbf{0.0180}_{\pm 0.0029}$	$\mathbf{0.2693}_{\pm 0.0001}$	$0.0306_{\pm 0.0007}$	$\mathbf{0.3033}_{\pm 0.0002}$	$0.0089_{\pm 0.0009}$	$0.2671_{\pm 0.0018}$	$0.0150_{\pm 0.0026}$	$0.3065_{\pm 0.0012}$

provides a concise, summary-style explanation that does not expose the underlying causal structure in detail. On the other hand, in low-complexity scenarios (Figure 7), GPT-4o generates a direct and sufficient confounder with a concise causal interpretation, while DeepSeek-R1 introduces additional intermediate mechanisms through step-wise reasoning. Although these steps are individually plausible, they offer limited additional insight for confounding relationships that can be captured at a more direct level.

E.5 Decoupling Confounder Generation and Outcome Imputation

Experimental Setup. We conduct an additional experiment on the Twins dataset to examine potential circularity in the unconfoundedness validation step by decoupling confounder generation and counterfactual outcome imputation. For confounder generation, we use two LLM backbones: GPT-4o and DeepSeek-R1. For each LLM backbone, we compare three outcome imputation settings: (i) **LLM + LLM**, where the same LLM is used for both confounder generation and counterfactual outcome imputation; (ii) **LLM + S-Learner**, where counterfactual outcomes are imputed using an S-Learner; and (iii) **LLM + CFR-Wass**, where counterfactual outcomes are imputed using a CFR-Wass model. The final CATE estimator is fixed as ESCFR model.

Results. From Table 7, we can see: Across both GPT-4o and DeepSeek-R1 backbones, replacing LLM-based outcome imputation with S-Learner or CFR-Wass results in comparable or slightly degraded performance. No consistent performance gains are observed after decoupling the outcome imputation stage. This provides evidence that the validation behavior is not merely an artifact of LLM self-consistency, alleviating concerns about circularity in the proposed framework.

F Prompt Templates

In this section, we present the detailed prompt templates used in the proposed ProCI framework. These prompts correspond to the four key components of our method: variable generation

($\mathcal{P}_{\text{var}}(X, Y, T)$ in Eq. (2), distribution type inference ($\mathcal{P}_{\text{dist}}(X, Y, T, \hat{U})$ in Eq. (3), parameter estimation ($\mathcal{P}_{\text{param}}(x_i, t_i, y_i)$ in Eq. (4), and counterfactual outcome imputation ($\mathcal{P}_{\text{out}}(x_i, u_i, y_i, t_i)$ in Eq. (5).

F.1 Prefix Prompt

The prefix prompt provides contextual information about the dataset, including a brief overview and detailed descriptions of the treatment, outcome, and confounding variables. This prompt serves as a foundation and should be included at the beginning of all subsequent prompts to ensure that the LLM is aware of the data background and variable semantics.

Prefix prompt: Dataset Introduction

Inputs: The dataset name $\mathcal{D}_{\text{name}}$ with a brief introduction $\mathcal{D}_{\text{intro}}$; variable names for confounders X_{name}, treatment T_{name}, and outcome Y_{name}; and their corresponding descriptions as provided by the original dataset: X_{desc}, T_{desc}, and Y_{desc}.

Prompt:

Brief introduction of the $\{\mathcal{D}_{\text{name}}\}$ dataset: $\{\mathcal{D}_{\text{intro}}\}$

This observational dataset contains:

(1) Treatment — $\{T_{\text{name}}\}$: $\{T_{\text{desc}}\}$

(2) Outcome — $\{Y_{\text{name}}\}$: $\{Y_{\text{desc}}\}$

(3) Confounders — $\{X_{\text{name}}\}$: $\{X_{\text{desc}}\}$

F.2 Variable Generation

In this prompt, we mainly utilize the name and description of variables X, T and Y in the observational dataset to infer a new confounder $\hat{U}$.

$\mathcal{P}_{\text{var}}(X, Y, T)$: Generating new confounder $\hat{U}$

Inputs: Prefix Prompt

Outputs: Confounder name $\hat{U}_{\text{name}}$, a brief explanation $\hat{U}_{\text{exp}}$

Prompts:

{Prefix prompt}

Based on your WORLD KNOWLEDGE, please propose one additional confounder which BOTH affects the treatment and outcome.

Make sure that the proposed confounder has a DIFFERENT MEANING compared to existing confounders.

For this proposed confounder, please provide:

(1) A clear name for the confounder.

(2) A brief explanation of why it affects both treatment and outcome.

F.3 Distribution type inference

After identifying the confounder variable, we leverage the commonsense knowledge embedded in LLMs to infer an appropriate distribution type for it. Instead of directly imputing its values using the

LLM, which often leads to degenerate or collapsed outputs when applied to tabular data, we defer value imputation to a subsequent structured process.

$\mathcal{P}_{\text{dist}}(X, Y, T, \hat{U})$: Inferring the distribution type of $\hat{U}$

Inputs: Prefix prompt, name of generated variable $\hat{U}_{\text{name}}$

Outputs: Distribution type $\mathcal{F}_{\hat{U}}$

Prompts:

{Prefix Prompt}

Based on your WORLD KNOWLEDGE, please provide the distribution type of confounder $\{\hat{U}_{\text{name}}\}$. For example:

(1) Continuous — e.g., Normal distribution

(2) Discrete — e.g., Multi-categorical distribution

(3) Binary — e.g., Bernoulli distribution

F.4 PARAMETER ESTIMATION

Given the inferred distribution type $\mathcal{F}_{\hat{U}}$ of the confounder $\hat{U}$, the next step is to estimate the corresponding distribution parameters. As there are various possible distribution families, we illustrate the parameter estimation process using the normal distribution as an example.

$\mathcal{P}_{\text{param}}(\mathcal{D}_X, \mathcal{D}_T, \mathcal{D}_Y)$: Generating the distribution parameter for each unit

Inputs: Prefix prompt, the confounder name $\hat{U}_{\text{name}}$, the values of confounder $\mathcal{D}_X$, treatment $\mathcal{D}_T$ and outcome $\mathcal{D}_Y$

Outputs: Distribution parameter $\theta_i = \{\mu_i, \sigma_i\}$ for each individual i

Prompts:

{Prefix prompt}

The values of existing confounders, treatments, and outcomes are given by:

(1) Confounder Values: $\{\mathcal{D}_X\}$

(2) Treatment Values: $\{\mathcal{D}_T\}$

(3) Outcome Values: $\{\mathcal{D}_Y\}$

For the confounder $\{\hat{U}_{\text{name}}\}$, please specify a normal distribution (mean and standard deviation) for each individual from which we can sample the confounder value.

To accommodate the token limitations of LLMs, this prompt is executed in a mini-batch manner. Once the distribution parameters (e.g., mean and standard deviation in the case of a normal distribution) are obtained, we sample concrete values of the confounder $\hat{U}$ from the personalized distribution for each instance.

F.5 COUNTERFACTUAL OUTCOME IMPUTATION

To evaluate the effectiveness of the generated confounders, we assess whether the unconfoundedness assumption holds after incorporating them. Since LLMs implicitly encode a wide range of commonsense and domain-specific knowledge—including information related to potential hidden confounders—we utilize the LLM to impute counterfactual outcomes, $\hat{Y}^0$ and $\hat{Y}^1$. These counterfac-

tuals are then used to perform an empirical test of the unconfoundedness assumption via conditional independence analysis.

$\mathcal{P}_{\text{out}}(\mathcal{D}_X, \mathcal{D}_{\hat{U}}, \mathcal{D}_T, \mathcal{D}_Y)$: Imputing Counterfactual Outcomes

Inputs: Prefix prompt, the values of confounder $\mathcal{D}_X$, imputed confounder $\mathcal{D}_{\hat{U}}$, treatment $\mathcal{D}_T$ and outcome $\mathcal{D}_Y$

Outputs: Counterfactual outcomes in $\hat{Y}^0$ and $\hat{Y}^1$

Prompts:

{Prefix prompt}

The values of existing confounders, treatments, and outcomes are given by:

(1) Confounders: $\{\mathcal{D}_X\}$

(2) Treatments: $\{\mathcal{D}_T\}$

(3) Outcomes: $\{\mathcal{D}_Y\}$

Based on the OBSERVED DATA and your WORLD KNOWLEDGE, please infer the values of the counterfactual outcome corresponding to the alternative value of treatment.

G CASE STUDY

In this section, we provide two running samples from Jobs dataset for both variable generation and value imputation.

G.1 CASE ON VARIABLE GENERATION

Generating New Variable {TRANSPORTATION ACCESS}

Prompts:

Brief introduction of the {Jobs} dataset:

{The Jobs dataset is widely used in causal inference research for evaluating the performance of treatment effect estimation methods. It is constructed by combining experimental and observational data from the National Supported Work (NSW) demonstration and comparison group data (e.g., PSID or CPS).}

This observational dataset contains:

(1) Treatment — {Participation in Job Training}: {$T \in \{0, 1\}$ indicating whether the individual participated or did not participate in the job training program.}

(2) Outcome — {Employment Status}: {The individual's employment observed after the treatment decision. $Y = 1$ means employed, $Y = 0$ means not employed.}

(3) Confounders — {Age, Marriage, Previous Earning,...}: {Features affecting both the treatment and the outcome, such as age, education, prior income, etc.}

Based on your WORLD KNOWLEDGE, please propose one additional confounder which BOTH affects the treatment and outcome.

Make sure that the proposed confounder has a different meaning compared to existing confounders.
For this proposed confounder, please provide:
(1) A clear name for the confounder.
(2) A brief explanation of why it affects both treatment and outcome.

Answers:
(1) Variable Name: Transportation Access
(2) Explanation: Access to reliable transportation can influence both participation in the job training program (treatment) and subsequent employment (outcome). Individuals without transportation may be less likely to enroll in or attend the program due to logistical barriers. Similarly, lack of transportation can hinder job search efforts and commuting to workplaces, reducing the likelihood of employment.

G.2 Case on Value Imputation

Inferring Distribution Type for New Variable {Transportation Access}

Prompts:
Brief introduction of the {Jobs} dataset:
{The Jobs dataset is widely used in causal inference research for evaluating the performance of treatment effect estimation methods. It is constructed by combining experimental and observational data from the National Supported Work (NSW) demonstration and comparison group data (e.g., PSID or CPS).}
This observational dataset contains:
(1) Treatment — {Participation in Job Training}: {$T \in \{0, 1\}$ indicating whether the individual participated or did not participate in the job training program.}
(2) Outcome — {Employment Status}: {The individual's employment observed after the treatment decision. $Y = 1$ means employed, $Y = 0$ means not employed.}
(3) Confounders — {Age, Marriage, Previous Earning,...}: {Features affecting both the treatment and the outcome, such as age, education, prior income, etc.}

Based on your WORLD KNOWLEDGE, please provide the distribution type of confounder {Transportation Access}. For example:
(1) Continuous — e.g., Normal distribution
(2) Discrete — e.g., Multi-categorical distribution
(3) Binary — e.g., Bernoulli distribution

Answers:
(1) Distribution Type: Binary
(2) Value Description: 0 indicates the individual lacks reliable transportation (e.g.,

no personal vehicle or inadequate public transit access), and 1 indicates they have reliable transportation.

Estimating Values for New Binary Confounder {TRANSPORTATION ACCESS}

Prompts:

Brief introduction of the {Jobs} dataset:

{The Jobs dataset is widely used in causal inference research for evaluating the performance of treatment effect estimation methods. It is constructed by combining experimental and observational data from the National Supported Work (NSW) demonstration and comparison group data (e.g., PSID or CPS).}

This observational dataset contains:

(1) Treatment — {Participation in Job Training}: {$T \in \{0, 1\}$ indicating whether the individual participated or did not participate in the job training program.}

(2) Outcome — {Employment Status}: {The individual's employment observed after the treatment decision. $Y = 1$ means employed, $Y = 0$ means not employed.}

(3) Confounders — {Age, Education Years, ..., Previous Earnings}: {Features affecting both the treatment and the outcome, such as age, education, prior income, etc.}

The values of existing confounders, treatments, and outcomes are given by:

(1) Confounder Values: {[47,12,...,0], ..., [40, 8, ..., 0]}

(2) Treatment Values: {0,0,1,0,...,1}

(3) Outcome Values: {1,0,0,1,...,0}

For the new confounder {Transportation Access}, please specify the parameter p in $Bernoulli(p_i)$ distribution for each individual.

Answers:

(1) Probabilities: [0.12, 0.08, ..., 0.91]

(2) Summary: The new confounder *Transportation Access* is modeled as a latent binary variable generated from a conditional Bernoulli distribution $X_i \sim \text{Bernoulli}(p_i), \quad p_i = \sigma\big(f(X_i^{\text{obs}}, \mathcal{K}_{\text{LLM}})\big)$ where X_i^{obs} denotes the observed covariates, and $\mathcal{K}_{\text{LLM}}$ represents external knowledge encoded in the pretrained language model. The function $f(\cdot)$ integrates both observed features and socioeconomic knowledge (e.g., relationships between income, infrastructure, and mobility) to estimate the probability of transportation access.

This formulation captures transportation access as a latent socioeconomic factor informed not only by observed variables but also by broader contextual knowledge,

allowing the generated confounder to reflect patterns that may not be explicitly present in the dataset.

H BROADER IMPACTS

This work introduces ProCI, a framework that uses LLMs to help mitigate hidden confounding when estimating treatment effects from observational data. By combining structured and unstructured data with the general knowledge and reasoning abilities of LLMs, ProCI offers a new and useful way to improve causal inference. In practice, the ProCI framework can improve decision-making in areas where running controlled experiments is difficult or unethical, such as healthcare, social programs, and economic policies. It can help uncover hidden confounders that affect treatment and outcomes, leading to fairer and more informed decisions—especially in places with limited resources. Also, by reducing the need for expert-designed tools or domain knowledge, ProCI makes causal analysis easier and more available to a wider group of researchers and practitioners.

I LIMITATIONS

While our proposed method consistently outperforms base models across benchmarks, it still exhibits several limitations that warrant further investigation:

- Our evaluation is conducted on two widely-used observational benchmarks—Twins and Jobs. To better assess the generalizability of the ProCI framework, future work should explore a broader range of real-world and domain-specific datasets.
- As noted in Theorem 1, the empirical unconfoundedness test using Kernel Conditional Independence Test (KCIT) on imputed counterfactuals approximates the true test only when the sample size is sufficiently large. More robust or distribution-free statistical tests may be needed to relax this assumption in smaller datasets.
- While our study extends confounder generation to include four distinct LLMs—GPT-4o, DeepSeek-R1, LLaMA 3–8B, and Qwen2.5–7B—these models primarily represent instruction-tuned decoders. Future work should further examine ProCI's applicability across a wider range of architectures, such as multilingual models, encoder-decoder frameworks, or smaller-scale LLMs, to comprehensively assess its robustness and scalability.