Improving Treatment Effect Estimation with LLM-Based Data Augmentation

Nicolas Huynh [*1] **Julianna Piskorz** [*1] **Jeroen Berrevoets** [1] **Max Ruiz Luyten** [1] **Mihaela van der Schaar** [1]

Abstract

We introduce GATE, a framework which improves conditional average treatment effects (CATE) estimation in small-sample regimes. Our framework augments datasets with synthetic *counterfactual* outcomes using *pre-trained* generative models. Doing so addresses the covariate shift problem when inferring CATE from observational data. By using pre-trained generative models, GATE augments downstream CATE models with knowledge *beyond the training data*. In particular, we instantiate GATE with large language models (LLMs), which we show to work exceptionally well. LLMs utilize rich contextual information, such as dataset metadata, to generate outcomes grounded in real-world contexts. We demonstrate, both theoretically and empirically, that restricting augmentation to a carefully chosen subset of the covariate space can achieve performance gains—*even with imperfect generated outcomes.*

1. Introduction

Treatment Effects (CATE) inference is an active area of machine learning research [41, 63], with critical applications in healthcare [28], economics [11], and marketing [32]. However, most modern machine learning methods, including those developed for CATE estimation [63, 64], are designed to leverage large datasets. while *real-world problems often lack this luxury.* Data scarcity amplifies a second critical challenge in CATE estimation: *covariate shift*. Non-random treatment assignment shifts the covariate distributions of treated individuals compared to the controlled individuals, resulting in biased or high-variance treatment effect estimates [40] particularly in small-sample regimes [4].

Existing CATE estimation methods typically entail new *model* specifications (e.g. inverse-propensity weighting [1], representation learning [39, 63], or both [8, 30]). Such methods are ill-equipped to handle the challenges of data scarcity and covariate shift as they suffer from high variance estimates [19]. Instead, we propose a *complementary and simple solution*, to be used *alongside* a CATE learner: addressing covariate shift by manipulating the *dataset* rather than the model. Augmenting the observational dataset with missing potential outcomes, sampled from a generative model, not only increases the sample size but also directly *mitigates the covariate shift*. While previous works have proposed imputing pseudo-outcomes using a GAN model [71], or local regression methods [6], we note a major difference: our generative model can rely on information *outside of the observed dataset*.

We propose GATE (*Generative Augmentation for Treatment Effect estimation*), a *flexible and simple* data augmentation framework that leverages pre-trained generative models, in particular large language models (LLMs). The key advantage of using LLMs as generative models lies in their rich prior knowledge obtained via extensive pre-training. For example, LLMs have access to metadata in observational datasets, these include textual descriptions of covariates or other contextual information. By utilizing these metadata, LLMs can generate potential outcomes grounded in real-world contexts.

But can LLMs be trusted? There are valid concerns about the robustness of employing LLMs in the causal setting, particularly given their propensity for hallucinations. As such, GATE *restricts* augmentation to the *admissible set*, a carefully selected subset of the covariate space where we expect the generative model's predictions to be most reliable. This is motivated by our theoretical analysis that shows a trade-off between covariate shift reduction and the accuracy of the generative model.

In our experiments, we demonstrate that GATE improves the performance of a range of CATE models on three datasets, while reducing the performance gap between learners.

2. GATE

2.1. Conditional average treatment effects (CATEs).

Let $\mathcal{D}^{(\text{obs})} = \{(X_i, T_i, Y_i)\}_{i=1}^n$ be an observational dataset such that $(X_i, T_i, Y_i) \overset{\text{i.i.d.}}{\sim} P(X, T, Y)$, where $Y_i \in \mathcal{Y}$ is a

[*]Equal contribution [1]Department of Applied Mathematics and Theoretical Physics, University of Cambridge. Correspondence to: Nicolas Huynh <nvth2@cam.ac.uk>, Julianna Piskorz <jp2048@cam.ac.uk>.

Proceedings of the 1st ICML Workshop on Foundation Models for Structured Data, Vancouver, Canada. 2025.

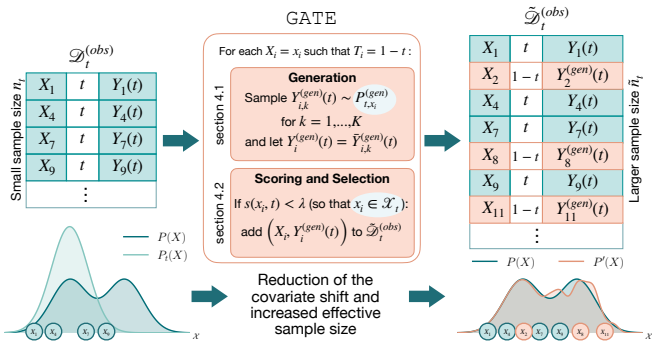

Figure 1: **Overview.** GATE enhances CATE estimation in the finite-sample regime through selective data augmentation. For a fixed treatment t, synthetic potential outcomes for samples with $T_i = 1 - t$ are generated using $P_{t,x}^{(\text{gen})}$ and scored via $s(x, t)$ to decide their inclusion in $\tilde{\mathcal{D}}_t^{(\text{obs})}$.

continuous or binary outcome, $X_i \in \mathcal{X} \subset \mathbb{R}^d$ is a vector of covariates and $T_i \in \{0, 1\}$ is a binary treatment assignment. For conciseness, we ignore the sample subscript i unless explicitly needed. We assume that there are two possible *potential outcomes*: $Y(0)$ (no treatment) and $Y(1)$ (under treatment) [59]. From $\mathcal{D}^{(\text{obs})}$ we aim to estimate the CATE:

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \,|\, X = x] = \mu_1(x) - \mu_0(x), \quad (1)$$

where $\mu_t(x) = \mathbb{E}[Y(t) \,|\, X = x]$. We make the standard [59] assumptions of *overlap* ($0 < \mathbb{P}(T = 1 \,|\, X = x) < 1 \;\forall x \in \mathcal{X}$), *ignorability* ($(Y(1), Y(0)) \perp T \,|\, X$), and *consistency* ($Y = Y(t)$ if $T = t$). We also define $P_t = P(X \mid T = t)$ and $\pi_t = \mathbb{P}(T = t)$. Based on eq. 1, we see that it is sufficient to estimate the Conditional Average Potential Outcomes (CAPOs) $\mu_0(x)$ and $\mu_1(x)$ from $\mathcal{D}_0^{(\text{obs})}$ and $\mathcal{D}_1^{(\text{obs})}$ respectively, where $\mathcal{D}_t^{(\text{obs})} = \{(X_i, Y_i) \in \mathcal{D}^{(\text{obs})} \,|\, T_i = t\}$. However, inferring CAPOs from finite observational data $\mathcal{D}_t^{(\text{obs})}$ is challenging for two main reasons [40]: *large variance* in *small datasets*, and *covariate shift*, where $P_t(X)$ is typically different from $P_{1-t}(X)$.

2.2. Data augmentation via potential outcomes

Our key insight is that both model variability and covariate shift can be mitigated easily via *data augmentation*. Thus we propose GATE, a framework where we generates the missing potential outcomes for individuals in $\mathcal{D}^{(\text{obs})}$ using **a generative model** $P_{t,x}^{(gen)}$ which allows to sample $Y^{(\text{gen})}(t) \mid X = x \sim P_{t,x}^{(gen)}$ for all $x \in \mathcal{X}$. We use this model to create an additional sample $(X, 1 - T, Y^{(\text{gen})}(1 - T))$ for every (X, T, Y) and add it to $\mathcal{D}^{(\text{obs})}$ to create the augmented dataset $\tilde{\mathcal{D}}^{(\text{obs})}$ (see Figure 1). A perfect generator fixes the covariate shift problem completely as the augmented dataset $\tilde{\mathcal{D}}_t^{(\text{obs})}$ would comprise the same covariates as $\tilde{\mathcal{D}}_{1-t}^{(\text{obs})}$. Furthermore, we would have $|\tilde{\mathcal{D}}_t^{(obs)}| = n \gg \pi_t \cdot n$, mitigating the variance problem.

However, realistically, the generator may be inaccurate in at least some areas of the covariate space $\mathcal{X}$. As such, we need to balance the bias introduced by $P_{t,x}^{(\text{gen})}$ with the benefits obtained by mitigating the covariate shift. To this end, it might be better to generate the counterfactual potential outcomes only in a select subset of the covariate space, which we call the admissible set $\mathcal{X}_t$. We formalize this intuition theoretically via a generalization bound on the expected risk of the CAPOs presented in Appendix C.

The main takeaway is that performance gains can be obtained even in the face of potential inaccuracy of $P_{t,x}^{(\text{gen})}$, and manipulating $\mathcal{X}_t$ for a given generative model $P_{t,x}^{(\text{gen})}$ allows to navigate the trade-off between the bias introduced by the inaccuracy of the generator, and the reduction of variance and covariate shift achieved via data augmentation. How to best construct $\mathcal{X}_t$ is discussed next.

3. Augmentation with LLMs

We describe an instantiation of GATE using LLMs, which defines the admissible set $\mathcal{X}_t$ using the LLM's uncertainty.

3.1. LLMs as potential outcome generators

While GATE can be used with generative models trained exclusively on observational data $\mathcal{D}^{(\text{obs})}$– such as local regression models [6]) –the utility of such solutions is inherently limited as such a setup ultimately utilises the same information as the downstream CATE model. To overcome this limitation, we propose to use GATE with *foundation models*, such as large language models (LLMs) [13]. Due to extensive pre-training LLMs encode rich domain knowledge, *outside the scope of* $\mathcal{D}^{(\text{obs})}$. They can leverage dataset metadata such as covariate descriptions or the context of data collection to align their outputs with the specific problem domain, integrating contextual relationships that may not be present in $\mathcal{D}^{(\text{obs})}$ [57, 65]. Additionally, LLMs excel at few-shot learning, allowing them to adapt to a given task when conditioned on $\mathcal{D}^{(\text{obs})}$.

Prompting strategies. We guide the extraction of the prior knowledge of the LLM by including in the prompts information such as: natural language descriptions of covariates, information about the data collection technique, the population of the study or more general context of the dataset. We also exploit the few-shot learning capabilities of the LLM by conditioning the generation on a randomly chosen subset of samples from the observational dataset $\mathcal{D}^{(obs)}$ to exploit the LLM's in-context learning abilities.

Stochastic nature of the LLM. We sample K potential outcomes from $P_{t,x}^{(\text{gen})}$: $Y_{i,k}^{(gen)}(t) \sim P_{t,x_i}^{(gen)}, k = 1, \dots, K$, for every X_i that requires augmentation. To improve the robustness of the generation, we then average these samples and set $Y_i^{(gen)}(t) = \bar{Y}_{i,k}^{(gen)}(t)$ (see Figure 1).

3.2. Choosing the admissible set $\mathcal{X}_t$

Constructing $\mathcal{X}_t$ can be guided by excluding regions of the covariate space $\mathcal{X}$ where the distribution $P_{t,x}^{(\text{gen})}$ significantly deviates from the true distribution $P_{t,x}$. However, assessing the statistical distance between these two distributions is challenging as the potential outcomes $Y(t)$ are not all observed. As such, we propose a scoring function $s(x,t)$, which is chosen as a *proxy* for the fidelity of $P_{t,x}^{(\text{gen})}$ at a given point $x \in \mathcal{X}$.

Choosing the scoring function s for LLMs. For stochastic models such as LLMs, we propose to rely on the variance in the generated outcomes to define the admissible set, by setting $s(x,t) := \text{Var}_{Y^{(\text{gen})}(t)\sim P_{t,x}^{(\text{gen})}}(Y^{(\text{gen})}(t) \,|\, X = x)$ (approximated by the empirical variance from K outcomes sampled per (x,t)). While this scoring function might not always be optimal, it reflects the heuristic that the accuracy of the $P_{t,x}^{(\text{gen})}$ might be lower in the areas where the generative model is less certain about its predictions. In the context of LLMs, it has been shown that uncertainty measures such as variance can be used to discriminate between factually correct and incorrect responses [26, 37], as well as predict the quality of a response [47] (see Appendix B.3 for more details). Given a fixed parameter $\alpha \in [0,1]$, we define an adaptive threshold $\lambda(\alpha, \mathcal{D}^{(\text{obs})}) = \text{Quantile}_\alpha(\{s(X_i,T_i) \mid i \in [n]\})$, corresponding to a percentile-based threshold to easily control the proportion of generated potential outcomes across datasets. We finally define the admissible set $\mathcal{X}_t$ as

$$\mathcal{X}_t = \{X_i \mid i \in [n], s(X_i,T_i) < \lambda(\alpha, \mathcal{D}^{(\text{obs})})\}, \quad (2)$$

comprising the samples for which the variance in the generated potential outcomes is below the α-quantile. We provide a detailed discussion on this definition in Appendix B.3.

4. Numerical Experiments

Data. Evaluating CATE models using observational data is challenging due to the lack of ground-truth CATE values. Standard benchmarks like IHDP [33] or News [39] address this by designing artificial potential outcome functions. However, since we aim to compare generative models trained on $\mathcal{D}^{(\text{obs})}$ with those trained on external (real-world) datasets, the outcome's relationship with treatment and covariates must be *reality-grounded*. Consequently, we utilize the following datasets: **Lalonde CPS1** [44]; **STAR Project** [3]; and **Hillstrom** [34]. More details are in Appendix D.

CATE Models. We compare the performance of downstream CATE models when trained on the original dataset $\mathcal{D}^{(\text{obs})}$ vs. when trained on the augmented $\tilde{\mathcal{D}}^{(\text{obs})}$. We use the S-, T-, X-, R-, IPW- and DR-learner [19, 43]. We also consider in Appendix F.1 the CFR-Wass, CFR-MMD algorithms [63], TARNet [63], DragonNet [64] and BART [9].

Instantiating GATE. We use GPT-3.5 Turbo [2] when instantiating GATE with LLMs. In Section 4.2, we compare the performance GATE with an LLM against a diverse set of models trained on $\mathcal{D}^{(\text{obs})}$: mean model, 1-nearest neighbour (1-NN), random forest (RF), GAN (following the approach of GANITE [71]). We also consider TabPFN v2 [35], a *foundation model* for tabular data. Detailed descriptions of the experiments can be found in Appendix D, anonymised code to reproduce the experiments can be found here.

4.1. Does GATE improve CATE estimation?

Setup. Each CATE model is trained on both the original dataset and the GATE-augmented dataset, and we compare the PEHE in each case. (additional parameters such as architecture or hyperparameters remain fixed across settings).

Results. Table 1 shows that *GATE consistently improves performance* across all the considered CATE models, with gains across the average PEHE [33] and its standard deviation. Furthermore, *GATE decreases the performance gap across CATE learners*, making it a *model agnostic* data pre-processing step that can aid model selection, usable with both one-step and two-step learners. Results for other learners are in Appendix F.1.

4.2. How do LLMs compare to other generative models?

Setup. We train each baseline $P_{t,x}^{(\text{gen})}$ on $\mathcal{D}_t^{(\text{obs})}$. Performance comparisons are conducted for the DR learner across three datasets of varying sizes, randomly sampled with proportions $\rho \in \{0.1, 0.5, 1.0\}$ from the original observational datasets. Each baseline uses the same admissible sets $\mathcal{X}_t$ as the LLM for a fair comparison. Results with $\mathcal{X}_t = \mathcal{X}$ can be found in Appendix F.6.

Results. In Figure 2, we present the average $\sqrt{\epsilon_{\text{PEHE}}}$ obtained across 3 seeds when instantiating GATE with each of the generative models. We note that multiple models can offer performance improvements to the downstream CATE model, compared to the no-augmentation baseline. Interestingly, the *LLM consistently outperforms the baselines trained on $\mathcal{D}^{(\text{obs})}$ only*, yielding lower average PEHE. As predicted, this performance gap is most evident in small sample regimes ($\rho = 0.1$), where LLM-derived prior knowledge proves most beneficial. Remarkably, comparing the PEHE across all three dataset sizes demonstrates that using GATE with the LLM allows to obtain performance levels which are close to optimal when using only a fraction of the original dataset. The performance gap between LLMs and other generative models is most pronounced in the STAR dataset, which exhibits high treatment effect heterogeneity. Conversely, this gap is minimal in the Hillstrom dataset, where outcome heterogeneity is low (cf. Appendix F.6). Low heterogeneity explains the strong performance of the mean imputation model, as the generated *constant* potential outcomes effectively regularize the downstream model.

Table 1: GATE improves the performance of different CATE learners across the datasets without data augmentation (✗), and with data augmentation (✓). Average $\sqrt{\epsilon_{\text{PEHE}}}$ and 1std is reported for 3 seeds (↓ is better)

Learner	Lalonde CPS1D		STAR		Hillstrom	
	✗	✓	✗	✓	✗	✓
S-learner	1.09 ± 0.07	0.95 ± 0.01	0.78 ± 0.10	0.56 ± 0.02	0.32 ± 0.03	0.25 ± 0.01
T-learner	1.28 ± 0.03	0.96 ± 0.01	0.81 ± 0.08	0.50 ± 0.03	0.4 ± 0.01	0.24 ± 0.01
X-learner	1.43 ± 0.10	0.95 ± 0.01	0.93 ± 0.05	0.49 ± 0.02	0.29 ± 0.01	0.24 ± 0.01
R-learner.	1.35 ± 0.42	0.95 ± 0.00	6.12 ± 2.57	0.47 ± 0.01	0.63 ± 0.21	0.26 ± 0.02
IPW-learner.	1.12 ± 0.03	0.95 ± 0.01	0.57 ± 0.06	0.47 ± 0.01	0.29 ± 0.01	0.25 ± 0.00
DR-learner	1.29 ± 0.02	0.95 ± 0.01	0.60 ± 0.11	0.48 ± 0.02	0.41 ± 0.02	0.25 ± 0.01

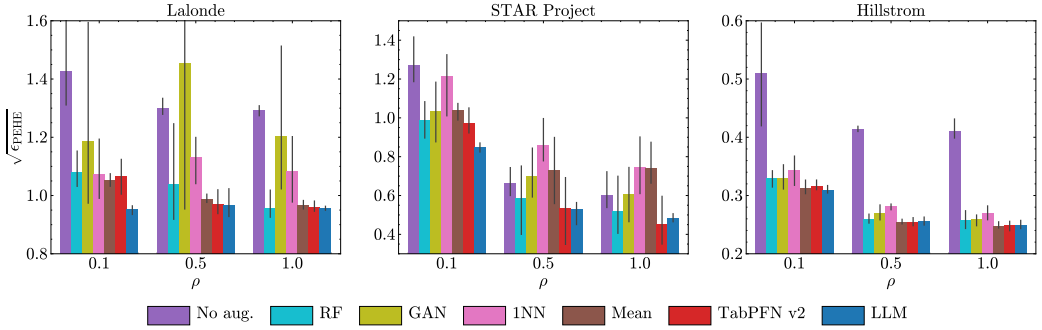

Figure 2: **Comparison of generative models in GATE.** The LLM outperforms the models trained on $\mathcal{D}^{(\text{obs})}$ across different proportions ρ. The error bars mark 1std, computed across 3 seeds.

4.3. Where does the benefit of using LLMs come from?

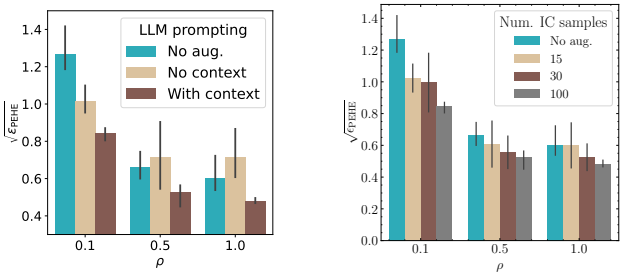

Figure 3: **Left:** Contextual information helps to achieve good performance in low-samples. **Right:** In-context learning. For both figures, the error bars mark 1std (3 seeds).

Setup. We quantify the influence of prior knowledge by performing an ablation where we remove all the contextual information in the prompt given to the LLM and give the features generic names (e.g. *Feature 1*). Hence effectively only the in-context samples are provided. We compare the performance of this context-deprived LLM with an LLM which is informed about the context of the dataset and feature names (see Appendix E for exact prompts). We do this for a varying number of samples between $\{15, 30, 100\}$. We perform both experiments on the STAR dataset, with the DR-learner as the learner across varying proportions, ρ.

Results. Figure 3 Left shows substantial performance gains when using STAR's meta-data to elicit the prior knowledge of the LLM, particularly when ρ is small. As ρ increases, the performance gap between context-informed LLM and the no-augmentation-baseline naturally becomes smaller. Furthermore, Figure 3 Right demonstrates that including more in-context samples in the prompt improves the downstream performance. Results for other datasets are in Appendix F.7.

4.4. Additional results.

In the interest of space, we relegated additional results in Appendix F, including the impact of augmentation on covariate shift reduction, a comparison with other selector functions and per-dataset results.

5. Discussion

We have presented GATE, a framework which improves CATE estimation in small-sample regimes using pre-trained generative models, in particular LLMs. We demonstrate that using generated potential outcomes is an effective way to inject prior knowledge beyond the observation data. Although in this work we have focused on binary treatment assignments, our framework could naturally be extended to settings with multiple, continuous or even time-varying treatments. Further work could also be dedicated to devising statistical methods to obtain confidence intervals for the CATE estimate, borrowing ideas from the supervised learning setting [7]. Finally new methods for the assessment of LLMs generations will naturaly benefit GATE.

References

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

[2] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Achilles, C. M., Bain, H. P., Bellott, F., Boyd-Zaharias, J., Finn, J., Folger, J., Johnston, J., and Word, E. Tennessee's Student Teacher Achievement Ratio (STAR) project, October 2008.

[4] Alaa, A. and van der Schaar, M. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *International Conference on Machine Learning*, pp. 129–138. PMLR, 2018.

[5] Alaa, A. M. and Van Der Schaar, M. Bayesian inference of individualized treatment effects using multi-task gaussian processes. *Advances in neural information processing systems*, 30, 2017.

[6] Aloui, A., Dong, J., Le, C. P., and Tarokh, V. Counterfactual Data Augmentation with Contrastive Learning, November 2023. URL http://arxiv.org/abs/2311.03630. arXiv:2311.03630 [cs, stat].

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

[8] Assaad, S., Zeng, S., Tao, C., Datta, S., Mehta, N., Henao, R., Li, F., and Carin, L. Counterfactual representation learning with balancing weights, 2021. URL https://arxiv.org/abs/2010.12618.

[9] Athey, S. and Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, July 2016. doi: 10.1073/pnas.1510489113. Publisher: Proceedings of the National Academy of Sciences.

[10] Ban, T., Chen, L., Wang, X., and Chen, H. From query tools to causal architects: Harnessing large language models for advanced causal discovery from data. *arXiv preprint arXiv:2306.16902*, 2023.

[11] Baum-Snow, N. and Ferreira, F. Causal inference in urban and regional economics. In *Handbook of regional and urban economics*, volume 5, pp. 3–68. Elsevier, 2015.

[12] Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.

[13] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[14] Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45, 2015.

[15] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[16] Choi, K., Cundy, C., Srivastava, S., and Ermon, S. LM-Priors: Pre-Trained Language Models as Task-Specific Priors, October 2022. arXiv:2210.12530 [cs].

[17] Cortes, C., Mansour, Y., and Mohri, M. Learning bounds for importance weighting. *Advances in neural information processing systems*, 23, 2010.

[18] Curth, A. and Van der Schaar, M. On inductive biases for heterogeneous treatment effect estimation. *Advances in Neural Information Processing Systems*, 34: 15883–15894, 2021.

[19] Curth, A. and Van der Schaar, M. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pp. 1810–1818. PMLR, 2021.

[20] Curth, A. and Van Der Schaar, M. In search of insights, not magic bullets: Towards demystification of the model selection dilemma in heterogeneous treatment effect estimation. In *International Conference on Machine Learning*, pp. 6623–6642. PMLR, 2023.

[21] Curth, A., Svensson, D., Weatherall, J., and van der Schaar, M. Really doing great at estimating cate? a critical look at ml benchmarking practices in treatment effect estimation. 2021.

[22] de Vassimon Manela, D., Battaglia, L., and Evans, R. Marginal causal flows for validation and inference. *Advances in Neural Information Processing Systems*, 37:9920–9949, 2024.

[23] Dehejia, R. H. and Wahba, S. Causal effects in non-experimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062, 1999.

[24] Dehejia, R. H. and Wahba, S. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, 84(1):151–161, 2002.

[25] Du, Y., Watkins, O., Wang, Z., Colas, C., Darrell, T., Abbeel, P., Gupta, A., and Andreas, J. Guiding Pre-training in Reinforcement Learning with Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 8657–8677. PMLR, July 2023. ISSN: 2640-3498.

[26] Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

[27] Gentzel, A. M., Pruthi, P., and Jensen, D. How and why to use experimental data to evaluate methods for observational causal inference. In *International Conference on Machine Learning*, pp. 3660–3671. PMLR, 2021.

[28] Gershon, A. S., Lindenauer, P. K., Wilson, K. C., Rose, L., Walkey, A. J., Sadatsafavi, M., Anstrom, K. J., Au, D. H., Bender, B. G., Brookhart, M. A., Dweik, R. A., Han, M. K., Joo, M. J., Lavergne, V., Mehta, A. B., Miravitlles, M., Mularski, R. A., Roche, N., Oren, E., Riekert, K. A., Schoenberg, N. C., Stukel, T. A., Weiss, C. H., Wunsch, H., Africk, J. J., and Krishnan, J. A. Informing healthcare decisions with observational research assessing causal effect. an official american thoracic society research statement. *American Journal of Respiratory and Critical Care Medicine*, 203(1):14–23, 2021. doi: 10.1164/rccm.202010-3943ST. PMID: 33385220.

[29] Hahn, P. R., Murray, J. S., and Carvalho, C. M. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

[30] Hassanpour, N. and Greiner, R. Counterfactual regression with importance sampling weights. In *IJCAI*, pp. 5880–5887. Macao, 2019.

[31] He, Q., Wang, Y., and Wang, W. Can Language Models Act as Knowledge Bases at Scale?, February 2024. arXiv:2402.14273 [cs].

[32] Hill, D. N., Moakler, R., Hubbard, A. E., Tsemekhman, V., Provost, F., and Tsemekhman, K. Measuring causal impact of online actions via natural experiments: Application to display advertising. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1839–1847, 2015.

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

[34] Hillstrom, K. The minethatdata e-mail analytics and data mining challenge, 2008. URL https://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html.

[35] Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T., and Hutter, F. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

[36] Horvitz, D. G. and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47 (260):663–685, 1952.

[37] Huang, Y., Song, J., Wang, Z., Chen, H., and Ma, L. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*, 2023.

[38] Jiralerspong, T., Chen, X., More, Y., Shah, V., and Bengio, Y. Efficient causal graph discovery using large language models. *arXiv preprint arXiv:2402.01207*, 2024.

[39] Johansson, F., Shalit, U., and Sontag, D. Learning Representations for Counterfactual Inference. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 3020–3029. PMLR, June 2016. ISSN: 1938-7228.

[40] Johansson, F. D., Shalit, U., Kallus, N., and Sontag, D. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *Journal of Machine Learning Research*, 23(166):1–50, 2022.

[41] Kennedy, E. H. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, January 2023. ISSN 1935-7524, 1935-7524. doi: 10.1214/23-EJS2157. Publisher: Institute of Mathematical Statistics and Bernoulli Society.

[42] Kennedy, E. H., Ma, Z., McHugh, M. D., and Small, D. S. Non-parametric Methods for Doubly Robust Estimation of Continuous Treatment Effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4):1229–1245, September 2017. ISSN 1369-7412. doi: 10.1111/rssb.12212.

[43] Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

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

[45] Lan, H. and Syrgkanis, V. Causal Q-aggregation for CATE model selection. In Dasgupta, S., Mandt, S., and Li, Y. (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4366–4374. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/lan24a.html.

[46] Li, Q., Yang, X., Wang, H., Wang, Q., Liu, L., Wang, J., Zhang, Y., Chu, M., Hu, S., Chen, Y., Shen, Y., Fan, C., Zhang, W., Xu, T., Gu, J., Zheng, J., and Group, G. Z. A. From Beginner to Expert: Modeling Medical Knowledge into General LLMs, January 2024. arXiv:2312.01040 [cs] version: 3.

[47] Lin, Z., Trivedi, S., and Sun, J. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023.

[48] Long, M., Cao, Y., Wang, J., and Jordan, M. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pp. 97–105. PMLR, 2015.

[49] Long, S., Schuster, T., and Piché, A. Can large language models build causal graphs?, February 2024. arXiv:2303.05279 [cs].

[50] Mahajan, D., Mitliagkas, I., Neal, B., and Syrgkanis, V. Empirical analysis of model selection for heterogeneous causal effect estimation. *arXiv preprint arXiv:2211.01939*, 2022.

[51] Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

[52] Nagalapatti, L., Iyer, A., De, A., and Sarawagi, S. Continuous treatment effect estimation using gradient interpolation and kernel smoothing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14397–14404, 2024.

[53] Neal, B., Huang, C.-W., and Raghupathi, S. Realcause: Realistic causal inference benchmarking. *arXiv preprint arXiv:2011.15007*, 2020.

[54] Nie, X. and Wager, S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2): 299–319, June 2021. ISSN 0006-3444. doi: 10.1093/biomet/asaa076.

[55] Nori, H., King, N., McKinney, S. M., Carignan, D., and Horvitz, E. Capabilities of GPT-4 on Medical Challenge Problems, April 2023. arXiv:2303.13375 [cs].

[56] Parikh, H., Varjao, C., Xu, L., and Tchetgen, E. T. Validating causal inference methods. In *International conference on machine learning*, pp. 17346–17358. PMLR, 2022.

[57] Requeima, J., Bronskill, J., Choi, D., Turner, R., and Duvenaud, D. K. Llm processes: Numerical predictive distributions conditioned on natural language. *Advances in Neural Information Processing Systems*, 37: 109609–109671, 2024.

[58] Richens, J. and Everitt, T. Robust agents learn causal world models, February 2024. arXiv:2402.10877 [cs].

[59] Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[60] Saito, Y. and Yasui, S. Counterfactual cross-validation: Stable model selection procedure for causal inference models. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8398–8407. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/saito20a.html.

[61] Schuler, A., Baiocchi, M., Tibshirani, R., and Shah, N. A comparison of methods for model selection when estimating individual treatment effects, 2018.

[62] Seedat, N., Huynh, N., van Breugel, B., and van der Schaar, M. Curated LLM: Synergy of LLMs and Data Curation for tabular augmentation in ultra low-data regimes, February 2024. arXiv:2312.12112 [cs].

[63] Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pp. 3076–3085. PMLR, 2017.

[64] Shi, C., Blei, D., and Veitch, V. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

[65] Shysheya, A., Bronskill, J., Requeima, J., Siddiqui, S. A., Gonzalez, J., Duvenaud, D., and Turner, R. E. Jolt: Joint probabilistic predictions on tabular data using llms. *arXiv preprint arXiv:2502.11877*, 2025.

[66] Takayama, M., Okuda, T., Pham, T., Ikenoue, T., Fukuma, S., Shimizu, S., and Sannai, A. Integrating large language models in causal discovery: A statistical causal approach. *arXiv preprint arXiv:2402.01454*, 2024.

[67] Wager, S. and Athey, S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523):1228–1242, July 2018. ISSN 0162-1459. doi: 10.1080/01621459.2017.1319839. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2017.1319839.

[68] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[69] Willig, M., Zečević, M., Dhami, D. S., and Kersting, K. Can Foundation Models Talk Causality?, December 2022. arXiv:2206.10591 [cs].

[70] Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[71] Yoon, J., Jordon, J., and Van Der Schaar, M. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International conference on learning representations*, 2018.

[72] Zečević, M., Willig, M., Dhami, D. S., and Kersting, K. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal, August 2023. arXiv:2308.13067 [cs].

[73] Zhu, M., Stanivuk, S., Petrovic, A., Nikolic, M., and Lio, P. Incorporating LLM Priors into Tabular Learners. November 2023. arXiv: 2311.11628.

Appendix: GATE

Table of Contents

A. Related Works

Methods for CATE estimation. Machine learning methods for CATE estimation can be broadly divided into two categories: model-specific and model-agnostic methods. **Method-specific** approaches rely on adjusting specific machine learning methods to the treatment effect setting. This gives rise to solutions based on neural networks [18, 39, 63, 64], Gaussian processes [5] or random forests and regression trees [9, 29, 33, 67].

In contrast, **model-agnostic** methods (so-called 'meta-learners' [19, 43]) are general learning strategies which can be instantiated with any base learner (e.g., neural network, random forest). Within the model-agnostic strategies we can

distinguish the *one-step learners*, which directly estimate the potential outcome surfaces, $\hat{\mu}_0$ and $\hat{\mu}_1$, and then obtain the CATE as: $\hat{\tau} = \hat{\mu}_1 - \hat{\mu}_0$. The alternative *two-step learners* [19, 42, 54] implicitly rely on ideas from data imputation. In their first step, two-step learners obtain pseudo outcomes $\tilde{Y}_\phi$, which are "proximal" target treatment effect values, composed from nuisance parameters $\phi = (\pi, \mu_0, \mu_1)$ estimated from the given observational dataset $\mathcal{D}^{(\text{obs})}$ (or a subset of it). In the second step, the final CATE model is obtained by regressing the pseudo-outcomes $\tilde{Y}_\phi$ on the covariates X. The pseudo-outcomes can be obtained using strategies relying on propensity weighting (IPW-learner [36]), regression-adjustment (X-learner [43]) or both of these combined (DR-learner [41]). We note that our framework GATE is a strategy which is complementary to these standard two-step learners.

How is GATE different from a standard two-step learner?

1. **Admissible set:** Two-step learners require obtaining the missing potential outcomes for *all* individuals in the observational dataset. As we demonstrate, this might introduce excessive bias if the generative model $P_{t,x}^{(\text{gen})}$ is inaccurate. As a solution to this problem, in our framework we introduce the concept of an admissible set $\mathcal{X}_t$, which allows to navigate the trade-off between the reduction of the covariate shift and the introduced bias. In the case where $\mathcal{X}_t \neq \mathcal{X}$, GATE does not allow to explicitly obtain treatment effect proxies for all individuals in the dataset, making it different from a standard two-step learner.

2. **External information:** A standard two-step learning strategy does not allow to utilise external sources of information to inform the generation of the pseudo-outcomes, as the nuisance parameters ϕ estimated in the first step are fitted using the observational data $\mathcal{D}^{(\text{obs})}$ only. In contrast, a key defining characteristic of GATE is that it allows to infuse the downstream CATE estimator with external knowledge, by training the generative model $P_{t,x}^{(\text{gen})}$ on datasets different from $\mathcal{D}^{(\text{obs})}$.

3. **Complementary inductive biases:** Considering our method as a pre-processing data augmentation method allows to aggregate the inductive biases imposed by the generative model $P_{t,x}^{(\text{gen})}$ and the downstream CATE learner used on the augmented dataset, $\hat{\tau}$. Particularly in small sample regimes, when the observational dataset $\mathcal{D}^{(\text{obs})}$ does not contain sufficient information to confidently estimate the CATE function, combining the inductive biases imposed by different methods might be particularly beneficial. This is why in our experiments (Section 4) we fit two-step learners on top of the GATE-augmented dataset, demonstrating performance improvements.

As such, GATE can be used as the first step in the meta-learning pipeline, without requiring any change to the standard meta-learners.

Data augmentation for CATE estimation. Other works have proposed alternative model-specific instantiation of the two-step learning strategy, relying for example on obtaining the pseudo-outcome using a GAN model [71], or local regression methods [6]. However, these imputation approaches are constrained by the amount of information present in the observational datasets. As a result, they are particularly vulnerable to scenarios with covariate shift, where there are significant differences between the distributions of the control and treated groups. Furthermore, these imputation methods (GAN and local regression) require large amounts of data to be accurate, which contrasts the small-sample regime tackled in this work. On the other hand, GATE provides a principled way of leveraging models trained on external data sources, such as the LLMs, and thus is able to take advantage of the dataset metadata to set the context, leading to helpful data augmentation as shown in Section 4.1. [52] uses imputation for CATE estimation, however it relies on the differentiability of the outcome functions with respect to the continuous treatment t, and thus is not directly applicable to our setting. An orthogonal strand of litterature [22, 56] uses generative models to validate, evaluate, or assess the performance of various causal inference methods, whereas GATE 's objective is to improve the performance of CATE learners.

LLMs as sources of prior knowledge for downstream tasks. As large language models (LLMs) have increased in parameter count and training set size, it has become clear that they are able to act as knowledge bases, showing great performance across a variety of knowledge-retrieval tasks [15, 31, 46, 55]. As such, LLMs have been proposed as tools for extracting prior knowledge about the world, which can be used to ground standard *data-driven* ML models in real-world contexts and encourage their outputs to be consistent with common-sense reasoning based on the meta-data [16]. Relying on the inductive biases generated by the LLMs from the task-specific metadata has been demonstrated to improve performance on tasks as diverse as reinforcement learning [16, 25], tabular learning [62, 73] as well as causal discovery [10, 16, 38, 66].

Furthermore, recent studies demonstrate that LLMs are in principle capable of gaining knowledge about the underlying causal structure of real-world data generating processes, despite not being explicitly trained to reason 'causally' [10, 58, 72**?**]. This finding is further supported by empirical research, demonstrating LLMs potential for answering causal queries [49, 69, 72]. This further motivates the use of LLMs as sources of prior knowledge and relevant inductive biases in *causal inference* tasks in particular.

Comparison with domain adaption. Our bound in **??** is related to a series of works studying generalisation theory for unsupervised domain adaptation [12, 48, 51], but differs in significant ways. These bounds involve the risk in the target domain ($D = 1$) using the observed risk in the source domain ($D = 0$) and the distance between the domains:

$$R_{D=1}(f) \leq R_{D=0}(f) + d_{\mathcal{H}}(P(X|D=1), P(X|D=0)) + \lambda_{\mathcal{H}},$$

where $\mathcal{H}$ is some function class and $\lambda_{\mathcal{H}}$ is a constant. Unique in our bound is the use of the distribution Q to split $d_{\mathcal{H}}$ into the terms $IPM_{\mathcal{L}}(Q, P_{1-t})$ and $\mathbb{E}_{X\sim Q}[IPM_{\mathcal{L}^x}(P(Y(t)|X = x), P(Y^{(gen)}(t)|X = x))]$ which correspond respectively to the covariate shift and the bias introduced by the generator. Through this, our bound offers the following novel insights which we validate experimentally:

1. **Insight**: Even when the generative model $P_{t,x}^{(gen)}$ is imperfect, using it to target covariate shift can improve performance.

   → **Experiment 6.5.2**: We explicitly compare the bias introduced by the generative model $P_{t,x}^{(\text{gen})}$ against the reduction of the covariate shift obtained by data augmentation, showing that these two effects can be balanced.
   → **Experiment 6.4**: We verify this by comparing the performance with and without GATE across three datasets and multiple CATE models.

2. **Insight**: Tuning the distribution Q via the admissible set $\mathcal{X}_t$ allows to balance the trade-off between the bias introduced by the generator, and the reduction of variance and covariate shift achieved via data augmentation.

   → **Experiment 6.5.2**: We verify that modulating $\mathcal{X}_t$ allows to navigate this trade-off.

3. **Insight**: Excluding from the admissible set regions of the covariate space where the generative model is particularly "incorrect" can improve performance.

   → **Experiment 6.5.2**: We propose to identify such regions using a proxy measure: the uncertainty in the generated outcomes. We verify that as we increase the allowed level of uncertainty of $P_{t,x}^{(gen)}$, the bias introduced by data augmentation increases, while the covariate shift decreases.

B. Details on GATE

B.1. Usage with CATE learners

GATE is a data augmentation method, which means that it is agnostic to the choice of the downstream CATE learner [18, 43]. As such, it can be used both with one-step learners and two-step learners. We illustrate in Algorithm 1 how to use it in practice.

One-step learners: Examples of one-step learners include the T-learner and the S-learner. For the T-learner, one can estimate separately each μ_t using the dataset $\tilde{\mathcal{D}}_t^{(obs)}$. For the S-learner, we define the concatenation $\tilde{\mathcal{D}}^{(obs)} = (\tilde{\mathcal{D}}_0^{(obs)}, \tilde{\mathcal{D}}_1^{(obs)})$ which shall be used to estimate $\mu(x, t)$, the average PO for treatment t and covariate x.

Two-step learners: Two-step learners require the estimation of the nuisance parameters μ_0 and μ_1 in their first step, which we propose to estimate on $\tilde{\mathcal{D}}^{(\text{obs})}$. In addition, some two-step learners (e.g. DR learner) require an estimation of the propensity score $\pi(x)$. Such estimator can be obtained by considering either the original dataset $\mathcal{D}^{(obs)}$ or the augmented dataset $\tilde{\mathcal{D}}^{(\text{obs})}$ (in our empirical experiments, we used the latter option). The second step of these learners does not require any change as the nuisance estimators are used as plug-in. As such, the pseudo-outcomes should be obtained for the observational dataset $\mathcal{D}^{(obs)}$.

B.2. The question of model selection and hyperparameter tuning

The fundamental problem of causal inference makes the standard approaches to model selection and hyperparameter tuning not applicable in CATE estimation. Because the ground truth CATE value is unobserved, one cannot simply choose a model

Algorithm 1 Using GATE with CATE meta-learners.

Input: observational dataset $\mathcal{D}^{(\text{obs})} = \{(X_i, T_i, Y_i)\}_{i=1}^n$, pretrained generative model $P_{t,x}^{(\text{gen})}$, admissible sets $\mathcal{X}_0$, $\mathcal{X}_1$
Output: CATE estimation model $\hat{\tau}(x)$

for $t \in \{0, 1\}$ **do**
  $\tilde{\mathcal{D}}_t^{(obs)} \leftarrow \mathcal{D}_t^{(obs)}$
  for $i = 1$ to n **do**
    if $T_i = 1 - t$ and $X_i \in \mathcal{X}_t$ **then**
      sample $Y^{(\text{gen})} \sim P_{t,X_i}^{(gen)}$
      $\tilde{\mathcal{D}}_t^{(obs)} \leftarrow \tilde{\mathcal{D}}_t^{(obs)} \cup \{(X_i, Y^{(\text{gen})})\}$
    end if
  end for
end for
T-learner: For $t = 0, 1$, fit $\hat{\mu}_t(x)$ on $\tilde{\mathcal{D}}_t^{(obs)}$, then $\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$;
S-learner: Fit $\hat{\mu}(x, t)$ on $(\tilde{\mathcal{D}}_0^{(obs)}, \tilde{\mathcal{D}}_1^{(obs)})$, then $\hat{\tau}(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0)$;
Two-step learners: For $t = 0, 1$, fit $\hat{\mu}_t(x)$ on $\tilde{\mathcal{D}}_t^{(\text{obs})}$ and fit $\hat{\pi}(x)$ on $\tilde{\mathcal{D}}^{(\text{obs})}$; perform the second step on $\mathcal{D}^{(\text{obs})}$ or on a held-out observational dataset.

which performs best on a held-out validation set. Instead, model selection for CATE estimation has to rely on heuristics, assumptions on the data generating process and general prior knowledge of the problem at hand [20]. Model selection procedures for causal inference models remain an active area of research [20, 45, 50, 60, 61].

The challenges of model selection in causal inference also apply to GATE. Deciding which of the available generative models should be used to augment the observational dataset at hand is non-trivial, and neither is the question of choosing the admissible set $\mathcal{X}_t$ (e.g. by specifying the value of α in our proposed instantiation) for a given generative model.

Choosing the generative model. We provide the following insights which might guide the selection of the generative model within the GATE framework:

1. **LLMs vs other models.** The performance gap between the LLM and the models trained on $\mathcal{D}^{(\text{obs})}$ seems to depend on the size of the dataset (with LLMs providing particular performance improvements in smaller datasets, cf. Figure 2), as well as on the variability in the (standardized) potential outcomes (the Hillstrom dataset – where the performance gap is particularly small – has $Var(Y(1)) = 0.04$ and $Var(Y(0)) = 0.02$, while in the STAR dataset – where the performance gap is particularly large – $Var(Y(1)) = 1.05$ and $Var(Y(0)) = 0.93$). In scenarios with low outcome heterogeneity and/or large sample sizes, a simple model such as mean imputation can already perform well.

2. **Auditing the outcomes generated with the LLM**: Contrasting other generative models, the use of LLMs with GATE permits to make the generation process more transparent. Beyond producing numerical values, LLMs can also detail verbal explanations of their generations. Indeed, alternative prompting strategies can be employed to elicit explicit causal reasoning chains underpinning outcome generation. This capability enables human-in-the-loop applications of GATE, where domain experts can evaluate the generated data by examining these reasoning traces against their domain knowledge. As such, we view the use prompting techniques for explicit reasoning, such as chain-of-thought [68] or tree of thought[70], as a promising direction for future work.

Tuning the hyperparameter α. With these challenges in mind, we propose three complementary strategies which allow to guide the selection of the value of α to define the admissible set $\mathcal{X}_t$ within the GATE framework:

1. As we explain above, relying on the **fixed-threshold definition of the admissible set** (rather than the percentile-based definition of the threshold) can allow to guide the selection of α using domain knowledge.

2. We further propose to guide the selection of α by **measuring the covariate shift** between the sets $\tilde{\mathcal{D}}_0^{(obs)}$ and $\tilde{\mathcal{D}}_1^{(obs)}$, using for example the sliced Wasserstein distance (an example of such an analysis can be found in our Figure 5, right). Then, we propose to choose the minimal value of α which allows to achieve significant reduction in a covariate shift

(which might correspond to the 'elbow' in the graph). While such an 'elbow' might not always exist, this criterion provides additional guidelines in certain circumstances.

3. Finally, the choice of α can be further guided by **standard methods for CATE model selection**, particularly those based on comparing the downstream CATE models using a pseudo-outcome surrogate criteria evaluated on a held-out validation set (see [20] for an overview). In particular, in view of strong covariate shift and small sample regime, we propose to rely on criteria which do not rely on estimating the propensity score, as these might lead to high variance in such cases.

Nevertheless, while finding the *optimal* value of α is non-trivial, our experiments showed the following: (1) Fixing $\alpha = 0.5$ consistently led to improved dowstream PEHE across the 3 datasets and the 11 CATE learners (cf. Table 1) (2) The strong effect of the reduction in covariate shift obtained with data augmentation (shown in Figure 5) reduces the sensitivity with respect to α. Indeed, Figure 5 (middle) highlights that any value $\alpha > 0$ leads to performance gains compared to $\alpha = 0$.

B.3. Using variance in the generated outcomes to select the admissible set

Sources of variance. We acknowledge that the variance in the outcomes generated by the LLM, which we use as a proxy to evaluate the LLM's uncertainty and hence guide the selection of the admissible set $\mathcal{X}_t$, might capture different types of uncertainty. Firstly, it might reflect the aleatoric uncertainty, which refers to the irreducible uncertainty of the outcome distributions. Secondly, it also captures the epistemic uncertainty, which accounts for both the insufficiency of observational data in some regions of the covariate space and insufficient semantic knowledge of the LLM. Our variance-based selection mechanism relies on the implicit assumption that the aleatoric uncertainty does not vary significantly across the covariate space $\mathcal{X}$. This implies that choosing the admissible set $\mathcal{X}_t$ based on the variance allows to capture the differences in the epistemic uncertainty of the LLM across $\mathcal{X}$, where higher epistemic uncertainty may indicate to a higher inaccuracy in the generated outcomes.

Definition of the admissible sets. We define the admissible sets $\mathcal{X}_0$ and $\mathcal{X}_1$ in Section 3.2. In our instantiation, these sets are kept equal. The rationale for this choice is that only samples with relatively low uncertainty should be kept. One can imagine the situation where the generative model performs significantly worse for one of the groups (i.e. treated or control) compared to the other one. If the quantile value $\lambda(\alpha, \mathcal{D}^{(\text{obs})})$ was computed separately for the treated and control groups, then the same ratio of samples would be kept in the augmented dataset in the two groups, despite the disparities across these groups. This justifies the computation of the quantile value using all the covariates, as explicited in eq. 2.

Fixed-value threshold for the scoring function. In the instantiation of GATE that we have used in the experiments, our main focus was to control the *number* of generated potential outcomes, and as a result we have decided to use a percentile-based definition of the scoring function $s(x, t)$ (where choosing $\alpha = 0.5$ guarantees that 50% of missing potential outcomes are generated, thus allowing to fix the proportion of generated outcomes across datasets).

However, in real-world applications a more optimal strategy might be to let α be a fixed variance threshold instead, the value of which can be guided by domain-knowledge or exploratory analysis of the data. This would more explicitly guardrail against the inclusion in the augmented dataset of particularly 'poor' generated outcomes. Then, we would define $\mathcal{X}_t = \{X_i \mid i \in [n], s(X_i, T_i) < \alpha_t\}$. We note that this in case, the proportion of generated outcomes depends on the properties of the generative model. In particular, if the model is particularly bad, no potential outcomes are generated and our method recovers the baseline performance.

C. Theoretical Results

We focus on augmenting the dataset $\mathcal{D}_t^{(\text{obs})}$ for each $t \in \{0, 1\}$, used to estimate μ_t. As our goal is estimating CATE using each μ_t, we prefer augmentations that are highly informed of the covariate space $\mathcal{X}$. However, as $P_{t,x}^{(\text{gen})}$ is a pre-trained model, we cannot assume that $P_{t,x}^{(\text{gen})}$ is properly adjusted for bias. To avoid introducing too much bias, we would like to use $P_{t,x}^{(\text{gen})}$ only in a selected subset of the covariate space, which we call the *admissible set* $\mathcal{X}_t \subseteq \mathcal{X}$. We theoretically analyze how the choice of $\mathcal{X}_t$ affects the trade-off between accuracy and bias. In what follows, we fix $t \in \{0, 1\}$.

To formalise the induced changes, we introduce the variables $(X', Y'(t), Z)$, such that the joint distribution over $(X, T, Y, X', Y', Z, Y^{(gen)}(0), Y^{(gen)}(1))$ is as follows: $Z \sim Q$ (where Q is a distribution supported on $\mathcal{X}_t$) is independent of the other random variables, $X' = 1(T = t)X + 1(T = 1 - t)Z$ and $Y' = 1(T = t)Y + 1(T = 1 - t)Y^{(gen)}(t)$

(where $Y^{(gen)}(t) \sim P^{(gen)}_{t,z}$ and $1(T=t)Y = 1(T=t)Y(t)$ by consistency). Hence, $X' = X$ when $T = t$, i.e. $X' \sim P_t(X)$ when $T = t$ (factual distribution), and $X' \sim Q$ when $T = 1-t$.

Having formalized a data augmentation process with the distribution Q, we now highlight the trade-off associated with choosing a good $\mathcal{X}_t$ (and a good Q supported on this $\mathcal{X}_t$) for CAPO estimation. We do so by deriving a generalization bound on the expected risk $R(f_t) = \mathbb{E}_{X,Y(t)}\left[L(Y(t), f_t(X))\right]$ for a hypothesis f_t and a loss function L, building on the bound presented in [40] to account for the effect of using the augmented dataset $\tilde{\mathcal{D}}^{(\text{obs})}_t = \{(X'_i, Y'_i(t))\}_{i=1}^{\tilde{n}_t}$. Let $f_t \in \mathcal{H}$ denote the hypothesis used to make the predictions for $Y(t)$, where $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Y}\}$ is a hypothesis class. Let $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be a loss function (e.g. the squared loss function $L(y, y') = (y - y')^2$). We define the following quantities:

- Pointwise loss: $\ell_{f_t}(x) := \mathbb{E}_{Y(t)|X=x}\left[L(Y(t), f_t(x))\right]$,
- Marginal risk: $R(f_t) := \mathbb{E}_X\left[\ell_{f_t}(X)\right] = \mathbb{E}_{Y(t),X}\left[L(Y(t), f_t(X))\right]$,
- Marginal risk for the augmented distribution: $\tilde{R}(f_t) := \mathbb{E}_{X'}\left[\ell_{f_t}(X')\right] = \mathbb{E}_{Y'(t),X'}\left[L(Y'(t), f_t(X'))\right]$,
- Factual risk: $R_t(f_t) := \mathbb{E}_{X|T=t}\left[\ell_{f_t}(X)\right] = \mathbb{E}_{Y(t),X|T=t}\left[L(Y(t), f_t(X))\right]$,
- Counterfactual risk: $R_{1-t}(f_t) := \mathbb{E}_{X|T=1-t}\left[\ell_{f_t}(X)\right] = \mathbb{E}_{Y(t),X|T=1-t}\left[L(Y(t), f_t(X))\right]$,
- Empirical risk on the augmented distribution: $\tilde{R}^{(\text{emp})}(f_t) := \frac{1}{\tilde{n}_t}\sum_{i=1}^{\tilde{n}_t} L(Y'_i(t), f_t(X'_i))$.

In addition to these notations related to the risk, we define the class of functions $\mathcal{L} \subset \{x \to \mathbb{R}_+\}$ comprising functions $g : x \mapsto \mathbb{E}_{Y(t)|X=x}\left[L(Y(t), f_t(x))|X = x\right]$ for all $f_t \in \mathcal{H}$. Furthermore, for any $x \in \mathcal{X}$, we define a class of functions $\mathcal{L}^x \subset \{\mathcal{Y} \to \mathbb{R}_+\}$ comprising the functions $l^x_{f_t} : y \mapsto L(y, f_t(x)) \in \mathcal{L}^x$ for all $f_t \in \mathcal{H}$. Finally, for a class of functions $\mathcal{S}$ and two distributions P and P', we write $\text{IPM}_{\mathcal{S}}(P, P') = \sup_{f\in\mathcal{S}}|\mathbb{E}_{V\sim P}[f(V)] - \mathbb{E}_{W\sim P'}[f(W)]|$ for the Integral Probability Metric between P and P' defined for the class $\mathcal{S}$.

We first recall the statement of the generalization bound:

Theorem C.1. *Generalization bound Assume access to an augmented dataset* $\{(X'_i, Y'_i(t))\}_{i=1}^{\tilde{n}_t} \overset{i.i.d.}{\sim} P'(X', Y'(t))$ *and assume that* $0 < \mathbb{E}_{X',Y'\sim P'}\left[L^2(Y', f_t(X')\right] < +\infty$. *Then with probability at least* $1-\delta$,

$$R(f_t) \leq \tilde{R}^{(\text{emp})}(f_t) + (1-\pi_t)\mathbb{E}_{X\sim Q}\left[\mathit{IPM}_{\mathcal{L}^X}\left(P(Y(t) \mid X), P^{(gen)}(Y^{(gen)}(t) \mid X)\right)\right] \tag{3}$$

$$+ (1-\pi_t)\mathit{IPM}_{\mathcal{L}}(Q, P_{1-t}) + V_{P'}\frac{\mathcal{C}^{\mathcal{H}}_{\tilde{n}_t,\delta}}{\tilde{n}_t^{3/8}}, \tag{4}$$

where $V_{P'} = \max\left(\sqrt{\mathbb{E}_{X',Y'(t)\sim P'}\left[L^2(Y'(t), f_t(X'))\right]}, \sqrt{\mathbb{E}_{X',Y'\sim \hat{P}'}\left[L^2(Y'(t), f_t(X'))\right]}\right)$, *with* $\hat{P}'$ *denoting the empirical distribution for* P', *and* $\mathcal{C}^{\mathcal{H}}_{\tilde{n}_t,\delta} = 2^{5/4}\left(\frac{d\log\frac{2e\tilde{n}_t}{d} + \log\frac{8}{\delta}}{\tilde{n}_t}\right)^{\frac{3}{8}}$, *with* d *the pseudo-dimension of* $\{(x, y) \mapsto L(y, f_t(x)) \mid f_t \in \mathcal{H}\}$.

Proof. To prove Theorem C.1 for a given hypothesis f_t, our goal lies in obtaining a finite-sample generalisation bound of the marginal risk $R(f_t)$. We further note that:

$$R(f_t) = \pi_t R_t(f_t) + (1-\pi_t)R_{1-t}(f_t).$$

In this decomposition, $R_t(f_t)$ is the factual risk which is identifiable from the observational data under the ignorability assumption, and as such can be estimated using the empirical risk. However, $R_{1-t}(f_t)$ is not identifiable from the observational data. Thus, bounding the marginal risk is possible only after bounding the counterfactual risk, which requires accounting for the covariate shift and the variance in the outcomes, as we demonstrate below.

Lemma C.2. *Let* $f_t \in \mathcal{H}$. *The following inequality holds:*

$$R(f_t) - \tilde{R}(f_t) \leq (1-\pi_t)\left(\mathit{IPM}_{\mathcal{L}}(P_{1-t}, Q) + \mathbb{E}_{X\sim Q}\left[\mathit{IPM}_{\mathcal{L}^X}\left(P(Y(t)|X), P^{(gen)}(Y^{(\text{gen})}(t)|X)\right)\right]\right) \tag{5}$$

Proof. As stated earlier, $R(f_t) = \pi_t R_t(f_t) + (1-\pi_t)R_{1-t}(f_t)$. We obtain a similar decomposition for $\tilde{R}(f_t)$:

$$\begin{aligned}
\tilde{R}(f_t) &= \mathbb{E}_{X',Y'(t)}\left[L(Y'(t), f_t(X'))\right] && (6)\\
&= \pi_t \mathbb{E}_{X',Y'(t)|A=t}\left[L(Y', f_t(X')\right] + (1-\pi_t)\mathbb{E}_{X',Y'(t)|A=1-t}\left[L(Y'(t), f_t(X')\right] && (7)\\
&= \pi_t \mathbb{E}_{X,Y(t)|T=t}\left[L(Y, f_t(X)\right] + (1-\pi_t)\mathbb{E}_{X',Y'(t)|A=1-t}\left[L(Y'(t), f_t(X')\right] && (8)\\
&= \pi_t R_t(f_t) + (1-\pi_t)\tilde{R}_{1-t}(f_t) && (9)
\end{aligned}$$

where line (10) follows by definition of $(A, X', Y'(t))$. Hence, $R(f_t) - \tilde{R}(f_t) = (1-\pi_t)\left(R_{1-t}(f_t) - \tilde{R}_{1-t}(f_t)\right)$.

We can then bound $R_{1-t}(f_t) - \tilde{R}_{1-t}(f_t)$ as follows:

$$\begin{aligned}
& R_{1-t}(f_t) - \tilde{R}_{1-t}(f_t) && (10)\\
&= \mathbb{E}_{X,Y(t)|T=1-t}\left[L(Y(t), f_t(X))\right] - \mathbb{E}_{X',Y'|A=1-t}\left[L(Y'(t), f_t(X'))\right] && (11)\\
&= \mathbb{E}_{X|T=1-t}\left[\mathbb{E}_{Y(t)|X}\left[L(Y(t), f_t(X))|X\right]\right] - \mathbb{E}_{X'|A=1-t}\left[\mathbb{E}_{Y'(t)|X',A=1-t}\left[L(Y'(t), f_t(X'))|X'\right]\right] && (12)\\
&= \mathbb{E}_{X\sim P_{1-t}}\left[\mathbb{E}_{Y(t)|X}\left[L(Y(t), f_t(X))|X\right]\right] - \mathbb{E}_{X'\sim Q}\left[\mathbb{E}_{Y'(t)|X',A=1-t}\left[L(Y'(t), f_t(X') \mid X'\right]\right] && (13)\\
&= \mathbb{E}_{X\sim P_{1-t}}\left[\mathbb{E}_{Y(t)|X}\left[L(Y(t), f_t(X))|X\right]\right] - \mathbb{E}_{X\sim Q}\left[\mathbb{E}_{Y(t)|X}\left[L(Y(t), f_t(X)) \mid X\right]\right] && (14)\\
&\quad + \mathbb{E}_{X\sim Q}\left[\mathbb{E}_{Y(t)|X}\left[L(Y(t), f_t(X))|X\right]\right] - \mathbb{E}_{X\sim Q}\left[\mathbb{E}_{Y^{(\text{gen})}(t)|X}\left[L(Y^{(\text{gen})}(t), f_t(X)) \mid X\right]\right] && (15)\\
&\leq \sup_{g\in\mathcal{L}}\left|\mathbb{E}_{X\sim P_{1-t}}\left[g(X)\right] - \mathbb{E}_{X\sim Q}\left[g(X)\right]\right| && (16)\\
&\quad + \mathbb{E}_{X\sim Q}\left[\sup_{\ell^X\in\mathcal{L}^X}\left|\mathbb{E}_{Y(t)|X}\left[\ell^X(Y(t))\right] - \mathbb{E}_{Y^{(\text{gen})}(t)|X}\left[\ell^X(Y^{(\text{gen})}(t))\right]\right|\right] && (17)\\
&= \text{IPM}_{\mathcal{L}}\left(P_{1-t}, Q\right) + \mathbb{E}_{X\sim Q}\left[\text{IPM}_{\mathcal{L}^X}\left(P(Y(t)|X), P^{(gen)}(Y^{(\text{gen})}(t)|X)\right)\right] && (18)
\end{aligned}$$

where in line (13) we used the fact that $\mathbb{E}_{X'\sim Q}\left[\mathbb{E}_{Y'(t)|X',A=1-t}\left[L(Y'(t), f_t(X') \mid X'\right]\right] = \mathbb{E}_{X\sim Q}\left[\mathbb{E}_{Y^{(\text{gen})}(t)|X}\left[L(Y^{(\text{gen})}(t), f_t(X)) \mid X\right]\right]$ (by definition of $Y'(t)$).

Multiplying the sum of the IPM terms by the factor $1-\pi_t$ then yields the result. □

Having bounded the difference in the marginal risk between the original P and the augmented distribution P', we now introduce the empirical risk to bound the marginal risk for P'.

Lemma C.3. *For any $f_t \in \mathcal{H}$, assume that $0 < \mathbb{E}_{X',Y'\sim P'}\left[L^2(Y', f_t(X')\right] < +\infty$. Let $0 < \delta < 1$, and consider an augmented dataset $\{(X'_i, Y'_i(t))\}_{i=1}^{\tilde{n}_t} \overset{i.i.d.}{\sim} P'(X', Y'(t))$. The following bound then holds with probability at least $1-\delta$:*

$$\tilde{R}(f_t) \leq \tilde{R}^{(\text{emp})}(f_t) + V_{P'}\frac{\mathcal{C}^{\mathcal{H}}_{\tilde{n}_t,\delta}}{\tilde{n}_t^{3/8}} \tag{19}$$

where $V_{P'} = \max\left(\sqrt{\mathbb{E}_{X',Y'(t)\sim P'}\left[L^2(Y'(t), f_t(X'))\right]}, \sqrt{\mathbb{E}_{X',Y'\sim\hat{P}'}\left[L^2(Y'(t), f_t(X'))\right]}\right)$, with $\hat{P}'$ denoting the empirical distribution for P', and $\mathcal{C}^{\mathcal{H}}_{\tilde{n}_t,\delta} = 2^{5/4}\left(\frac{d\log\frac{2e\tilde{n}_t}{d} + \log\frac{8}{\delta}}{\tilde{n}_t}\right)^{\frac{3}{8}}$, with d the pseudo-dimension of $\{(x,y) \mapsto L(y, f_t(x)) \mid f_t \in \mathcal{H}\}$.

Proof. This result directly follows from Corollary 2 in the supplementary material of [17]. □

By summing the bounds involved in the Theorem C.2 and Theorem C.3, we then obtain Theorem C.1.

□

Interpretation. Although Theorem C.1 does not offer a guarantee on *downstream performance* (as the noise of the generator is unknown), it illustrates the different mechanisms which can affect the generalization error involved in using generative models for data augmentation.

Covariate shift and Variance: $\text{IPM}_{\mathcal{L}}(Q, P_{1-t})$ measures the distance between the distribution Q of the covariates for which the generative model is used and the counterfactual distribution P_{1-t}. We note that Theorem C.1 considers a simple setup where a fixed proportion $\pi_t = \mathbb{P}(T = t)$ of the augmented dataset consists of factual samples (in practice, this may vary due to selection). This allows to show that when the generator is perfect, it is optimal to have $Q = P_{1-t}$. The term $V_{P'} C^{\mathcal{H}}_{\tilde{n}_t,\delta}/\tilde{n}_t^{3/8}$ quantifies the variance stemming from the finite-sample regime, emphasizing that potential outcome generation is particularly impacted in the small-sample regime. Both terms can be reduced by performing data augmentation for *all* the samples with treatment $T = 1 - t$.

Noise of the generator: While the minimization of the above two terms suggests that we should augment the observational dataset with as many generated samples as possible, this ignores the impact of the generator's inaccuracy. This inaccuracy is highlighted by the term involving $\text{IPM}_{\mathcal{L}^X}$ which quantifies how close the distribution $Y^{(\text{gen})}(t)$ is to the ground-truth distribution of the potential outcome $Y(t)$, conditioned on the covariates X.

D. Experimental details

D.1. Reproducibility

All our code can be found in the following (anonymised) code-repository: https://anonymous.4open.science/r/GATE-8849. We note all the hyperparameters of models used in Appendix B and used prompts in Appendix E.

D.2. License for existing assets

The following existing assets were used to produce the experimental results:

- *Hillstrom* dataset [34]: available from https://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html
- *STAR Project* dataset [3]: CC0 1.0 License
- *Lalonde* dataset [23, 24, 44]: CC BY-NC 2.0 DEED License
- *RealCause* python library [53]: MIT License
- *CATENets* python library [18, 19, 21]: BSD 3-Clause License
- *COCOA* code [6]: Apache 2.0 license
- *TabPFN v2* model [35]: Apache 2.0 license

D.3. Dataset details

- **Lalonde** [44]: The covariates comprise several demographic variables (e.g. age, degree, marital status). The treatment corresponds to attending a job training program. The outcome is the real earnings obtained in 1978. We generate the dataset using the trained models in [53].
- **STAR project** [3]: The individuals correspond to students, and we use the following covariates: *Gender*, *Race*, *Birth year*, *G3 Surban*, *G3 Free lunch*, *G3 Present*, *Aided class*, *G3 Teacher gender*, *G3 Teacher race*, *G3 Teacher high degree*, *G3 Teach years of experience*, *G3 Teacher training*, where G3 denotes Grade 3. The treatment corresponds to putting the student in a small class. In our analysis we have only included students who were assigned to the same treatment group through all grades K-3. The outcome is the SAT score of the student.
- **Hillstrom** [34]: The covariates correspond to different customers' attributes such as the months since last purchase or the zip code of the customer. The treatment corresponds to sending an email for men's merchandise. The outcome corresponds to whether or not the customer visited the website in the following two weeks.

We provide an overview of the datasets' characteristics in Table 2.

Table 2: **Details on the datasets.**

Dataset	Type of obs. dataset	# Samples (obs.)	Covariate dim.	Label
Lalonde	Semi-synthetic	7279	8	Continuous
STAR project	Subsampled from RCT	1429	12	Continuous
Hillstrom	Subsampled from RCT	1070	8	Binary

Dataset subsampling. While the Lalonde dataset is semi-synthetic, the STAR and Hillstrom observational datasets used throughout the experiments in Section 4 are obtained by subsampling from their respective RCT data. We follow the same procedure as in [27], by defining a biasing function, with the desideratum that this biasing function should introduce a covariate shift between the treated and control groups. More precisely, given an original dataset $\{(X_i, T_i, Y_i) \mid i \in [n]\}$, we define an encoder r such that $r(x)$ is the first PCA component score for x, obtained with the set of covariates $\{X_i \mid i \in [n]\}$. Given this encoder, we then compute $\gamma = \text{Median}(\{r(X_i) \mid i \in [n]\})$. This permits to construct the datasets $\mathcal{S}_0 = \{(X_j, T_j, Y_j) \mid r(X_j) < \gamma, j \in [n]\}$ and $\mathcal{S}_1 = \{(X_j, T_j, Y_j) \mid r(X_j) \geq \gamma, j \in [n]\}$. Intuitively, these two groups have a substantial difference in terms of covariates, as is captured by the encoder r. Finally, we obtain the observational dataset using the subsampling mechanism of [27] and keep the individuals in $\mathcal{S}_0$ with treatment equal to 0, and individuals in $\mathcal{S}_1$ with treatment equal to 1, i.e. $\mathcal{D}^{(\text{obs})} = \{(X_j, T_j, Y_j) \mid (X_j, T_j, Y_j) \in \mathcal{S}_0, T_j = 0\} \bigcup \{(X_j, T_j, Y_j) \mid (X_j, T_j, Y_j) \in \mathcal{S}_1, T_j = 1\}$.

In Appendix F.2.1, we adjust the biasing intensity to modulate the covariate shift. To do so, we consider a probability $p \in [0, 1]$. We then define $\mathcal{D}^{(\text{obs})}(p) = \{(X_j, T_j, Y_j) \mid (X_j, T_j, Y_j) \in \mathcal{S}_0, B_j \sim \text{Ber}(1-p), T_j = B_j\} \bigcup \{(X_j, T_j, Y_j) \mid (X_j, T_j, Y_j) \in \mathcal{S}_1, B_j \sim \text{Ber}(p), T_j = B_j\}$.

Intuitively, higher values of p yields a more pronounced covariate shift. We consider $p \in [0.5, 0.8, 1]$ in Appendix F.2.1.

Ground-truth CATE. We fit two random forest models to half of the original and large STAR and Hillstrom datasets, which permits to estimate the two potential outcome surfaces for each of the datasets. This approach is not biased because these original datasets are RCTs. Equipped with the fitted potential surfaces, we then take their difference to define the ground-truth CATE values used for model evaluation. The other half of the datasets is then used to define an observational dataset (used to train the CATE learners) and a test set (used to evaluate the CATE learners).

D.4. Implementation details for the CATE learners

Hardware. All the experiments were performed on a machine equipped with a 64-Core AMD Ryzen Threadripper and a NVIDIA RTX A4000. Fitting one CATE learner for one given dataset took in the worst case 3 minutes, and generating the augmented datasets with the LLM took a maximum of 17 minutes and 43 seconds per dataset and seed.

We now detail the hyperparameters used for the different CATE learners used in Section 4, which use neural networks backbones.

- **TNet**: Following [19], each hypothesis function has 3 layers with 200 units. The output head consists of 2 additional layers with 100.
- **SNet**: We use 3 layers with 100 hidden units for the shared layers, 2 layers with 100 units for the output head of the hypothesis functions, and 2 layers with 100 units for the output of the propensity network.
- **XNet**: *First stage:* We use the T strategy to estimate the nuisance parameters. We use 3 layers with 100 units for the representation, 2 layers with 100 units for the output head, *Second stage:* We use 2 layers with 100 units for the output, 3 layers with 200 units for the representation.
- **DRNet**: *First stage:* We use the T strategy to estimate the nuisance parameters. We use 3 layers with 200 units for the representation, 2 layers with 100 units for the output head, *Second stage:* We use 2 layers with 100 units for the output, 3 layers with 200 units for the representation.
- **CFR-Wass**: We use 3 layers with 200 units for the representation layers and 3 layers with 100 units per hypothesis function. We use the Wasserstein-1 distance for the regularization, with the regularization coefficient α set to 3.
- **CFR-MMD**: We use 3 layers with 200 units for the representation layers and 3 layers with 100 units per hypothesis function. We use the MMD for the regularization, with the regularization coefficient α set to 3.

- **RNet**: We use 3 layers with 200 units for the representation, 2 layers with 100 units for the output head, for the two stages.
- **IPW**: *First stage:* We use the T strategy to estimate the nuisance parameters. We use 3 layers with 200 units for the representation, 2 layers with 100 units for the output head, *Second stage:* We use 2 layers with 100 units for the output, 3 layers with 200 units for the representation.

The batch size is set to 500, the learning rate is set to 0.0001 with the Adam optimizer and we use early stopping with a validation split proportion equal to 0.3.

Hyperparameters for the instanciation of GATE with LLMs. We use `GPT-3.5` as the LLM throughout our experiments, which we access using the API, version `2023-07-01-preview`. We use a temperature of 0.9 throughout our experiments. We define the admissible set $\mathcal{X}_t$ with the variance-based criterion (Section 3.2), using a fixed threshold $\alpha = 0.5$ unless otherwise stated. We use 100 in-context samples in each prompt, and set $K = 10$ unless otherwise stated.

D.5. Implementation details for the generative models

The following models perform augmentation by training $P_{0,x}^{(gen)}$ on $\mathcal{D}_0^{(obs)}$ and $P_{1,x}^{(gen)}$ on $\mathcal{D}_1^{(obs)}$ respectively. In particular:

- Mean imputation: $P_{t,x}^{(gen)} = \delta(\frac{1}{n_t}\sum_{i=1}^{n} Y_i \mathbb{1}(T_i = t))$, where $\{X_i, T_i, Y_i\} \in D_t^{(obs)}$
- Random Forest: $P_{t,x}^{(gen)} = \delta(f_t^{(RF)}(x))$, where $f_t^{(RF)}$ is a random forest model trained on $\mathcal{D}_t^{(obs)}$
- Nearest-neighbor: $P_{t,x}^{(gen)} = \delta(f_t^{(NN)}(x))$ where $f_t^{(NN)}$ is a nearest-neighbor predictor trained on $\mathcal{D}_t^{(obs)}$.

The GAN augmentation uses a single model trained on $\mathcal{D}^{(obs)}$. We refer to [71] for the details of the method. We use the following parameters: { hidden dimension: 100, batch size: 256, iteration: 10000, α : 1, learning rate: 0.001 }.

We use the TabPFN v2 model [35] released publicly by the authors, and use the default parameters.

D.6. Metrics

Assessing covariate shift with the sliced Wasserstein distances. In the experiments in Appendix F.2.1 and Appendix F.2.2, we quantify the covariate shift between the treated and control group using the sliced Wasserstein distance. It is a metric which can compare two high-dimensional distributions [14]. To compute it, we perform random projections on vectors of the unit sphere. For two distributions μ_1 and μ_2, the sliced Wasserstein distance of order p is defined as:

$$SW_p(\mu_1, \mu_2) \coloneqq \int_{\mathbb{S}^{d-1}} W_p(P_u \# \mu_1, P_u \# \mu_2) du \tag{20}$$

where $\mathbb{S}^{d-1}$ denotes the unit sphere in dimension d, $P_u(x) = u \cdot x$ denotes the projection of the vector x on u, $P_u \# \mu$ is the push-forward of μ by P_u, and W_p is the Wasserstein distance of order p. In our experiments, we use a Monte-Carlo estimate by randomly sampling $n = 5000$ random vectors $\{u_i | i \in [n]\}$ in $\mathbb{S}^{d-1}$ and consider $p = 2$.

Assessing the inaccuracy of the generated potential outcomes. Let us consider an augmented dataset $\{(O_i, X_i', T_i', Y_i')\}_{i=1}^{\tilde{n}}$, where T_i' denotes the observed (factual) treatment, $O_i = 0$ if Y_i' is the observed potential outcome and $O_i = 1$ if Y_i' was generated with $P_{t,x}^{(\text{gen})}$. We assess in Appendix F.2.2 the inaccuracy of the generated potential outcomes in the augmented dataset by computing:

$$\Delta = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} (Y_i' - \mathbb{E}\left[Y_i(1 - T_i') \mid X_i'\right])^2 \mathbb{1}(O_i = 1)$$

PEHE. Our results in Section 4 evaluate the performance of the models on $\mathcal{D}_{\text{test}}$ using the Precision in Estimation of Heterogeneous Effect (PEHE), defined as $\epsilon_{\text{PEHE}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}\left[Y_i(1) - Y_i(0) | X = X_i\right] - (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)))^2$. We report its square root $\sqrt{\epsilon_{\text{PEHE}}}$ [33].

E. LLM Prompts

Prompt design. When instantiating GATE with LLMs, we consider a prompt structure which includes the following important elements:

- **Task context**: We include context about the task (CATE estimation). We also provide information about the covariates, the treatment, and the outcomes.
- **Statistics on the outcomes**: we provide the average outcomes in both the control and treatment group, as well as the range of the outcomes to help the LLM generate realistic outcomes.
- **In-context samples**: we serialize the observational data in their raw format. The covariates are provided as (feature name, feature value) tuples, followed by (treatment name, treatment value), and (outcome name, outcome value). The in-context samples are randomly shuffled in the prompt to avoid any generation artifacts stemming from the ordering of the samples. We use 100 in-context samples per prompt.

The prompt structure is summarized in Figure 4.

> You are an expert in causal inference. Your goal is to produce counterfactuals from observational data. I will give you the covariates, the treatment and the outcome from the observational data. Leverage your knowledge about {Task context: general}. The covariates consist of {Task context: covariates description} The treatment indicator (binary) corresponds to { Task context: treatment description}. The outcome is { Task context: outcomes }. To help you, I am providing some statistics about the data. {Statistics treatment group} {Statistics control group} Your response should only contain the generated counterfactuals in the format ## outcome ##. {In-context examples}

Figure 4: Prompt structure.

Prompt example. We provide an example of the prompt used for the Lalonde dataset in Listing 1.

Listing 1: **Prompt example.** On `Lalonde` dataset.

```
You are an expert in causal inference. Your goal is to produce
   counterfactuals from observational data. I will give you the
   covariates, the treatment and the outcome from the
   observational data. Leverage your knowledge about job
   training and real earnings to produce counterfactuals. The
   covariates consist of a number of demographic variables: age,
    measured in years; education, measured in years; black,
   indicating race (1 if black, 0 otherwise);hispanic,
   indicating race (1 if Hispanic, 0 otherwise);married,
   indicating marital status (1 if married, 0 otherwise);
   nodegree, indicating high school diploma (1 if no degree, 0
   otherwise); re74, real earnings in 1974; re75, real earnings
   in 1975. The treatment indicator (binary) corresponds to job
   training. The outcome is real earnings in the year 1978,
   denoted as re78. To help you, I am providing some statistics
   about the data. In the presence of the treatment (treat: 1),
   the average re78 (outcome) in the observational data is
   4576.24, the min re78 is 0.0, the max re78 is 26354.16. In
   the absence of the treatment (treat: 0), the average re78 (
   outcome) in the observational data is 14868.48, the min re78
   is 0.0, the max re78 is 28609.63. Your response should only
   contain the generated counterfactuals in the format ##
   outcome ##
```

```
Covariates: age: 26.0, education: 11.0, black: 0.0, hispanic:
    0.0, married: 1.0, nodegree: 1.0, re74: 25862.32, re75:
    16650.0
treat: 0
re78: ## 24058.61 ##
Covariates: age: 23.0, education: 7.0, black: 1.0, hispanic: 0.0,
     married: 1.0, nodegree: 1.0, re74: 18350.49, re75: 14967.1
treat: 0
re78: ## 8564.2 ##

...
Covariates: age: 30.0, education: 16.0, black: 0.0, hispanic:
    0.0, married: 1.0, nodegree: 0.0, re74: 695.54, re75: 930.97
treat: 1
re78:
```

No context prompt We provide in Listing 2 the prompt used throughout Section 4.3, where the contextual information is removed.

Listing 2: **Prompt example without contextual information.** On `Lalonde` dataset.

```
You are an expert in causal inference. Your goal is to produce
    counterfactuals from observational data. I will give you the
    covariates, the treatment and the outcome from the
    observational data. To help you, I am providing some
    statistics about the data. In the presence of the treatment (
    treat: 1), the average re78 (outcome) in the observational
    data is 4576.24, the min re78 is 0.0, the max re78 is
    26354.16. In the absence of the treatment (treat: 0), the
    average re78 (outcome) in the observational data is 14868.48,
     the min re78 is 0.0, the max re78 is 28609.63. Your response
     should only contain the generated counterfactuals in the
    format ## outcome ##
Covariates: Feature_0: 26.0, Feature_1: 11.0, Feature_2: 0.0,
    Feature_3: 0.0, Feature_4: 1.0, Feature_5: 1.0, Feature_6:
    25862.32, Feature_7: 16650.0
treat: 0
outcome: ## 24058.61 ##
Covariates: Feature_0: 23.0, Feature_1: 7.0, Feature_2: 1.0,
    Feature_3: 0.0, Feature_4: 1.0, Feature_5: 1.0, Feature_6:
    18350.49, Feature_7: 14967.1
treat: 0
outcome: ## 8564.2 ##

...
Covariates: Feature_0: 30.0, Feature_1: 16.0, Feature_2: 0.0,
    Feature_3: 0.0, Feature_4: 1.0, Feature_5: 0.0, Feature_6:
    695.54, Feature_7: 930.97
treat: 1
outcome:
```

Dataset splitting. Since the LLM context window limits the number of tokens which can be used in the prompt, we cannot feed all the available observational data into a single prompt. To bypass this issue, we randomly partition the observational dataset into different groups of in-context samples, each of these groups making one prompt. Each group is populated by $n_{\text{ICL}} = 100$ samples. Having split the observational data into different groups, we construct the prompts as follows. For

each individual, we identify the group it belongs to, and construct a prompt where the individual appears at the end of the prompt, with the rest of the group passed as in-context examples above it in a random order to avoid any ordering bias. The LLM then generates $m = 10$ outcomes for each individual and its associated constructed prompt.

Memorization risks. A natural question is whether or not the LLM is returning outcomes which have been memorized and seen during its pretraining stage. We note that this is very unlikely to be the case, since by definition, the LLM is used to output missing potential outcomes, which are not present in the observational datasets and hence not part of the pretraining corpora of the LLM. We also remark that the Lalonde dataset is semi-synthetic, meaning that it is also very unlikely that it has been memorized by the LLM.

Effect of the LLM temperature. The LLM temperature controls the randomness in the generated outcomes. Sampling with low temperature does not faithfully capture the outcome distribution, as it limits the variability in the generated outcomes and as such makes it difficult to faithfully estimate the LLM uncertainty to guide the selection of the admissible set. In contrast, choosing high temperature increases the diversity of the predictions, potentially leading to the generation of outlier values which can decrease the quality of predictions. In our experiments we consistently set the temperature to 0.9, as initial tests showed that this leads to optimal performance.

F. Additional results

F.1. Results for other CATE learners

In the interest of space, we put in Table 3 the results of the experiment conducted in Section 4.1 with additional CATE learners: CFR-Wass, and CFR-MMD [63], TARNet [63], DragonNet [64] and BART [9].

Table 3: GATE improves the performance of different CATE learners across the datasets without data augmentation (✗), and with data augmentation (✓). Average $\sqrt{\epsilon_{\text{PEHE}}}$ and 1std is reported for 3 seeds (↓ is better)

Learner	Lalonde CPS1D		STAR		Hillstrom	
	✗	✓	✗	✓	✗	✓
CFR-Wass.	0.99 ± 0.03	0.95 ± 0.02	0.61 ± 0.15	0.41 ± 0.01	0.24 ± 0.0	0.24 ± 0.0
CFR-MMD	1.00 ± 0.03	0.95 ± 0.00	0.64 ± 0.16	0.44 ± 0.00	0.24 ± 0.00	0.24 ± 0.00
TARNet	1.20 ± 0.03	0.96 ± 0.01	0.49 ± 0.1	0.48 ± 0.04	0.39 ± 0.02	0.24 ± 0.00
DragonNet	0.97 ± 0.02	0.95 ± 0.02	0.90 ± 0.26	0.48 ± 0.04	0.41 ± 0.04	0.24 ± 0.01
BART	1.36 ± 0.03	1.35 ± 0.00	0.70 ± 0.09	0.56 ± 0.02	0.27 ± 0.02	0.25 ± 0.01

F.2. Does GATE conform to the theoretical intuition?

Having shown that GATE can consistently improve CATE estimation, we now further verify whether the empirical results agree with the theoretical intuition provided in Appendix C. In particular, we investigate whether GATE addresses the covariate shift problems, and whether the gains from reducing covariate shift counterbalances the potential bias introduced by a generative model.

F.2.1. High covariate shift settings

Goal. We investigate the correlation between GATE's performance gains and the intensity of the *covariate shift* between the treated and control groups in $\mathcal{D}^{(\text{obs})}$.

Setup. We control the covariate shift's strength with the biasing intensity in the subsampling mechanism proposed by [27]. This manipulation yields three distinct datasets, derived from the original STAR dataset. We quantify the strength of the covariate shift with the sliced Wasserstein distance (SW) between the covariates of individuals in $\mathcal{D}_0^{(\text{obs})}$ and $\mathcal{D}_1^{(\text{obs})}$. For each dataset, we calculate the relative gain in $\sqrt{\epsilon_{\text{PEHE}}}$ obtained with CATE learners trained on $\tilde{\mathcal{D}}^{(\text{obs})}$ compared to $\mathcal{D}^{(\text{obs})}$.

Results. In Figure 5 (left), *the performance gain obtained by GATE increases with the SW* across models which employ specialised regularisation techniques (i.e. CFR-Wass and DR-learner). This shows that GATE can be particularly helpful in strong covariate shift settings. Contrary to other meta-learners, we find that the S-learner is least affected by data augmentation. We believe this to be due to its data efficiency (using all data for each PO estimate).

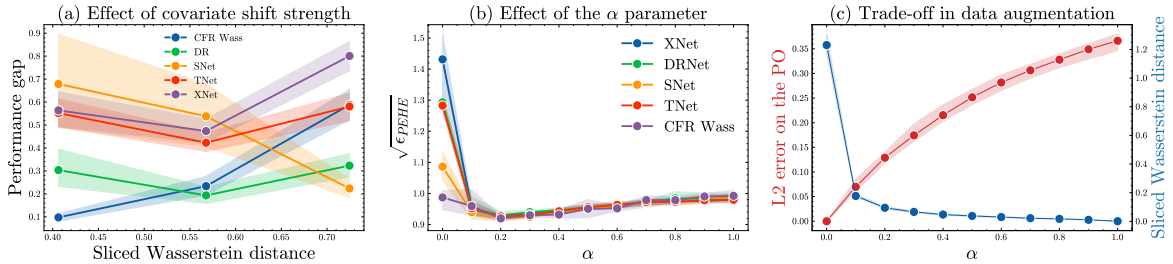

Figure 5: **Left:** The performance gains offered by GATE increase across the majority of learners as the strength of the covariate shift increases (the shaded regions denote 95% confidence intervals computed over 30 seeds) **Middle:** The value of the hyperparameter α allows to navigate the trade-off involved in data augmentation. **Right:** The bias introduced by $P_{t,x}^{(\text{gen})}$ is counterbalanced by the reduction in covariate shift obtained when using GATE.

F.2.2. Counterbalancing Bias through Covariate Shift Reduction

Goal. We further verify whether, as indicated by Theorem C.1, the benefits obtained from the reduction in covariate shift can counterbalance the bias potentially introduced by $P_{t,x}^{(\text{gen})}$, thus offering performance benefits to the downstream CATE model. We also check whether the hyperparameter α allows to navigate the trade-off between the covariate shift reduction and the bias induced by $P_{t,x}^{(\text{gen})}$.

Setup. We vary the quantile value α used by the selector (eq. 2) across the range $(0, 1)$. For each α, we compute the performance when using GATE (with LLM) for the different CATE models (Figure 5, middle). Furthermore, we explicitly quantify the covariate shift in $\tilde{\mathcal{D}}^{(\text{obs})}$ using SW, and the bias introduced by $P_{t,x}^{(\text{gen})}$ by computing the average error in the potential outcomes in $\tilde{\mathcal{D}}^{(\text{obs})}$ compared to the ground-truth values (Appendix D.6 for more details). We show how these quantities vary in α (Figure 5, right). We report averages and 95% confidence intervals for 3 seeds.

Results. Both the middle and right plots in Figure 5 verify our intuition that there exists an optimal choice of α ($\alpha \simeq 0.2$) for the Lalonde dataset which allows to balance the gains obtained by addressing the covariate shift with the losses suffered by introducing bias with $P_{t,x}^{(\text{gen})}$. Figure 5 (middle) also demonstrates that *GATE offers performance gains for most choices of* $\alpha > 0$ when compared to the no-augmentation baseline ($\alpha = 0$). Further, as we increase the allowed level of uncertainty of $P_{t,x}^{(gen)}$, the average L^2 error in potential outcomes over $\tilde{\mathcal{D}}^{(obs)}$ (quantifying how much noise is introduced by data augmentation) increases, while the covariate shift decreases.

F.3. Local regression results

We compare GATE with COCOA [6]. As discussed in Appendix A, COCOA employs a local regression model which is trained on the observational data only. This limitation can make COCOA particularly susceptible to covariate shift scenarios or when operating in a small-sample regime, where the available data may not sufficiently capture the underlying distribution of outcomes.

We report the results in Table 4, comparing the LLM-instantiated GATE with COCOA, which shows that the LLM-instantiated GATE consistently outperforms COCOA across almost all of the datasets and meta-learners. The performance gap is particularly noticeable for the Lalonde dataset, where the control and treated groups are imbalanced, making the local regression model in COCOA significantly less useful than the prior-knowledge-empowered LLMs.

F.4. Comparison on the IHDP dataset

We evaluate the benefits of GATE instantiated with an LLM for the IHDP dataset [63]. We note that the outcomes for this dataset are synthetic. Therefore, the objective of this experiment is to assess the in-context learning abilities of the LLM, and the importance of covariate shift reduction via data augmentation. We report the results in Table 5, showing that GATE improves the performance of almost all the CATE learners.

Table 4: **Comparison with COCOA [6].** Performance comparison across the datasets for COCOA and GATE . Average $\sqrt{\epsilon_{\text{PEHE}}}$ and 1std is reported for 3 seeds (↓ is better).

Learner	Lalonde CPS1D		STAR		Hillstrom	
	COCOA	GATE	COCOA	GATE	COCOA	GATE
R-learner	1.66 ± 0.42	0.95 ± 0.00	0.58 ± 0.03	0.47 ± 0.01	0.30 ± 0.02	0.26 ± 0.02
IPW-learner	1.12 ± 0.05	0.95 ± 0.01	0.59 ± 0.07	0.47 ± 0.01	0.34 ± 0.11	0.25 ± 0.00
TARNet	1.26 ± 0.08	0.96 ± 0.01	0.45 ± 0.06	0.48 ± 0.04	0.27 ± 0.01	0.24 ± 0.00
DragonNet	1.04 ± 0.06	0.95 ± 0.02	0.51 ± 0.03	0.48 ± 0.04	0.27 ± 0.01	0.24 ± 0.01
CFR-MMD	1.01 ± 0.01	0.95 ± 0.00	0.58 ± 0.15	0.44 ± 0.00	0.24 ± 0.00	0.24 ± 0.00
BART	1.32 ± 0.01	1.35 ± 0.00	0.62 ± 0.07	0.56 ± 0.02	0.26 ± 0.01	0.25 ± 0.01
T-learner	1.35 ± 0.06	0.96 ± 0.01	0.66 ± 0.08	0.50 ± 0.03	0.28 ± 0.03	0.24 ± 0.01
S-learner	1.04 ± 0.14	0.95 ± 0.01	0.88 ± 0.13	0.56 ± 0.02	0.28 ± 0.02	0.25 ± 0.01
X-learner	1.38 ± 0.15	0.95 ± 0.01	0.73 ± 0.04	0.49 ± 0.02	0.27 ± 0.01	0.24 ± 0.01
DR-learner	1.35 ± 0.05	0.95 ± 0.01	0.62 ± 0.2	0.48 ± 0.02	0.31 ± 0.02	0.25 ± 0.01
CFR-Wass.	0.98 ± 0.04	0.95 ± 0.02	0.55 ± 0.15	0.41 ± 0.01	0.24 ± 0.0	0.24 ± 0.0

Table 5: **Comparison on IHDP.** Performance comparison for the IHDP dataset, between No augmentation and GATE . Average $\sqrt{\epsilon_{\text{PEHE}}}$ and 1std is reported for 3 seeds (↓ is better).

Learner	IHDP	
	No aug.	GATE
S-learner	0.71 ± 0.10	0.54 ± 0.03
T-learner	0.70 ± 0.13	0.40 ± 0.06
X-learner	0.68 ± 0.10	0.33 ± 0.04
R-learner	0.68 ± 0.04	0.37 ± 0.01
IPW-learner	0.85 ± 0.04	0.38 ± 0.04
DR-learner	0.61 ± 0.06	0.37 ± 0.04
TARNet	0.47 ± 0.03	0.31 ± 0.04
DragonNet	0.41 ± 0.02	0.31 ± 0.04
CFR-MMD	0.29 ± 0.01	0.27 ± 0.01
CFR-Wass.	0.28 ± 0.01	0.29 ± 0.05
BART	0.56 ± 0.00	0.59 ± 0.01

F.5. Sensitivity with respect to α

We complement the results shown in Appendix F.2.2, with Figure 6 and Figure 7, which present the impact of varying the quantile α used to define the admissible set $\mathcal{X}_t$. We note that the results of the trade-off experiment presented here (Figure 7) and in the main text (Figure 5) were obtained using the DR-learner. For both the STAR Project and Hillstrom datasets, we see that incorporating the generated outcomes helps improve the PEHE. However, unlike for the Lalonde dataset, there is no clear cutoff value for α after which the PEHE starts increasing. This observation can be made more intuitive by examining Figure 7. Indeed, we notice that the covariate shift reduction obtained by increasing α is less pronounced than in the case of the Lalonde dataset, while the noise introduced with the generated outcomes increases at a similar rate. This explains why setting higher values of α is not harmful: the effect of the reduction in covariate shift balances the increased inaccuracy in the generated potential outcomes.

F.6. Comparison with the baselines under no selection

While the results in Figure 2 used the same LLM-based admissible set for all the augmentation methods for a fair comparison, in Figure 8 we provide additional results under no selection for all the methods – i.e. we choose the admissible set $\mathcal{X}_1 = \mathcal{X}_0 = \mathcal{X}$. The results show that LLMs outperform baseline methods both with and without selection. Further, selection improves performance in the Lalonde dataset, which aligns with the results presented in Figure 6. We note that the results of the comparison experiment presented here (Figure 8) and in the main text (Figure 2) were obtained using the DR-learner. We notice that the performance gap on the Hillstrom dataset is negligible. This aligns with our observation

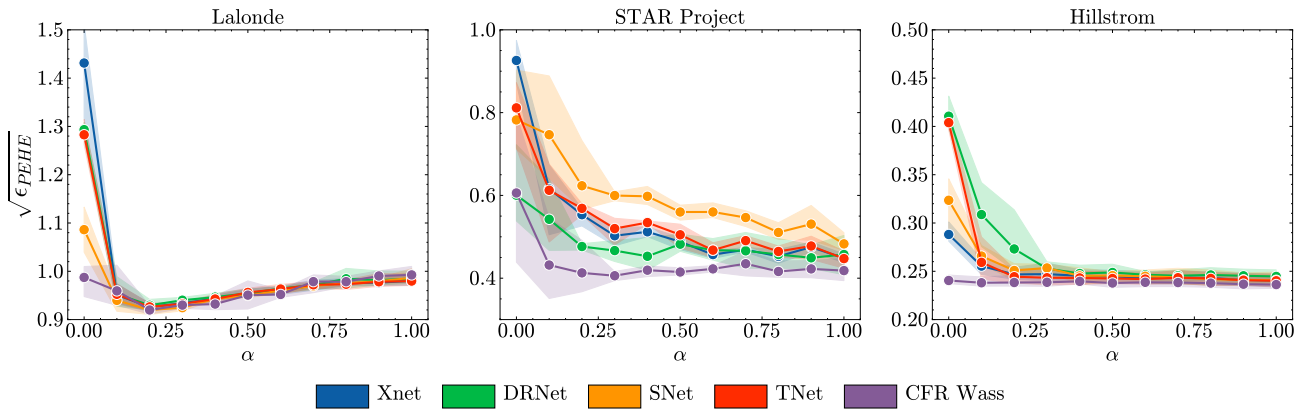

Figure 6: Sensitivity with respect to α

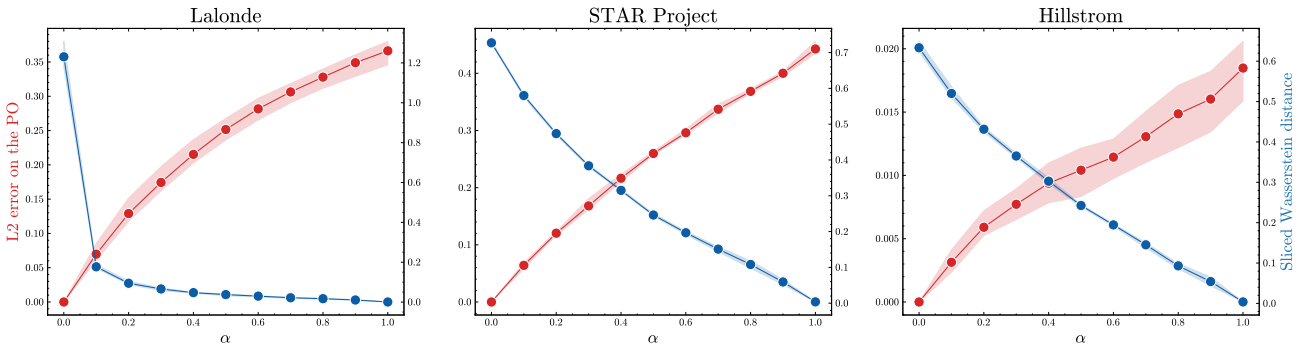

Figure 7: Tradeoff between covariance shift and potential outcome generation inaccuracy

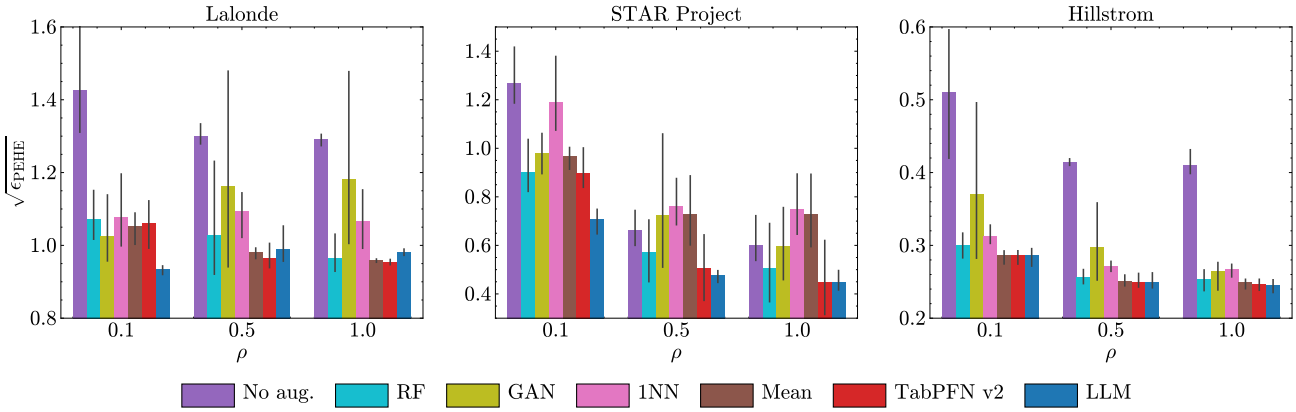

Figure 8: Comparison of the LLM with the baselines (no selection)

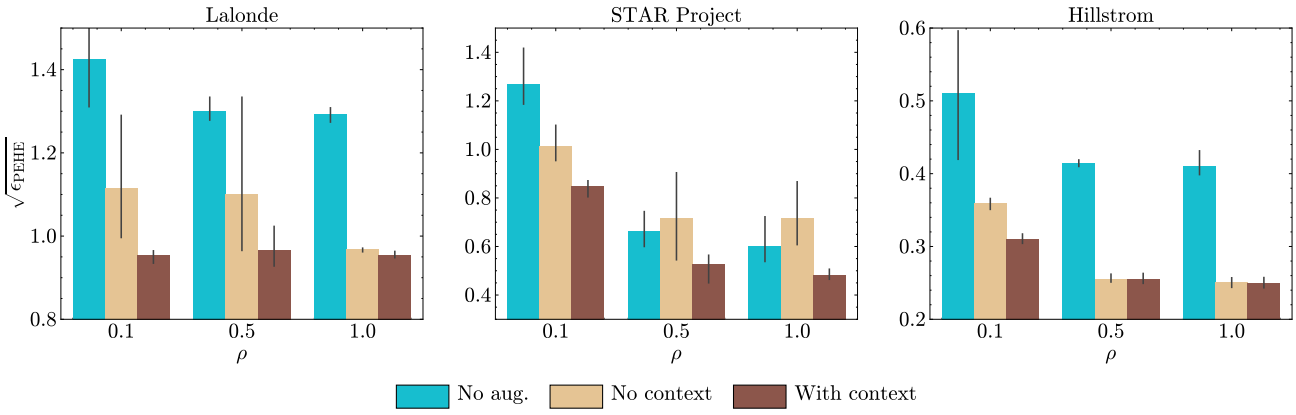

Figure 9: Comparison of using DR-learner fitted on the data augmented with GATE, using the LLM prompted with and without context. The error bars mark 1std computed over 3 seeds.

that the treatment effect is very small for this dataset. Indeed, the Average Treatment Effect, defined as $\mathbb{E}\left[Y(1) - Y(0)\right]$ is equal to 0.08. Furthermore, the variability in the outcome is negligible, with $\text{Var}(Y(1)) = 0.04$ and $\text{Var}(Y(0)) = 0.02$, explaining why the mean imputation baseline performs competitively with respect to the LLM. In contrast, the ATE for the STAR dataset is equal to 0.15, and $\text{Var}(Y(1)) = 1.05$ and $\text{Var}(Y(0)) = 0.93$ (computed on the normalized outcomes), where the larger variability explains the performance gap between the LLM and the mean baseline.

F.7. Importance of contextual information

Following the same experimental setup as in Section 4.3, we assess the importance of the contextual information to improve the potential outcome generation for the Lalonde and Hillstrom datasets. We report the results in Figure 9. We note that the results of the context experiment presented here (**??**) and in the main text (Figure 9) were obtained using the DR-learner. For the Lalonde dataset, we notice that the gains obtained using contextual information are especially noticeable in the small-sample regime (i.e. $\rho = 0.1$), echoing the observations made for the STAR Project dataset. The performance gap narrows down with an increasing ρ, as the increased sample size in factual data makes the CATE learner more robust with respect to the inaccuracy of the generated potential outcomes. On the other hand, the performance gap on the Hillstrom dataset is negligible. This aligns with our observation that the treatment effect is very small for this dataset.

F.8. Statistical Tests of Improvements

Experiment setting. In order to assess the statistical significance of the results in Table 1, we conduct two-sample t-tests on the $\sqrt{\epsilon_{PEHE}}$ obtained with and without `GATE` (instantiated with LLMs).

Results. We report the p-values in Table 6, showing that the performance gains obtained with GATE are statistically significant at the 0.05 level across the majority of CATE learners and datasets.

Table 6: **Statistical significance of GATE's performance gains.** We report the p-values of the two-sample t-tests, where bolded entries represent statistical significance at the 0.05 level.

Learner	Lalonde CPS1D	STAR	Hillstrom
S-learner	$\mathbf{3.0 \times 10^{-2}}$	$\mathbf{2.2 \times 10^{-2}}$	$\mathbf{1.3 \times 10^{-2}}$
T-learner	$\mathbf{4.0 \times 10^{-5}}$	$\mathbf{4.0 \times 10^{-3}}$	$\mathbf{1.3 \times 10^{-5}}$
X-learner	$\mathbf{1.3 \times 10^{-3}}$	$\mathbf{1.5 \times 10^{-4}}$	$\mathbf{4.1 \times 10^{-3}}$
R-learner	$\mathbf{1.1 \times 10^{-2}}$	$\mathbf{1.5 \times 10^{-7}}$	$\mathbf{4.9 \times 10^{-4}}$
IPW-learner	$\mathbf{1.0 \times 10^{-4}}$	$\mathbf{1.9 \times 10^{-5}}$	$\mathbf{7.4 \times 10^{-9}}$
DR-learner	$\mathbf{6.8 \times 10^{-6}}$	1.3×10^{-1}	$\mathbf{1.4 \times 10^{-4}}$
CFR-Wass.	3.7×10^{-1}	$\mathbf{3.4 \times 10^{-6}}$	3.8×10^{-1}
CFR-MMD.	$\mathbf{4.9 \times 10^{-6}}$	1.6×10^{-1}	6.0×10^{-1}
TARNet	$\mathbf{1.1 \times 10^{-5}}$	2.8×10^{-1}	$\mathbf{3.2 \times 10^{-16}}$
DragonNet	1.7×10^{-1}	$\mathbf{7.3 \times 10^{-6}}$	$\mathbf{1.8 \times 10^{-7}}$
BART	$\mathbf{1.3 \times 10^{-6}}$	6.6×10^{-2}	3.0×10^{-1}

F.9. Convergence of CATE models after augmentation

As shown in Table 1, data augmentation reduces the performance gaps between the CATE learners. A key reason is that it mitigates the covariate shift problem (cf. Figure 7), reducing the importance of regularisation strategies (e.g. balancing representations in CFR, importance weighting in IPW).

An explanation of why CATE errors do not reach zero after augmentation is that there exists irreducible noise in the prediction task – uncontrollable in our setting due to the use of real-world (rather than fully synthetic) outcomes. To confirm this, we evaluate performance using imputed ground truth counterfactuals in Table 7. Even with oracle augmentation and no covariate shift, errors remain nonzero.

Table 7: **Comparison with an oracle augmentation.** Average $\sqrt{\epsilon_{\text{PEHE}}}$ and 1std is reported for 3 seeds

Learner	Lalonde CPS1D		STAR		Hillstrom	
	LLM	Oracle	LLM	Oracle	LLM	Oracle
S-learner	0.95 ± 0.01	0.91 ± 0.01	0.56 ± 0.02	0.29 ± 0.04	0.25 ± 0.01	0.24 ± 0.01
T-learner	0.96 ± 0.01	0.91 ± 0.01	0.50 ± 0.03	0.22 ± 0.02	0.24 ± 0.01	0.24 ± 0.01
X-learner	0.95 ± 0.01	0.91 ± 0.01	0.49 ± 0.02	0.21 ± 0.02	0.24 ± 0.01	0.24 ± 0.01
R-learner.	0.95 ± 0.00	0.91 ± 0.01	0.47 ± 0.01	0.20 ± 0.01	0.26 ± 0.02	0.25 ± 0.01
IPW-learner.	0.95 ± 0.01	0.91 ± 0.01	0.47 ± 0.01	0.22 ± 0.00	0.25 ± 0.00	0.24 ± 0.01
DR-learner	0.95 ± 0.01	0.92 ± 0.01	0.48 ± 0.02	0.20 ± 0.00	0.25 ± 0.01	0.24 ± 0.01
CFR-Wass.	0.95 ± 0.02	0.91 ± 0.01	0.41 ± 0.01	0.23 ± 0.01	0.24 ± 0.00	0.24 ± 0.01
CFR-MMD	0.95 ± 0.00	0.91 ± 0.01	0.44 ± 0.00	0.18 ± 0.00	0.24 ± 0.00	0.24 ± 0.01
TARNet	0.96 ± 0.01	0.92 ± 0.01	0.48 ± 0.04	0.21 ± 0.01	0.24 ± 0.00	0.24 ± 0.01
DragonNet	0.95 ± 0.02	0.92 ± 0.01	0.48 ± 0.04	0.21 ± 0.01	0.24 ± 0.01	0.24 ± 0.01
BART	1.35 ± 0.00	1.32 ± 0.00	0.56 ± 0.02	0.51 ± 0.05	0.25 ± 0.01	0.25 ± 0.01

F.10. Alternative Selection of In-context Samples

Experimental setting. We consider an instantiation GATE with LLM where the in-context samples used in the prompts are k nearest-neighbours of the samples considered for augmentation. More specifically, given a sample (x, t), we define $\mathcal{S}_{x,t} = \text{NN}_k(X, \mathcal{D}^{(obs)}_{1-t})$ as the set of in-context samples for (x, t). We set $k = 50$ and use a DR-learner for downstream CATE estimation.

Results. As presented in Table 8, our results demonstrate that random sampling of in-context samples from $\mathcal{D}^{(obs)}$

(encompassing both control and treated groups) consistently yields superior performance compared to the nearest neighbor baseline. This is intuitive given the covariate shift between the two groups,which inherently limits the utility of nearest-neighbor information drawn from the opposing treatment group. Random sampling, by contrast, enables the incorporation of individuals from both groups – a particularly advantageous approach when prior knowledge exists regarding the relationship between Y^1 and Y^0 (e.g. difference in expectation).

Table 8: **Comparison of in-context samples' selection.** Results reported for 3 seeds.

IC sampling	Lalonde CPS1D	STAR	Hillstrom
	$\rho = 0.1$		
Nearest neighbor	1.09 ± 0.12	0.99 ± 0.08	0.39 ± 0.06
Random sampling	0.95 ± 0.02	0.85 ± 0.04	0.31 ± 0.01
	$\rho = 0.5$		
Nearest neighbor	1.10 ± 0.04	0.62 ± 0.09	0.28 ± 0.02
Random sampling	0.97 ± 0.05	0.53 ± 0.07	0.26 ± 0.01
	$\rho = 1$		
Nearest neighbor	1.09 ± 0.10	0.58 ± 0.02	0.26 ± 0.01
Random sampling	0.95 ± 0.01	0.48 ± 0.02	0.25 ± 0.01

F.11. Comparison against Other Selectors

Experimental setting. We compare the variance-based selector used in our LLM instantiation of `GATE` with two additional selectors: (1) a selector which selects the samples uniformly at random in the observational dataset (*Random*) and (2) a propensity-based selector (*Propensity*), which defines the score function as $s(x,t) = P(T = t|X = x)$, intuitively favouring samples exhibiting characteristics similar to those from the opposite treatment group. For all the selectors, we set $\alpha = 0.5$, and use a DR-Learner for downstream CATE estimation.

Results. We report the results in Table 9, showing that the variance-based selector achieves optimal performance most consistently out of the considered selection criteria, with performance gains especially noticeable in the small-sample regime ($\rho = 0.1$).

Table 9: **Comparison against other selectors.** Results reported for 3 seeds.

Selector	Lalonde CPS1D	STAR	Hillstrom
	$\rho = 0.1$		
Random	0.95 ± 0.02	0.90 ± 0.21	0.35 ± 0.03
Propensity	0.95 ± 0.01	0.87 ± 0.14	0.39 ± 0.02
Variance	0.95 ± 0.02	0.85 ± 0.04	0.31 ± 0.01
	$\rho = 0.5$		
Random	1.00 ± 0.06	0.54 ± 0.02	0.25 ± 0.01
Propensity	1.00 ± 0.04	0.50 ± 0.07	0.35 ± 0.00
Variance	0.97 ± 0.05	0.53 ± 0.07	0.26 ± 0.01
	$\rho = 1$		
Random	0.98 ± 0.01	0.49 ± 0.03	0.25 ± 0.01
Propensity	1.02 ± 0.06	0.47 ± 0.10	0.33 ± 0.03
Variance	0.95 ± 0.01	0.48 ± 0.02	0.25 ± 0.01

G. Broader Impacts

In Section 4 we learn that GATE may enable practical adoption of CATE estimation in low-sample settings, possibly yielding a positive impact in fields where data is costly. Furthermore, GATE helps address problems such as covariate shift

(particularly in low-sample regimes), further aiding the adoption of CATE inference in practice. However, extra care should be taken before relying on the LLM to guide decision-making in high-stakes domains.