A Computational Theory for Efficient Mini Agent Evaluation with Causal Guarantees

Anonymous Author(s)
Affiliation
Address
email

Abstract

In order to reduce the cost of experimental evaluation for agents, we introduce a computational theory of evaluation for mini agents: build evaluation model to accelerate the evaluation procedures. We prove upper bounds of generalized error and generalized causal effect error of given evaluation models for infinite agents. We also prove efficiency, and consistency to estimated causal effect from deployed agents to evaluation metric by prediction. To learn evaluation models, we propose a meta-learner to handle heterogeneous agents space problem. Comparing with existed evaluation approaches, our (conditional) evaluation model reduced 24.1% to 99.0% evaluation errors across 12 scenes, including individual medicine, scientific simulation, social experiment, business activity, and quantum trade. The evaluation time is reduced 3 to 7 order of magnitude per subject comparing with experiments or simulations.

1 Introduction

Throughout the history of computer science, manually handing the data has been replaced by computational approach, such as computational linguistic (OpenAI et al. [2024], DeepSeek-AI et al. [2025]), computational biology (Abramson et al. [2024]), and computational learning (Valiant [1984]). Here we extend the computational theory to the ubiquitous domain of evaluation.

Randomized controlled trials (Fisher [1974], Box et al. [2005]) are widely regarded as the gold standard for evaluating the efficacy of interventions, such as drugs and therapies, due to their ability to minimize bias and establish causal relationships between interventions and outcomes. By randomly assigning participants to intervention group or control group, RCTs effectively balance both observed and unobserved confounding variables, thereby mitigating the influence of hidden common causes (Reichenbach [1999]). However, conducting RCTs to assess every potential agent is often prohibitively expensive and, in certain cases, unfeasible as shown in appendix A. This challenge is particularly pronounced in fields like artificial intelligences, where models may possess hundreds of parameters, resulting in an infinite and high-dimensional space of potential interventions. Evaluating such agents through randomized experiments is resource-intensive, demanding substantial investments of time, finances, personnel, and materials (Speich et al. [2018]). Moreover, the extended duration of these experiments hampers the rapid iteration and optimization of agents, rendering the process inefficient. Given these limitations, alternative methodologies that balance rigorous evaluation with practical feasibility are essential, especially in rapidly evolving domains requiring swift assessment and deployment of new agents.

An improvement is to build the centralized A/B test platform, which has been proven successes in application updates and recommendation in Byte Dance (Byte-Dance). The implementation of centralized A/B testing platforms can significantly reduce the costs associated with conducting randomized trials for researchers. However, this cost reduction does not inherently enhance the utility derived from each individual trial and it still can not handle the infinite evaluation subjects problem. Notably, the lower cost and increased accessibility of such platforms can lead to a substantial rise in the number of randomized trials conducted. This proliferation of experiments raises ethical concerns regarding participant exposure to potential harm during the randomization process. For instance, as highlighted by Zhou et al. [2024b], the automation and scalability of A/B testing frameworks, while beneficial for rapid experimentation, necessitate careful consideration of ethical implications to safeguard participant well-being. Polonioli et al. [2023] emphasizes the need for ethical and responsible experimentation to protect users and society.

In addition to other agent evaluation works in appendix B, evaluatology (Zhan et al. [2024]) emerges as an independent filed in recent years. However, there is still a lack of a theory to help us find an effective and efficient evaluation system that satisfies the concerns and interests of the stakeholders. Here, we introduce a computational theory to evaluate the effect of given agents. The main benefit to evaluate by computation is its efficiency and low cost. If we learned the connection between the experiment result and pre-experiment agents by a computational model, the evaluation cost will be reduced dramatically comparing with performing experiments.

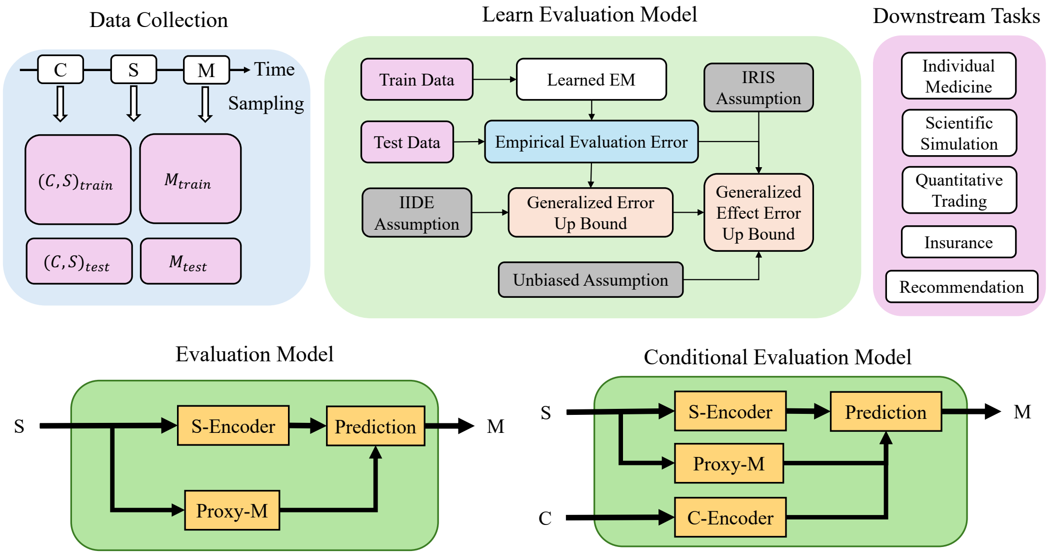


Figure 1: Procedure of computational evaluation. C is evaluation condition, S is evaluation subject (agent), M is evaluation metric, EM is evaluation model.

The procedure of our computational evaluation is listed in figure 7. First, data of evaluation condition, deployed agents, and evaluation metric are collected. If causal relationship is needed, the subject is required to be independently, randomly, identically sampled (IRIS) so that the causal effect from the subject to the metric can be bounded in the following steps. Next, evaluation model is learned from the collected data, and generalized evaluation error is upper bounded under independently, identically distributed error (IIDE) assumption. Also, the evaluation model is assumed unbiased to bound the causal effect from agent to metric. Then, learned evaluation model is applied to the downstream applications once the upper bounds are low enough to satisfy the real need. The mentioned assumptions are assessed by testing their necessary conditions.

In addressing on on the heterogeneous agents space problem in evaluation model learning, we propose a meta algorithm to parameterize agents, learn (conditional) evaluation models, and inference separately. The upper bounds of generalized error for both metric and its first order difference on agents are minimized by minimizing the empirical error. Also, learning evaluation model from evaluation samples directly is very challenging in some cases. So we add a proxy module for the agents in addressing the challenge. The proxy module will create some proxy metrics according some existing computational evaluation methods, which can help us to improve the performance of evaluation model.

In our evaluation experiment, we test evaluation model on 11 scenes including individual medicine, scientific simulation, insurance, advertisement, trade. Comparing with the existed computational evaluation methods, our method reduced 24.1% to 99.0% evaluation errors while remain the advantage of computational evaluation efficiency. We also test our conditional evaluation model in the quantitative trade scene. Comparing with baseline backtesting strategy, our evaluation model reduced 89.4% evaluation errors. The evaluation acceleration ratio is ranged from 1000 times to 10000000 times comparing with experimental or simulation based evaluation in the 12 scenes.

2 Notation

Variable	Name	Description
S	evaluation subject (agents)	the operable and indivisible unit that we want to deploy, remove, understand, and modify
C	evaluation condition	set of pre-evaluation features of evaluation subject, including inner attributes and environments that identify a real evaluation process
M	evaluation metric	subset of stakeholders' buy-in metrics
EM or $\hat{f}$	evaluation model	a function that takes an evaluation subject and evaluation conditions as inputs and produces evaluation metrics as its output
L	evaluation error function (ranged from 0 to 1)	a function to measure the discrepancy between output of evaluation model and true metric
E_{emp}	empirical evaluation error	average evaluation error of empirical sample
E_{gen}	generalized evaluation error	expected evaluation error of super-population

Table 1: Notation table for used variables.

Table 1 shows the notation we used in this paper. In order to simplify the evaluation problem, evaluation metric and difference of evaluation metric are assumed as measurable and pairwise comparable. The computational evaluation problem is modeled as computing the evaluation metrics for evaluation subject with certain evaluation conditions. Unless otherwise specified, all evaluation subjects in this paper are mini agents.

3 Theoretical analysis

3.1 Upper bound of generalized evaluation error

Theorem 1. ***Upper bound****. Given any evaluation model $\hat{f}$, $P(E_{gen}(\hat{f}) < E_{emp}(\hat{f}) + \sqrt{\frac{1}{2n}\ln(\frac{1}{\sigma})}) \geq 1-\sigma$ where n is number of independent identical distributed error (IIDE) measurements, $0 < 1-\sigma < 1$ is confidence.*

In Theorem 1, we show that the generalization error E_{gen} is bounded by the empirical error E_{emp}. the number of error measurements n, and the significance level $1-\sigma$ under the IIDE assumption. This result underpins our ability to use a finite number of evaluation samples to compare the performance of evaluation models across an infinite subject and condition space. For example, when the evaluation subject space is extremely large or infinite—as in the cases of AI diagnosis agents, AI treatment agents, or quantitative strategy agents—Theorem 1 allows us to generalize our empirical analysis to the entire subject space. Similarly, in evaluation condition spaces, such as those involving AI-driven discovery of unseen materials based on molecular and structural conditions, the generalization errors for any evaluation model can be bounded using the same theorem.

Once the upper bound is smaller than given error tolerance, evaluation model f is regarded to meet the practical needs with high probability $1-\sigma$. Additionally, Theorem 1 provides a method to estimate the number of error measurements (or evaluation samples) required prior to sampling.

We emphasize that IIDE can also be satisfied if the non-IID components of (C, S, M) was offset by the evaluation model or the evaluation error function (such as non-misspecified time-series model on time-series data) although independent identical distributed (C, S, M) is a sufficient condition of IIDE. In fact, the assumption of independent and identically distributed samples is considered excessively strong in this paper, particularly with respect to IID conditions and IID outcomes.

3.2 Hidden common cause

The hidden common cause between subject and metric may introduce the confounding bias if the evaluation task is causality sensitive, such as individual medicine. Causal inference is often used to handle hidden common cause from data. However, applying existed methods to handle unmeasured hidden common cause problem in evaluation faced those challenges:

- **Hidden identical common cause (HIC)**. If the assignment of S is determined by an unlisted identical common cause E which can be denoted as $S := E$, and there exists effect from E to M. We may mistakenly attributing the effect on M of E to S. The identical hidden confounder can *never* be excluded from statistical analysis due to the identical property if S was never controlled by randomization. Therefore, effect from S to M is unidentifiable in all kinds of existed non-intervention based causal analysis, including diagram-based identification (Pearl [1995], Tian and Pearl [2002], Shpitser and Pearl [2006], Bareinboim and Pearl [2012], YAN [2025]), neural-based identification (Xia et al. [2021]), statistical-based identification (Jaber et al. [2019]), observation study (ROSENBAUM and RUBIN [1983]), causal representation learning (Schölkopf et al. [2021]), and causal discovery (Spirtes and Glymour [1991], Zhang and Hyvärinen [2009]).
- **High-dimensional evaluation subject**. Potential outcomes (Splawa-Neyman et al. [1990], Holland [1986], Imbens and Rubin [2015]) often assumes each evaluation subject must be sampled once which often do not established for high-dimensional evaluation subjects. For instance, parameters of AI model can be hundreds or thousands and more complex which structure is different from continuous dose-level analysis (Schwab et al. [2020], Nie et al. [2021], Zhu et al. [2024], Nagalapatti et al. [2024]) with low dimension.

For causality sensitive task, any hidden common cause between subject and metric is excluded including HIC if the subject was randomly, independently, identically sampled or assigned as shown in theorem 2 and theorem 4. For high-dimensional evaluation subject, the causal effect of evaluation error is bounded and evaluation result can be generalized to the whole subject space whatever how large the space size as shown in theorem 4.

Theorem 2. ***Strict causal advantage***. *Given $\{c_i, s_i, m_i\}_{i=1}^{n}$, where c is evaluation condition, s is independent from c, and e, and s is independently, randomly, identically sampled (IRIS) from distribution $Pr(S)$, m is generate by an unknown and inaccessible function $m = f(c, e, s)$, e is other unlisted random variable, evaluation error function L is mean square error, $\forall \hat{f}_1$, $\forall \hat{f}_2$ where $\hat{f}_1$ and $\hat{f}_2$ are unbiased estimation of f given c (CU), we have* ***Efficiency***: *if $E_{gen}(\hat{f}_1) < E_{gen}(\hat{f}_2)$ for any s, then $\forall s_a \in \mathbf{S}$, $\forall s_b \in \mathbf{S}$, $\forall c \in \mathbf{C}$, $E_{gen}(\hat{f}_1(c, s_a) - \hat{f}_1(c, s_b)) < E_{gen}(\hat{f}_2(c, s_a) - \hat{f}_2(c, s_b))$;* ***Consistency***: *if $\lim_{\hat{f} \to f} E_{gen}(\hat{f}) = 0$, then $\forall s_a \in \mathbf{S}$, $\forall s_b \in \mathbf{S}$, $\forall c \in \mathbf{C}$, $\lim_{\hat{f} \to f} E_{gen}(\hat{f}(c, s_a) - \hat{f}(c, s_b)) = 0$.*

From theorem 2, minimization of generalized mean square causal effect evaluation error can be performed by minimizing generalized evaluation error greedily until the evaluation error is close to 0. However, we can not get the generalized evaluation errors in reality. An upper bound with empirical errors was inferred in the theorem 3.

Theorem 3. ***Causal bound with positivity***. *$\forall s_a \in S, \forall s_b \in S$, given any unbiased evaluation model $\hat{f}(c, s_a)$ and $\hat{f}(c, s_b)$,*

$$P(E_{gen}(\Delta\hat{f}) < 2\max\{E_{emp}(a) + \sqrt{\frac{1}{2n_a}\ln(\frac{2}{\sigma})}), E_{emp}(b) + \sqrt{\frac{1}{2n_b}\ln(\frac{2}{\sigma})}\}) \geq 1 - \sigma$$

, where $\Delta\hat{f} = \hat{f}(c, s_a) - \hat{f}(c, s_b)$, n_a and n_b are number of independently, randomly, identically sampled error measurements (IIDE) where $s = s_a$ and $s = s_b$, $0 < 1 - \sigma < 1$ is confidence, evaluation error function L is mean square error ranged from 0 to 1.

In practice, n_a and n_b are typically zero for infinite evaluation subject (non-positivity Imbens and Rubin [2015]); in such cases, upper bound of generalized effect error is given by Theorem 4.

Theorem 4. ***Causal bound with non-positivity***. *Given unbiased evaluation model $\hat{f}$, then $\forall s_a \in S, \forall s_b \in S$,*

$$P(E_{gen}(\Delta\hat{f}) < 2(E_{emp}(\hat{f}) + \sqrt{\frac{1}{2n}\ln(\frac{1}{\sigma})})) \geq 1 - \sigma$$

, where $\Delta \hat{f} = \hat{f}(c, s_a) - \hat{f}(c, s_b)$, n is number of independently, randomly, identically sampled error measurements (IIDE), $0 < 1 - \sigma < 1$ is confidence, s is independently, randomly, identically sampled (IRIS), evaluation error function L is mean square error ranged from 0 to 1.

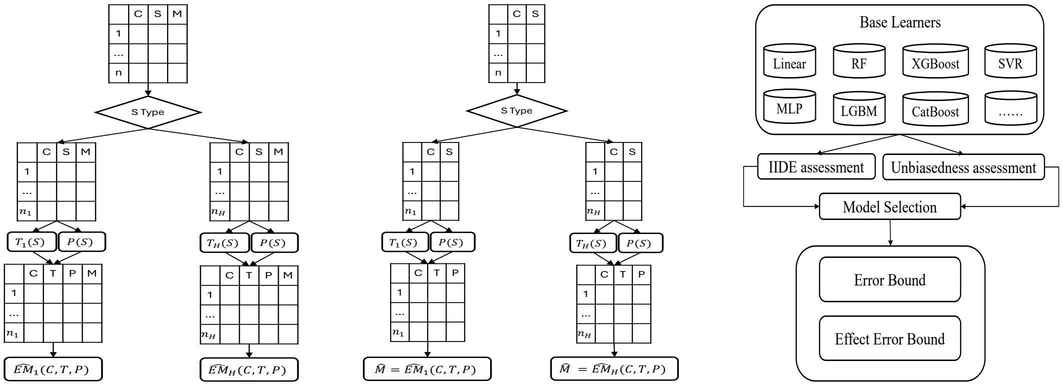
Theorem 4 provides the rationale for why our evaluation models can learn the causal effect from subjects to metrics rather than mere correlations when the evaluation subjects are independently, randomly, and identically sampled (IRIS), the evaluation models are unbiased, and the evaluation errors are independently, randomly, identically distributed (IIDE). Also, effect error is upper bounded by its bias even the model is not unbiased. Proofs of all the theorems were given in appendix C.

Comparing with Saito and Yasui [2020] and Alaa and Van Der Schaar [2019], our unbiased assumption has potential to be rejected from real data testing as shown in the experiment part while they did not provide approaches to reject their unbiased causal effect assumption (Saito and Yasui [2020]) or min-max optimal effect estimator assumption (Alaa and Van Der Schaar [2019]) from real data without counterfactual. Also, IID samples assumption in those works (Saito and Yasui [2020], Johansson et al. [2022], Alaa and Van Der Schaar [2019], Shalit et al. [2017], Li and Pearl [2022], Wang et al. [2022]) is too strong for causal evaluation, especially IID conditions and IID outcomes.

4 Learning evaluation models

In evaluation processes, two critical factors are evaluation error and evaluation cost. Our objective is to reduce evaluation cost while maintaining an acceptable level of generalized evaluation (causal) error through computational evaluation. To achieve this, we can minimize empirical evaluation error because upper bound size is fixed given number of error measurements n and confidence requirement $1 - \sigma$ in theorem 1 and theorem 4.

Directly learning an evaluation model from evaluation samples presents significant challenges. To address this, we introduce a proxy module that generates surrogate metrics using established computational methods, such as dataset-based approaches including bootstrapping, hold-out validation, and cross-validation. These techniques enhance the performance of the evaluation model by providing more robust and reliable features. Furthermore, to effectively manage heterogeneous agents, we employ vectorization strategies, develop specialized sub-models, and conduct inferences tailored to each agent type. This approach ensures that the evaluation model can accommodate the diverse characteristics inherent in heterogeneous agent spaces.



(a) Evaluation model learning. (b) Evaluation model inference. (c) Evaluation model Selection.

Figure 2: Learn evaluation models from data. P is proxy metrics of evaluation subject, and T_i is tensorization function for subject type i.

The proposed meta-learning algorithm for evaluation model is illustrated in Figure 2. Initially, data are classified based on subject type. Subsequently, distinct vectorization methods are applied corresponding to each subject type. Proxy metrics are then computed using predefined proxy functions, facilitating the development of tailored evaluation models. For the assessment of unseen subjects, the process involves computing proxy values, determining the subject type, selecting the appropriate vectorization function, and inputting these into the corresponding evaluation model to predict relevant metrics. The base learners for the evaluation models can vary and include algorithms such as linear models, Random Forests (RF), Multi-Layer Perceptrons (MLP), CatBoost, XGBoost, and LightGBM. Following the development of these evaluation models, an upper bound is calculated to check whether the models meet practical requirements and constraints.

5 Assessing assumptions

The IRIS assumption can be satisfied in data collection while the other assumptions may not be satisfied in real world. While the IIDE assumption could be trivially met by employing a random number generator as the outputs for IID outcome, this approach is devoid of practical utility, as it fails to reflect any substantive patterns or domain-specific knowledge required for realistic evaluation modeling. In order to assess the IIDE and unbiased assumptions, statistical analyses can be employed to verify necessary conditions under those assumptions, providing credibility to the theorems derived from this assumption for someone. It is important to note that these assessments do not conclusively prove the assumptions. Instead, they serve as methods to potentially refute the assumptions under specific conditions, rather than confirming its validity.

5.1 Assessing IIDE assumption

In highly controlled environments, such as CPU evaluation, the assumption can be rigorously satisfied by independently and randomly selecting tested chips and configurations from an identical distribution. However, in many open environments, the IIDE assumption is not directly testable on real-world data, such as individual medicine, and social experiment.

First, the normalized means of error subsets should follow a common Gaussian distribution, as implied by the central limit theorem. To test this, we randomly generate 30 subsets and apply the D'Agostino-Pearson test to assess whether their normalized means adhere to a single Gaussian distribution. Second, error distributions across different splits should be identical. We randomly partition the errors into 30 subsets and use the Kolmogorov-Smirnov test to determine whether the distributions across these subsets are statistically indistinguishable.

5.2 Assessing Unbiased assumption

Our effect error bounds requires unbiased evaluation models while sometime the evaluation model may not be unbiased. To assess the unbiasedness of the evaluation model, we employ the following statistical procedures: 1) Global Unbiasedness Assessment: We conduct a one-sample t-test on the entire dataset to determine if the model's errors have a mean of zero, indicating unbiasedness across the universal set; 2) Subset Unbiasedness Assessment: We randomly select 30 subsets from the dataset and perform one-sample t-tests on each to evaluate whether the mean errors within these subsets deviate significantly from zero. This step examines the consistency of the model's unbiasedness across different data partitions.

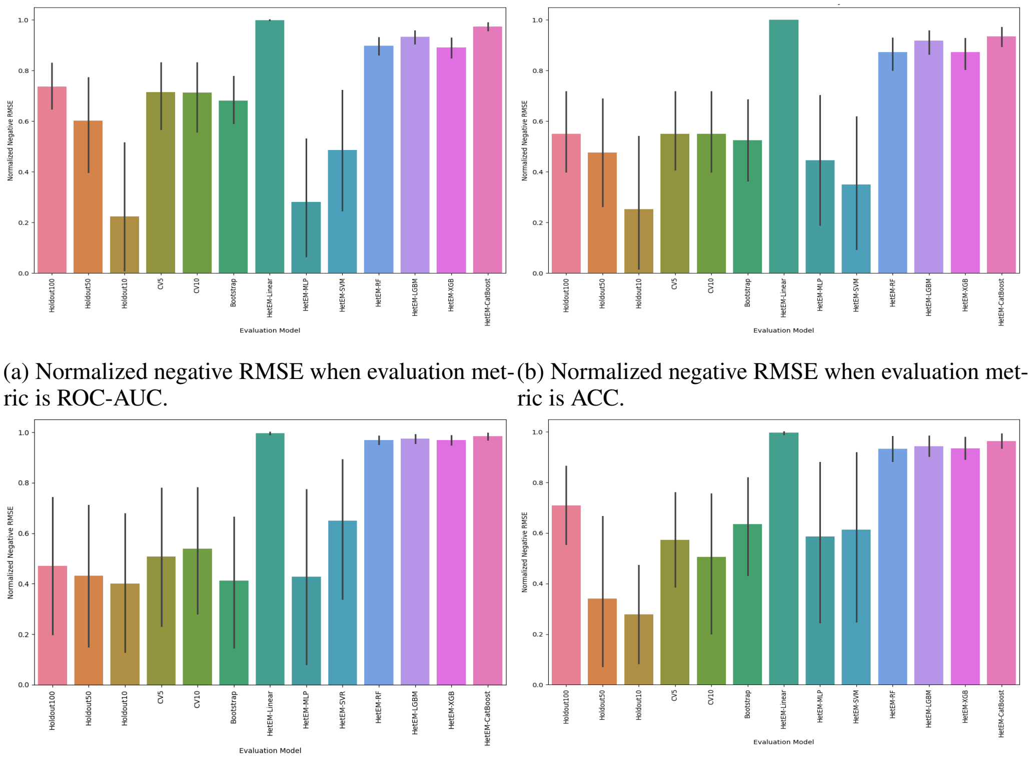
6 Experiment

We performed experiments for evaluation model without condition and evaluation model with condition respectively in 12 scenes of mini agent evaluation, including randomized experiment prediction, scientific simulation, advertising, insurance, and quantum trade. The detailed evaluation scenes (including experiment setting), assumption assessing, hyperparameter tuning, and evaluation results are listed in the appendix D, E, F, G.

6.1 Evaluation model

To demonstrate the generalizability of our evaluation model in agent space, we consider 11 distinct scenes. In these scenes, computational evaluation can be helpful to reduce experimental cost, shorten simulation time, or enhance the efficiency of advertising and insurance sales.

Although subjects are independently, randomly, and identically sampled from the subject space, they cannot be randomly deployed in real-world settings for direct measurement of benefit-related metrics (such as mortality, rehabilitation time) due to experimental limitation. Instead, our evaluation system defines the evaluation metric as the post-experiment metric on unseen test data with randomized treatment. The metrics capture the causal effect from subject to post-experiment outcome, as established in Theorem 4. Empirical studies (Ai et al. [2022], Zhou et al. [2023], Zhou et al. [2024a], Cheng et al. [2022], Gentzel et al. [2021]) have also demonstrated the correlation between the agents' predictive performance on randomized experimental data and its performance following randomized deployment. Accordingly, we employ these predictive metrics to approximate randomized post-deployment performance, which can then be integrated into the AI decision via an appropriate utility function.



(a) Normalized negative RMSE when evaluation metric is ROC-AUC.

(b) Normalized negative RMSE when evaluation metric is ACC.

(c) Normalized negative RMSE when evaluation metric is RMSE.

(d) Normalized negative RMSE when evaluation metric is R^2.

Figure 3: Normalized empirical negative RMSE of evaluation models crossing scenes (6 scenes in figure 3a and figure 3b, 5 scenes in figure 3c and figure 3d. The confidence level of confidence interval bar is set as 95%. The baselines are holdout 100%, holdout 50%, holdout 10%, 5-fold cross-validation, 10-fold cross-validation, and bootstrap. We test different base learners (Linear, MLP, SVM/SVR, Random forest, XGBoost, LighGBM, and CatBoost) for heterogeneous subject space.

Across the 11 scenes, we compare our evaluation model against 6 existing validation methods that do not incorporate prior information from data. The normalized negative empirical RMSE is illustrated in Figure 3. Our best evaluation model (Het-Linear) achieves error reductions of 94.7% to 99.0% in RMSE, 81.7% to 95.5% in R^2, 24.1% to 77.0% in ROC-AUC, and 39.2% to 89.5% in ACC relative to the Holdout-100 validation method. This performance improvement is attributable to the effective modeling of the relationships among the subject, its proxy metrics, and the evaluation metrics. Notably, models that capture spurious correlations (MLP and SVR without hyperparameter tuning) fail to reduce evaluation error.

6.2 Conditional evaluation model

In order to demonstrate the generalizbility of conditional evaluation model on the condition space further, we introduce the trade backtesting task of A-share market in China to test the performance of conditional evaluation model. In this scene, the computational evaluation can be helpful to exclude the influence of hidden common cause between trade strategy and return of invest (RoI), and to reduce the evaluation cost of trade strategy's deployment.

The evaluation subjects are trade models whose inputs are last day's variables of the stock and output are one of three decisions (buy 1 hand, sell 1 hand, and hold) at open this day. The metric is Return of Invest of a subject in a time slot (10 days) in future. The condition is the stock's time-series variables in the last time slot. In this scene, we not only measure the subject and its metric, but also consider the pre-trade variables (condition) of the subject in an attempt to reduce evaluation error further. We use the unseen future data and performance of unseen subjects for conditional evaluation model testing.

Similarly, we build the evaluation systems by combining the real data and a float model whose input is subject vector and last day's variables of the stock, and output is a small float ratio of opening price. We want to use it to simulate the exclusion of hidden common cause on the stock market by the data without the randomized deployment of the subject.

Backtesting method	RMSE of RoI
Baseline (Last10Days)	3.067
HetEM (Linear)	2.163
HetEM (CatBoost)	0.527

Table 2: Performance of backtesting methods.

We compare our conditional evaluation models with a baseline backtesting methods as shown in the table 2. The reason we did not take the time cross-validation, k-fold cross-validation, and combinatorial symmetry cross-validation (Bailey et al. [2013]) into baselines is that those methods are train-valid fused rather than evaluation-targeted methods. When the validation part is separated, they will give the same trivial constant prediction which can not utilize the evaluation conditions for each time slot.

Our best conditional evaluation models (HetEM-CatBoost) for heterogeneous subjects reduced 89.4% evaluation errors comparing with the baseline method. Even the linear conditional evaluation model can reduce the 34.0% RMSE of estimated RoI. The average RMSE of estimated RoI in a future time slot is about 0.527, which may be reduced further when adding more factors of stocks, and designing novel neural network architectures. From the result of different evaluation methods, the computational evaluation approach has higher potential to solve the backtest over-fitting problem (Bailey et al. [2013]) in stock market.

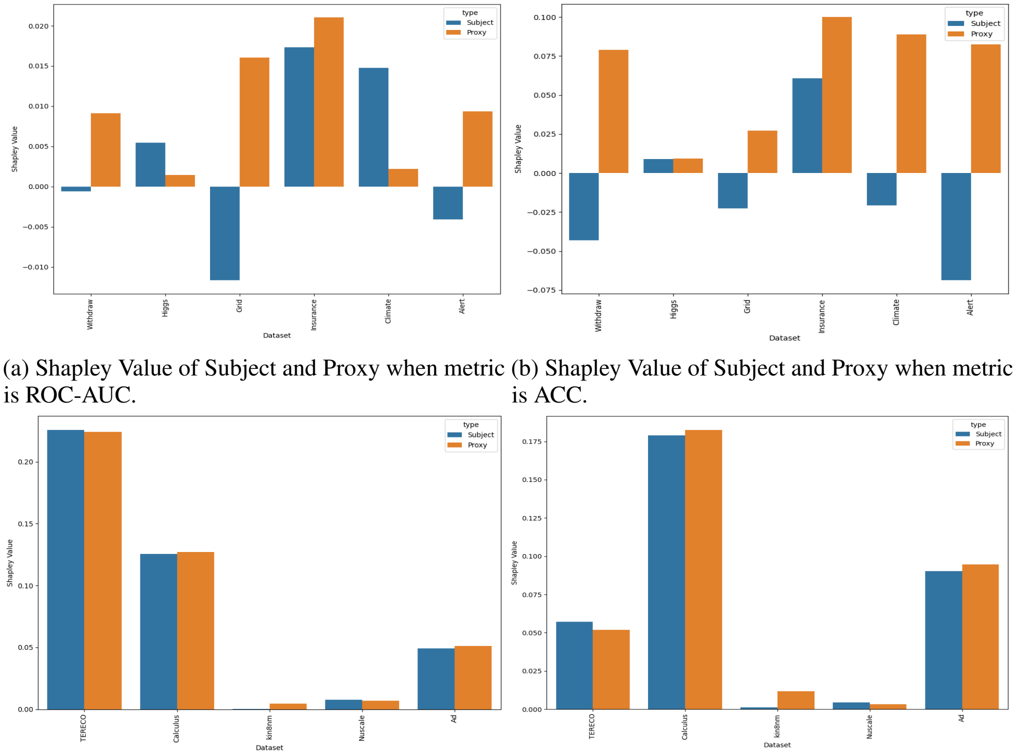
6.3 Interpretability

In order to understand the contribution of subject vector, and proxy metric in the evaluation of our best evaluation model respectively, we visualize the Shapley value (Shapley [1997], Lundberg and Lee [2017]) for subject vector and proxy metric of all models as shown in figrue 4. The null set's value is set as the negative RMSE of holdout-100. From the visualization, the main contribution for heterogeneous subject whose metrics are ROC-AUC and ACC, is from the relationship between the proxy metric and true metric despite of an example in climate prediction scene. For heterogeneous subjects whose metrics are RMSE and R^2, the contribution of subject and proxy metric is almost equal to each other. It reveals the effectiveness of the introduced proxy module and subject vectorization module in our learning algorithm respectively.

In order to understand the contribution of the condition vector, subject vector, and proxy metrics in the evaluation of our best conditional evaluation model, we use Shapley values to explain the contribution of modules (Shapley [1997], Lundberg and Lee [2017]). The null set is set as the baseline backtesting method. The contribution rate of condition vector, subject vector, and proxy metrics for the HetEM(CatBoost) are 28.8% (0.729 uplift RMSE of RoI), 35.6% (0.905 uplift RMSE of RoI), and 35.6% (0.906 uplift RMSE of RoI) respectively. It reveals the importance of evaluation condition to reduce the evaluation errors in the A-share trade scenes of China.

6.4 Evaluation cost

The primary advantage of our computational evaluation method is its efficiency and low cost. Rather than conducting an unbounded number of experiments for infinitely many subjects (e.g., AI models or other agents), our approach only requires the collection of initial data, training an evaluation model, and subsequently applying it in real-world scenes.



(a) Shapley Value of Subject and Proxy when metric is ROC-AUC.

(b) Shapley Value of Subject and Proxy when metric is ACC.

(c) Shapley Value of Subject and Proxy when metric is RMSE.

(d) Shapley Value of Subject and Proxy when metric is R^2.

Figure 4: Shapley value of subject vector and proxy metrics on 11 scenes (6 scenes in figure 4a and figure 4b, 5 scenes in figure 4c and figure 4d. The outcome of null set is set as the performance of holdout-100, and other outcomes are the negative RMSE of the linear evaluation models.

The acceleration ratio of time achieved by our computational method ranges from 1,000 to 10,000,000 per evaluation subject in the 12 scenes as shown in appendix H. Although experimental and simulation-based evaluations can be accelerated via parallel processing (excluding additional evaluation costs), the same parallelization strategies can be readily applied to computational evaluation approach.

7 Limitation and conclusion

Limitation. First, our theoretical guarantees rely on three assumptions, which may not hold for all real-world scenes and agents. Second, while our method has been validated across multiple domains, its generalization in extremely high-dimensional settings (such as LLM) require further investigation. Third, lower error and tighter bound are required to apply evaluation models into real applications with small sample. Four, scalability of evaluation models across scenes is another problem in future study. Five, the post-experiment metric (such as mortality) of *randomized deployed mini-agents* with high utility should be collected to increase the value of computational evaluation.

Conclusion. In this work, we propose a novel computational framework for the evaluation of mini-agents that rigorously derives upper bounds on generalized evaluation error and effect error, and employs a meta-learning strategy to address heterogeneity in agent space. Our extensive experimental results, conducted over 12 diverse scenes—including individual medicine, scientific simulation, social experiments, business activity, and quantum trade—demonstrate significant error reductions (ranging from 24.1% to 99.0%) and substantial acceleration (up to 10^7) compared to traditional evaluation methods. These findings highlight the potential of our technical route to drastically reduce evaluation costs for rapid model iteration. Future work will focus on more benefit-related metrics, multi-source inputs, novel neural architectures for evaluation tasks and enhancing the validation of key assumption. The upper bound table and paradigm of computational evaluation were given in the appendix I and J.

References

Michael D Abramoff, Noelle Whitestone, Jennifer L Patnaik, Emily Rich, Munir Ahmed, Lutful Husain, Mohammad Yeadul Hassan, Md Sajidul Huq Tanjil, Dena Weitzman, Tinglong Dai, Brandie D Wagner, David H Cherwek, Nathan Congdon, and Khairul Islam. Autonomous artificial intelligence increases real-world specialist clinic productivity in a cluster-randomized trial. *NPJ Digit. Med.*, 6(1):184, October 2023.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, Sebastian W Bodenstein, David A Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B Fuchs, Hannah Gladman, Rishub Jain, Yousuf A Khan, Caroline M R Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, June 2024.

Meng Ai, Biao Li, Heyang Gong, Qingwei Yu, Shengjie Xue, Yuan Zhang, Yunzhou Zhang, and Peng Jiang. Lbcf: A large-scale budget-constrained causal forest algorithm. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2310–2319, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450390965. doi: 10.1145/3485447.3512103. URL https://doi.org/10.1145/3485447.3512103.

Ahmed Alaa and Mihaela Van Der Schaar. Validating causal inference models via influence functions. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 191–201. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/alaa19a.html.

Rabia Musheer Aziz, Mohammed Farhan Baluch, Sarthak Patel, and Abdul Hamid Ganie. LGBM: a machine learning approach for ethereum fraud detection. *Int. J. Inf. Technol.*, 14(7):3321–3331, December 2022.

David H Bailey, Jonathan M Borwein, Marcos Lopez de Prado, and Qiji Jim Zhu. The probability of back-test over-fitting. *SSRN Electron. J.*, 2013. URL http://dx.doi.org/10.2139/ssrn.2326253.

P Baldi, P Sadowski, and D Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nat. Commun.*, 5(1):4308, July 2014.

Elias Bareinboim and Judea Pearl. Causal inference by surrogate experiments: z-identifiability. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12, page 113–120, Arlington, Virginia, USA, 2012. AUAI Press. ISBN 9780974903989.

Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Paul Oka, Miruna Oprescu, and Vasilis Syrgkanis. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/py-why/EconML, 2019. Version 0.x.

Ioana Bica, Ahmed Alaa, and Mihaela Van Der Schaar. Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 884–895. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/bica20a.html.

Patrick Blöbaum, Peter Götz, Kailash Budhathoki, Atalanti A. Mastakouri, and Dominik Janzing. Dowhy-gcm: An extension of dowhy for causal inference in graphical causal models. *Journal of Machine Learning Research*, 25(147):1–7, 2024. URL http://jmlr.org/papers/v25/22-1258.html.

Jian Yang Bochen Lyu, Xuefeng Li. YLearn: A Python Package for Causal Inference. https://github.com/DataCanvasIO/YLearn, 2022. Version 0.2.x.

Ben Bogin, Kejuan Yang, Shashank Gupta, Kyle Richardson, Erin Bransom, Peter Clark, Ashish Sabharwal, and Tushar Khot. SUPER: Evaluating agents on setting up and executing tasks from research repositories. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12622–12645, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.702. URL https://aclanthology.org/2024.emnlp-main.702/.

George E P Box, J Stuart Hunter, and William G Hunter. *Statistics for experimenters*. Wiley Series in Probability and Statistics. Wiley-Blackwell, Chichester, England, 2 edition, May 2005.

Leo Breiman. *Mach. Learn.*, 45(1):5–32, 2001.

Byte-Dance. Understanding a/b testing products - a/b testing - volcano engine. https://www.volcengine.com/docs/6287/65798. Accessed: 2024-10-16.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

Lu Cheng, Ruocheng Guo, Raha Moraffah, Paras Sheth, K. Selçuk Candan, and Huan Liu. Evaluation methods and measures for causal learning algorithms. *IEEE Transactions on Artificial Intelligence*, 3(6):924–943, 2022. doi: 10.1109/TAI.2022.3150264.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

David K. Eng, Nishith B. Khandwala, Jin Long, Nancy R. Fefferman, Shailee V. Lala, Naomi A. Strubel, Sarah S. Milla, Ross W. Filice, Susan E. Sharp, Alexander J. Towbin, Michael L. Francavilla, Summer L. Kaplan, Kirsten Ecklund, Sanjay P. Prabhu, Brian J. Dillon, Brian M. Everist, Christopher G. Anton, Mark E. Bittman, Rebecca Dennis, David B. Larson, Jayne M. Seekins, Cicero T. Silva, Arash R. Zandieh, Curtis P. Langlotz, Matthew P. Lungren, and Safwan S. Halabi. Artificial intelligence algorithm improves radiologist performance in skeletal age assessment:

A prospective multicenter randomized controlled trial. *Radiology*, 301(3):692–699, 2021. doi: 10.1148/radiol.2021204021. URL https://doi.org/10.1148/radiol.2021204021. PMID: 34581608.

Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Wang Siyuan, Zhongyu Wei, and Fei Huang. AI hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10183–10213, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.680/.

Sebastian Fischer. kin8nm. https://www.openml.org/search?type=data&status=active&id=44980, 2022. Accessed: March 13, 2025.

Sir Ronald Aylmer Fisher. *The design of experiments*. 1974.

Janos Galambos. Bonferroni inequalities. *The Annals of Probability*, 5(4):577–581, 1977. ISSN 00911798, 2168894X. URL http://www.jstor.org/stable/2243081.

Amanda M Gentzel, Purva Pruthi, and David Jensen. How and why to use experimental data to evaluate methods for observational causal inference. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3660–3671. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/gentzel21a.html.

Gokagglers. Sales conversion optimization. https://www.kaggle.com/datasets/loveall/clicks-conversion-tracking/data, 2018. Accessed: March 13, 2025.

Google. Causalimpact: An r package for causal inference using bayesian structural time-series models. https://github.com/google/CausalImpact. Accessed: 2025-5-8.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2282952.

Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986. doi: 10.1080/01621459.1986.10478354. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1986.10478354.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, January 2025.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. Mlagentbench: evaluating language agents on machine learning experimentation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Nguyen Huu Tiep. Lattice-physics (PWR fuel assembly neutronics simulation results). UCI Machine Learning Repository, 2024. DOI: https://doi.org/10.24432/C5BK64.

Guido W Imbens and Donald B Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, Cambridge, England, May 2015.

inovex. Justcause. https://github.com/inovex/justcause. Accessed: 2025-5-8.

Amin Jaber, Jiji Zhang, and Elias Bareinboim. Causal identification under Markov equivalence: Completeness results. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2981–2989. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/jaber19a.html.

Jingru Jia, Zehua Yuan, Junhao Pan, Paul E. McNamara, and Deming Chen. Decision-making behavior evaluation framework for llms under uncertain context. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 113360–113382. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/cda04d7ea67ea1376bf8c6962d8541e0-Paper-Conference.pdf.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 3020–3029, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/johansson16.html.

Fredrik D. Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *Journal of Machine Learning Research*, 23(166):1–50, 2022. URL http://jmlr.org/papers/v23/19-511.html.

Diviyan Kalainathan, Olivier Goudet, and Ritik Dutta. Causal discovery toolbox: Uncovering causal relationships in python. *Journal of Machine Learning Research*, 21(37):1–5, 2020. URL http://jmlr.org/papers/v21/19-187.html.

Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann. Causal inference using graphical models with the r package pcalg. *Journal of Statistical Software*, 47(11):1–26, 2012. doi: 10.18637/jss.v047.i11. URL https://www.jstatsoft.org/index.php/jss/article/view/v047i11.

Nathan Kallus. DeepMatch: Balancing deep covariate representations for causal inference using adversarial training. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5067–5077. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/kallus20a.html.

Roni W. Kobrosly. causal-curve: A python causal inference package to estimate causal dose-response curves. *Journal of Open Source Software*, 5(52):2523, 2020. doi: 10.21105/joss.02523. URL https://doi.org/10.21105/joss.02523.

Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat. Med.*, 24(11):1716–1720, November 2018.

Laird Kramer, Edgar Fuller, Charity Watson, Adam Castillo, Pablo Duran Oliva, and Geoff Potvin. Establishing a new standard of care for calculus using trials with randomized student allocation. *Science*, 381(6661):995–998, 2023. doi: 10.1126/science.ade9803. URL https://www.science.org/doi/abs/10.1126/science.ade9803.

Arun Kumar. customer. https://www.kaggle.com/datasets/arun0309/customer, 2023. Accessed: March 13, 2025.

Jaron J. R. Lee, Rohit Bhattacharya, Razieh Nabi, and Ilya Shpitser. Ananke: A python package for causal inference using graphical models, 2023. URL https://arxiv.org/abs/2301.11477.

Ang Li and Judea Pearl. Bounds on causal effects and application to high dimensional data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(5):5773–5780, Jun. 2022. doi: 10.1609/aaai.v36i5.20520. URL https://ojs.aaai.org/index.php/AAAI/article/view/20520.

Fengzi Li. Causal effect engine. https://github.com/L-F-Z/CEE. Accessed: 2025-5-8.

Jian'an Li, Wenguang Xia, Chao Zhan, Shouguo Liu, Zhifei Yin, Jiayue Wang, Yufei Chong, Chanjuan Zheng, Xiaoming Fang, Wei Cheng, and Jan D Reinhardt. A telerehabilitation programme in post-discharge COVID-19 patients (TERECO): a randomised controlled trial. *Thorax*, 77(7): 697–706, July 2022.

Sheng Li and Yun Fu. Matching on balanced nonlinear representations for treatment effects estimation. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/b2eeb7362ef83deff5c7813a67e14f0a-Paper.pdf.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/94b5bde6de888ddf9cde6748ad2523d1-Paper.pdf.

D. D. Lucas, R. Klein, J. Tannahill, D. Ivanova, S. Brandon, D. Domyancic, and Y. Zhang. Failure analysis of parameter-induced simulation crashes in climate models. *Geoscientific Model Development*, 6(4):1157–1171, 2013. doi: 10.5194/gmd-6-1157-2013. URL https://gmd.copernicus.org/articles/6/1157/2013/.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Nan Lv, Thomas Kannampallil, Lan Xiao, Corina R Ronneberg, Vikas Kumar, Nancy E Wittels, Olusola A Ajilore, Joshua M Smyth, and Jun Ma. Association between user interaction and treatment response of a voice-based coach for treating depression and anxiety: Secondary analysis of a pilot randomized controlled trial. *JMIR Hum Factors*, 10:e49715, Nov 2023. ISSN 2292-9495. doi: 10.2196/49715. URL https://humanfactors.jmir.org/2023/1/e49715.

A Luke MacNeill, Shelley Doucet, and Alison Luke. Effectiveness of a mental health chatbot for people with chronic diseases: Randomized controlled trial. *JMIR Form Res*, 8:e50025, May 2024. ISSN 2561-326X. doi: 10.2196/50025. URL https://formative.jmir.org/2024/1/e50025.

David C Mohr, Stephen M Schueller, Kathryn Noth Tomasino, Susan M Kaiser, Nameyeh Alam, Chris Karr, Jessica L Vergara, Elizabeth L Gray, Mary J Kwasny, and Emily G Lattie. Comparison of the effects of coaching and receipt of app recommendations on depression, anxiety, and engagement in the intellicare platform: Factorial randomized controlled trial. *J Med Internet Res*, 21(8): e13609, Aug 2019. ISSN 1438-8871. doi: 10.2196/13609. URL http://www.jmir.org/2019/8/e13609/.

Lokesh Nagalapatti, Akshay Iyer, Abir De, and Sunita Sarawagi. Continuous treatment effect estimation using gradient interpolation and kernel smoothing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(13):14397–14404, Mar. 2024. doi: 10.1609/aaai.v38i13.29353. URL https://ojs.aaai.org/index.php/AAAI/article/view/29353.

Ashwin Nayak, Sharif Vakili, Kristen Nayak, Margaret Nikolov, Michelle Chiu, Philip Sosseinheimer, Sarah Talamantes, Stefano Testa, Srikanth Palanisamy, Vinay Giri, and Kevin Schulman. Use of voice-based conversational artificial intelligence for basal insulin prescription management among patients with type 2 diabetes: A randomized clinical trial. *JAMA Network Open*, 6(12): e2340232–e2340232, 12 2023. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2023.40232. URL https://doi.org/10.1001/jamanetworkopen.2023.40232.

Lizhen Nie, Mao Ye, qiang liu, and Dan Nicolae. {VCN}et and functional targeted regularization for learning causal effects of continuous treatments. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=RmB-88r9dL.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

paras2612. Causebox-a-causal-inference-toolbox-for-benchmarkingtreatment-effect-estimators-with-machine-learning-methods. https://github.com/paras2612/CauseBox. Accessed: 2025-5-8.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2337329.

John D. Piette, Sean Newman, Sarah L. Krein, Nicolle Marinec, Jenny Chen, David A. Williams, Sara N. Edmond, Mary Driscoll, Kathryn M. LaChappelle, Robert D. Kerns, Marianna Maly, H. Myra Kim, Karen B. Farris, Diana M. Higgins, Eugenia Buta, and Alicia A. Heapy. Patient-centered pain care using artificial intelligence and mobile health tools: A randomized comparative effectiveness trial. *JAMA Internal Medicine*, 182(9):975–983, 09 2022. ISSN 2168-6106. doi: 10.1001/jamainternmed.2022.3178. URL https://doi.org/10.1001/jamainternmed.2022.3178.

Andrea Polonioli, Riccardo Ghioni, Ciro Greco, Prathm Juneja, Jacopo Tagliabue, David Watson, and Luciano Floridi. The ethics of online controlled experiments (A/B testing). *Minds Mach. (Dordr.)*, 33(4):667–693, September 2023.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf`.

quantumblacklabs. Causalnex: A toolkit for causal reasoning with bayesian networks. `https://github.com/quantumblacklabs/causalnex`. Accessed: 2025-5-8.

Nicholas Radcliffe. Using control groups to target on predicted lift: Building and assessing uplift model. 2007. URL `https://api.semanticscholar.org/CorpusID:22535399`.

Joseph D Ramsey, Kun Zhang, Madelyn Glymour, Ruben Sanchez Romero, Biwei Huang, Imme Ebert-Uphoff, Savini Samarasinghe, Elizabeth A Barnes, and Clark Glymour. Tetrad—a toolbox for causal discovery.

Hans Reichenbach. *The direction of time*. Dover Books on Physics. Dover Publications, Mineola, NY, July 1999.

PAUL R. ROSENBAUM and DONALD B. RUBIN. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 04 1983. ISSN 0006-3444. doi: 10.1093/biomet/70.1.41. URL `https://doi.org/10.1093/biomet/70.1.41`.

Yuta Saito and Shota Yasui. Counterfactual cross-validation: Stable model selection procedure for causal inference models. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8398–8407. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/saito20a.html`.

B Schäfer, C Grabow, S Auer, J Kurths, D Witthaut, and M Timme. Taming instabilities in power grid networks by decentralized control. *Eur. Phys. J. Spec. Top.*, 225(3):569–582, May 2016.

Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim M. Buhmann, and Walter Karlen. Learning counterfactual representations for estimating individual dose-response curves. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5612–5619, Apr. 2020. doi: 10.1609/aaai.v34i04.6014. URL `https://ojs.aaai.org/index.php/AAAI/article/view/6014`.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, May 2021. ISSN 1558-2256. doi: 10.1109/JPROC.2021.3058954.

Marco Scutari. Learning bayesian networks with the bnlearn r package. *Journal of Statistical Software*, 35(3):1–22, 2010. doi: 10.18637/jss.v035.i03. URL `https://www.jstatsoft.org/index.php/jss/article/view/v035i03`.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3076–3085. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/shalit17a.html`.

L. Shapley. *7. A Value for n-Person Games. Contributions to the Theory of Games II (1953) 307-317.*, pages 69–79. Princeton University Press, Princeton, 1997. ISBN 9781400829156. doi: doi:10.1515/9781400829156-012. URL `https://doi.org/10.1515/9781400829156-012`.

Amit Sharma and Emre Kiciman. Dowhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/8fb5f8be2aa9d6c64a04e3ab9f63feee-Paper.pdf`.

Yishai Shimoni, Chen Yanover, Ehud Karavani, and Yaara Goldschmnidt. Benchmarking framework for performance-evaluation of causal inference analysis, 2018. URL https://arxiv.org/abs/1802.05046.

Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, page 1219–1226. AAAI Press, 2006. ISBN 9781577352815.

Benjamin Speich, Belinda von Niederhäusern, Nadine Schur, Lars G Hemkens, Thomas Fürst, Neera Bhatnagar, Reem Alturki, Arnav Agarwal, Benjamin Kasenda, Christiane Pauli-Magnus, Matthias Schwenkglenks, and Matthias Briel. Systematic review on costs and resource use of randomized clinical trials shows a lack of transparent and comprehensive data. *J. Clin. Epidemiol.*, 96:1–11, April 2018.

Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72, 1991. doi: 10.1177/089443939100900106. URL https://doi.org/10.1177/089443939100900106.

Jerzy Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472, 1990. ISSN 08834237, 21688745. URL http://www.jstor.org/stable/2245382.

Johannes Textor, Juliane Hardt, and Sven Knüppel. DAGitty. *Epidemiology*, 22(5):745, September 2011.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Eighteenth National Conference on Artificial Intelligence*, page 567–573, USA, 2002. American Association for Artificial Intelligence. ISBN 0262511290.

Santtu Tikka and Juha Karvanen. Identifying causal effects with the r package causaleffect. *Journal of Statistical Software*, 76(12):1–30, 2017. doi: 10.18637/jss.v076.i12. URL https://www.jstatsoft.org/index.php/jss/article/view/v076i12.

Santtu Tikka, Antti Hyttinen, and Juha Karvanen. Causal effect identification from multiple incomplete data sources: A general search-based approach. *Journal of Statistical Software*, 99 (5):1–40, 2021. doi: 10.18637/jss.v099.i05. URL https://www.jstatsoft.org/index.php/jss/article/view/v099i05.

L. G. Valiant. A theory of the learnable. In *Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing*, STOC '84, page 436–445, New York, NY, USA, 1984. Association for Computing Machinery. ISBN 0897911334. doi: 10.1145/800057.808710. URL https://doi.org/10.1145/800057.808710.

Xin Wang, Shengfei Lyu, Xingyu Wu, Tianhao Wu, and Huanhuan Chen. Generalization bounds for estimating causal effects of continuous treatments. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 8605–8617. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/390bb66a088d37f62ee9fb779c5953c2-Paper-Conference.pdf.

Yixin Wang and David M. Blei and. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019. doi: 10.1080/01621459.2019.1686987. URL https://doi.org/10.1080/01621459.2019.1686987.

F Perry Wilson, Melissa Martin, Yu Yamamoto, Caitlin Partridge, Erica Moreira, Tanima Arora, Aditya Biswas, Harold Feldman, Amit X Garg, Jason H Greenberg, Monique Hinchcliff, Stephen Latham, Fan Li, Haiqun Lin, Sherry G Mansour, Dennis G Moledina, Paul M Palevsky, Chirag R Parikh, Michael Simonov, Jeffrey Testani, and Ugochukwu Ugwuowo. Electronic health record alerts for acute kidney injury: multicenter, randomized clinical trial. *BMJ*, 372:m4786, January 2021.

F Perry Wilson, Yu Yamamoto, Melissa Martin, Claudia Coronel-Moreno, Fan Li, Chao Cheng, Abinet Aklilu, Lama Ghazi, Jason H Greenberg, Stephen Latham, Hannah Melchinger, Sherry G Mansour, Dennis G Moledina, Chirag R Parikh, Caitlin Partridge, Jeffrey M Testani, and Ugochukwu Ugwuowo. A randomized clinical trial assessing the effect of automated medication-targeted alerts on acute kidney injury outcomes. *Nat. Commun.*, 14(1):2826, May 2023.

Kevin Xia, Kai-Zhan Lee, Yoshua Bengio, and Elias Bareinboim. The causal-neural connection: Expressiveness, learnability, and inference. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 10823–10836. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/5989add1703e4b0480f75e2390739f34-Paper.pdf.

Hedong YAN. Awesome causality identification. https://github.com/hedongyan/awesome-causality-identification, 2025. Version 0.x.

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/a50abba8132a77191791390c3eb19fe7-Paper.pdf.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ByKWUeWA-.

Jianfeng Zhan, Lei Wang, Wanling Gao, Hongxiao Li, Chenxi Wang, Yunyou Huang, Yatao Li, Zhengxin Yang, Guoxin Kang, Chunjie Luo, Hainan Ye, Shaopeng Dai, and Zhifei Zhang. Evaluatology: The science and engineering of evaluation. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 4(1):100162, 2024. ISSN 2772-4859. doi: https://doi.org/10.1016/j.tbench.2024.100162. URL https://www.sciencedirect.com/science/article/pii/S2772485924000140.

Keli Zhang, Shengyu Zhu, Marcus Kalander, Ignavier Ng, Junjian Ye, Zhitang Chen, and Lujia Pan. gcastle: A python toolbox for causal discovery, 2021.

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, page 647–655, Arlington, Virginia, USA, 2009. AUAI Press. ISBN 9780974903958.

Yang Zhao and Qing Liu. Causal ml: Python package for causal inference machine learning. *SoftwareX*, 21:101294, 2023. ISSN 2352-7110. doi: https://doi.org/10.1016/j.softx.2022.101294. URL https://www.sciencedirect.com/science/article/pii/S2352711022002126.

Yujia Zheng, Biwei Huang, Wei Chen, Joseph Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. Causal-learn: Causal discovery in python. *Journal of Machine Learning Research*, 25(60):1–8, 2024.

Hao Zhou, Shaoming Li, Guibin Jiang, Jiaqi Zheng, and Dong Wang. Direct heterogeneous causal learning for resource allocation problems in marketing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):5446–5454, Jun. 2023. doi: 10.1609/aaai.v37i4.25677. URL https://ojs.aaai.org/index.php/AAAI/article/view/25677.

Hao Zhou, Rongxiao Huang, Shaoming Li, Guibin Jiang, Jiaqi Zheng, Bing Cheng, and Wei Lin. Decision focused causal learning for direct counterfactual marketing optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6368–6379, New York, NY, USA, 2024a. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3672353. URL https://doi.org/10.1145/3637528.3672353.

Wenru Zhou, Miranda Kroehl, Maxene Meier, and Alexander Kaizer. An automated platform trial framework for a/b testing. *Contemporary Clinical Trials Communications*, 42:101388, 2024b. ISSN 2451-8654. doi: https://doi.org/10.1016/j.conctc.2024.101388. URL https://www.sciencedirect.com/science/article/pii/S2451865424001352.

Minqin Zhu, Anpeng Wu, Haoxuan Li, Ruoxuan Xiong, Bo Li, Xiaoqing Yang, Xuan Qin, Peng Zhen, Jiecheng Guo, Fei Wu, and Kun Kuang. Contrastive balancing representation learning for heterogeneous dose-response curves estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(15):17175–17183, Mar. 2024. doi: 10.1609/aaai.v38i15.29663. URL https://ojs.aaai.org/index.php/AAAI/article/view/29663.

A Investigation of individual medical AI in randomized controlled trial

From January 2022 and January 2023, there are total 380 completed phase III interventional trials (adult) with results and study documents on ClinicalTrials.gov. The statistical estimates can be categorized into six classes: geometric mean, mean, square mean, median, rank, and percentage, as shown in the figure 5. It is important to note that these estimates can not reflect the probability of a group or an individual benefiting from the treatment. For example, for treatment effect with Gaussian distribution, the mean difference can still be significant even the treatment is harmful for 49% and benefited for 51% if n is large enough. The randomized controlled trails can not be applied to development of individual treatment AI directly.

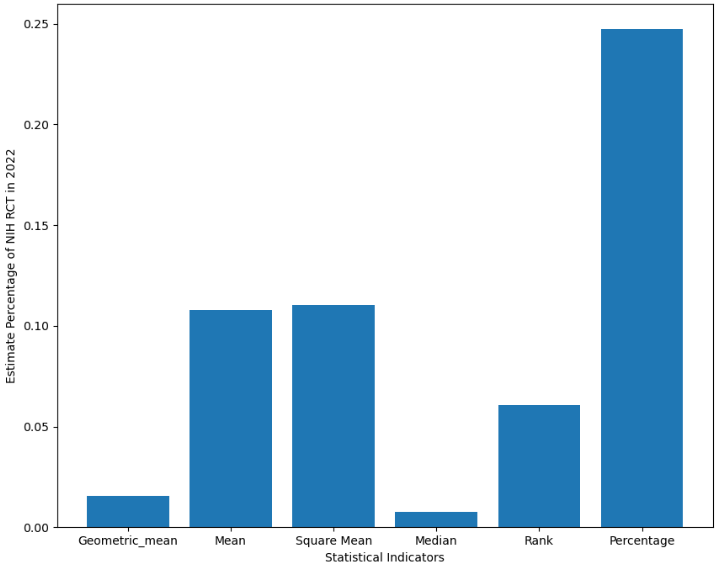


Figure 5: Estimates percentage of completed phase III interventional trials (adult) with results and study documents on https://clinicaltrials.gov.

Another query was conducted on ClinicalTrials.gov using the keyword **"Artificial Intelligence"** under the *Condition or Disease* field. Filters were applied to select only **completed**, **interventional studies**, **with posted results**, and **available statistical analysis plans**, with **no restriction on completion date** at January of 2025.

Only a total of 13 studies were retrieved. After manually excluding *single-arm trials* (i.e., studies without control groups or randomized design), 7 randomized clinical trials (RCTs) (Nayak et al. [2023], Abramoff et al. [2023], Lv et al. [2023], MacNeill et al. [2024], Eng et al. [2021], Piette et al. [2022], Mohr et al. [2019]) remained that investigated the clinical efficacy of AI-based healthcare interventions.

Among these, only 3 trials reported statistically significant primary outcomes, including:

- An AI-powered blood glucose reminder system (Nayak et al. [2023]),
- The Wysa AI chatbot for mental health support (MacNeill et al. [2024]),
- An AI system for bone age prediction (Eng et al. [2021]).

These results highlight the *challenge of translating AI agent systems into demonstrably effective clinical interventions*, emphasizing the need for rigorous evaluation, robust study design, and careful integration into complex healthcare workflows.

B Investigation of agent evaluation

B.1 Existed agent evaluation works

The theoretical foundations of agent evaluation, particularly with causal guarantees, remain underdeveloped. To date, we are not aware of any systematic theory that rigorously supports causal inference for agent performance evaluation. While several benchmarking efforts exist, they primarily focus on performance measurement rather than causal understanding. MLAgentBench (Huang et al. [2024]) uses 13 tasks (including Canonical Tasks, Classic Kaggle, Kaggle Challenges, Recent Research, and Code Improvement) to evaluate the performance of large language model. SUPER Benchmark (Bogin et al. [2024]) is used to evaluate the agents' ability to reproduce results from research repositories. AI Clinician (Komorowski et al. [2018]) and AI Hospital (Fan et al. [2025]) use synthetic data to test the performance of clinical agents. Jia et al. [2024] explores the LLMs' output decisions' alignment with human norms and ethical expectations. However, a common limitation across these works is the lack of control over unobserved confounders, such as participants' pre-experiment training or researchers' design biases. These hidden common causes may simultaneously influence both the deployed agent and its measured performance in real-world settings. As a result, these benchmarks cannot identify the (conditional) causal effect of an agent on real-world outcomes—such as clinical mortality—limiting their utility for causal evaluation.

B.2 Causal mini agent evaluation and potential outcome

Despite targeting different problem settings, the causal computational evaluation framework and the potential outcomes framework (Imbens and Rubin [2015]) for agent evaluation share one key connection and exhibit two fundamental differences. The connection lies in the fact that the unconfoundedness (ignorability/exchangeability) assumption in potential outcomes can be directly implied by the IRIS assumption. Furthermore, IRIS can be satisfied in practice by deploying mini agents in a randomized manner. The first major difference is that the Stable Unit Treatment Value Assumption (no interference and consistency)—a common requirement in potential outcome formulations—is not necessary under our framework, as long as the IIDE (Independent and Identically Distributed Evaluations) assumption holds. The second key difference is that the positivity assumption, which typically requires a non-zero probability of receiving each treatment, is also not required. Nevertheless, the causal effect of mini agents on evaluation metrics can still be bounded in the limit, as formally established in Theorem 4. These distinctions suggest that the proposed framework can be viewed as a generalized and more flexible extension of the potential outcomes framework, tailored specifically for the causal evaluation of mini agent behaviors.

B.3 Existed causal inference benchmark and packages

Although a number of benchmark datasets and open-source packages have been developed to address general causal inference problems—such as CausalML (Zhao and Liu [2023]), EconML (Battocchi et al. [2019]), DoWhy (Blöbaum et al. [2024], Sharma and Kiciman [2020]), CauseBox (paras2612), CausalNex (quantumblacklabs), Causal Curve (Kobrosly [2020]), CausalDiscovery (Kalainathan et al. [2020]), pcalg (Kalisch et al. [2012]), bnlearn (Scutari [2010]), TETRAD (Ramsey et al.), CausalityBenchmark (Shimoni et al. [2018]), JustCause (inovex), CausalEffect (Tikka and Karvanen [2017]), Ananke (Lee et al. [2023]), Dagitty (Textor et al. [2011]), YLearn (Bochen Lyu [2022]), CausalImpact (Google), causal-learn (Zheng et al. [2024]), gCastle (Zhang et al. [2021]), dosearch (Tikka et al. [2021]), CEE (Li), awesome-causality-identification (YAN [2025]) —these tools are primarily designed for causal discovery or treatment effect estimation. Despite extensive investigation, we did not identify any existing algorithms or frameworks that directly address the problem of causal evaluation of agents, particularly in the context of evaluating agent-driven decision-making processes.

C Proofs

Theorem 1. ***Upper bound***. *Given any evaluation model $\hat{f}$, $P(E_{gen}(\hat{f}) < E_{emp}(\hat{f}) + \sqrt{\frac{1}{2n}\ln(\frac{1}{\sigma})}) \geq 1-\sigma$ where n is number of independent identical distributed error (IIDE) measurements, $0 < 1-\sigma < 1$ is confidence.*

Proof. Denote $\epsilon_0 \cong \sqrt{\frac{1}{2n}\ln(\frac{1}{\sigma})}$, inequality in theorem 1 is equivalent to

$$P(E_{gen}(\hat{f}) - E_{emp}(\hat{f}) < \epsilon_0) \geq 1-\sigma$$

. Because event $E_{gen}(\hat{f}) - E_{emp}(\hat{f}) < \epsilon_0$ and event $E_{gen}(\hat{f}) - E_{emp}(\hat{f}) \geq \epsilon_0$ are complementary events, so

$$P(E_{gen}(\hat{f}) - E_{emp}(\hat{f}) < \epsilon_0) = 1 - P(E_{gen}(\hat{f}) - E_{emp}(\hat{f}) \geq \epsilon_0)$$

. Bring it in, we have

$$1 - P(E_{gen}(\hat{f}) - E_{emp}(\hat{f}) \geq \epsilon_0) \geq 1-\sigma$$

. Then it can be simplified as

$$P(E_{gen}(\hat{f}) - E_{emp}(\hat{f}) \geq \epsilon_0) \leq \sigma$$

. Because evaluation error is distributed on $[0,1]$, and Hoeffding inequality (Hoeffding [1963]),

$$\begin{aligned} &P(E_{gen}(\hat{f}) - E_{emp}(\hat{f}) \geq \epsilon_0) \\ &\leq 2e^{-\frac{2\epsilon_0^2 n^2}{\Sigma_{i=1}^{n}(1-0)^2}} = 2e^{-\frac{2(\sqrt{\frac{1}{2n}(\ln(\frac{1}{\sigma}))})^2 n^2}{\Sigma_{i=1}^{n}(1-0)^2}} = \sigma \end{aligned}$$

□

Theorem 2. ***Strict causal advantage***. *Given $\{c_i, s_i, m_i\}_{i=1}^{n}$, where c is evaluation condition, s is independent from c, and e, and s is independently, randomly, identically sampled (IRIS) from distribution $Pr(S)$, m is generate by an unknown and inaccessible function $m = f(c,e,s)$, e is other unlisted random variable, evaluation error function L is mean square error, $\forall \hat{f}_1$, $\forall \hat{f}_2$ where $\hat{f}_1$ and $\hat{f}_2$ are unbiased estimation of f given c (CU), we have* ***Efficiency****: if $E_{gen}(\hat{f}_1) < E_{gen}(\hat{f}_2)$ for any s, then $\forall s_a \in \mathbf{S}$, $\forall s_b \in \mathbf{S}$, $\forall c \in \mathbf{C}$, $E_{gen}(\hat{f}_1(c,s_a) - \hat{f}_1(c,s_b)) < E_{gen}(\hat{f}_2(c,s_a) - \hat{f}_2(c,s_b))$;* ***Consistency****: if $\lim_{\hat{f}\to f} E_{gen}(\hat{f}) = 0$, then $\forall s_a \in \mathbf{S}$, $\forall s_b \in \mathbf{S}$, $\forall c \in \mathbf{C}$, $\lim_{\hat{f}\to f} E_{gen}(\hat{f}(c,s_a) - \hat{f}(c,s_b)) = 0$.*

Proof. $\forall k = 1,2$, denote $\epsilon_k(c,s,e) = \hat{f}_k(c,s) - f(c,e,s)$, according to conditional unbiasness, we have $E(\epsilon_k(c,s)) = 0$. The error is denoted as

$$\Delta\hat{f}_k(s) = \hat{f}_k(c,s_a) - \hat{f}_k(c,s_b) - (f(c,s_a,e) - f(c,s_b,e)) = \epsilon_k(c,s_a,e) - \epsilon_k(c,s_b,e)$$

, the square error is denoted as

$$(\Delta\hat{f}_k)^2 = (\epsilon_k(c,s_a,e) - \epsilon_k(c,s_b,e))^2$$

, the mean square error is denoted as

$$E((\Delta\hat{f}_k)^2) = E(\epsilon_k(c,s_a,e)^2) + E(\epsilon_k(c,s_b,e)^2) - 2E(\epsilon_k(c,s_a,e)\epsilon_k(c,s_b,e))$$

. According to the IRIS assumption and condition unbiasness assumption, the third term

$$\begin{aligned} E(\epsilon_k(c,s_a)\epsilon_k(c,s_b)) &= E_c(E(\epsilon_k(c,s_a,e)\epsilon_k(c,s_b,e)|c)) \\ &= E_c(E(\epsilon(c,s_a)|c)E(\epsilon(c,s_b)|c)) = E_c(0*0) = E_c(0) = 0 \end{aligned}$$

, then because $\forall s$, $E_{gen}(\hat{f}_1) < E_{gen}(\hat{f}_2)$, that is $E(\epsilon_1(c,s,e)^2) < E(\epsilon_2(c,s,e)^2)$ for any s, then

$$\begin{aligned} E((\Delta\hat{f}_1)^2) &= E(\epsilon_1(c,s_a,e)^2) + E(\epsilon_1(c,s_b,e)^2) \\ &< E(\epsilon_2(c,s_a,e)^2) + E(\epsilon_2(c,s_b,e)^2) = E((\Delta\hat{f}_2)^2) \end{aligned}$$

, which can be write as $E_{gen}(\hat{f}_1(c,s_a) - \hat{f}_1(c,s_b)) < E_{gen}(\hat{f}_2(c,s_a) - \hat{f}_2(c,s_b))$ when error function is mean square error.

Following, the consistency can be proved spontaneously by the Squeeze Theorem in Calculus. □

Theorem 3. ***Causal bound with positivity****. $\forall s_a \in S, \forall s_b \in S$, given any unbiased evaluation model $\hat{f}(c, s_a)$ and $\hat{f}(c, s_b)$,*

$$P(E_{gen}(\Delta \hat{f}) < 2 \max\{E_{emp}(a) + \sqrt{\frac{1}{2n_a} \ln(\frac{2}{\sigma})}), E_{emp}(b) + \sqrt{\frac{1}{2n_b} \ln(\frac{2}{\sigma})}\}) \geq 1 - \sigma$$

, where $\Delta \hat{f} = \hat{f}(c, s_a) - \hat{f}(c, s_b)$, n_a and n_b are number of independently, randomly, identically sampled error measurements (IIDE) where $s = s_a$ and $s = s_b$, $0 < 1 - \sigma < 1$ is confidence, evaluation error function L is mean square error ranged from 0 to 1.

Proof. Because evaluation error function is mean square error,

$$E_{gen}(\Delta \hat{f}) \cong E((\Delta \hat{f} - \Delta f)^2) = E((\hat{f}(c, s_a) - \hat{f}(c, s_b) - (f(c, s_a, e) - f(c, s_b, e)))^2)$$

. Denote $\epsilon_a(c, e) \cong \hat{f}(c, s_a) - f(c, s_a, e)$ and $\epsilon_b(c, e) \cong \hat{f}(c, s_b) - f(c, s_b, e)$,

$$E_{gen}(\Delta \hat{f}) = E(\epsilon_a(c, e)^2 + \epsilon_b(c, e)^2 - 2\epsilon_a(c, e)\epsilon_b(c, e))$$

. Due to properties of expectation,

$$E_{gen}(\Delta \hat{f}) = E(\epsilon_a(c, e)^2) + E(\epsilon_b(c, e)^2) - 2E(\epsilon_a(c, e)\epsilon_b(c, e))$$

. Due to IIDE assumption, the sub population of errors where $s = a$ and $s = b$ is also independently sampled,

$$E_{gen}(\Delta \hat{f}) = E(\epsilon_a(c, e)^2) + E(\epsilon_b(c, e)^2) - 2E(\epsilon_a(c, e))E(\epsilon_b(c, e))$$

. Due to the unbiased assumption,

$$E_{gen}(\Delta \hat{f}) = E(\epsilon_a(c, e)^2) + E(\epsilon_b(c, e)^2)$$

. Denote $\epsilon_0 = 2 \max\{\nu_a, \nu_b\}$ where $\nu_a = E_{emp}(a) + \sqrt{\frac{1}{2n_a} \ln(\frac{2}{\sigma})}$ and $\nu_b = E_{emp}(b) + \sqrt{\frac{1}{2n_b} \ln(\frac{2}{\sigma})}$, due to Bonferroni inequality (Galambos [1977]),

$$P(E_{gen}(\Delta \hat{f}) < \epsilon_0) = P(E(\epsilon_a(c, e)^2) + E(\epsilon_b(c, e)^2) < \epsilon_0)$$

$$\geq P(E(\epsilon_a(c, e)^2) < \frac{\epsilon_0}{2} \bigcap E(\epsilon_b(c, e)^2) < \frac{\epsilon_0}{2})$$

$$\geq P(E(\epsilon_a(c, e)^2) < \frac{\epsilon_0}{2}) + P(E(\epsilon_b(c, e)^2) < \frac{\epsilon_0}{2}) - 1$$

$$\geq P(E(\epsilon_a(c, e)^2) < \nu_a) + P(E(\epsilon_b(c, e)^2) < \nu_b) - 1$$

$$\geq 1 - \frac{\sigma}{2} + 1 - \frac{\sigma}{2} - 1$$

$$= 1 - \sigma$$

□

Theorem 4. ***Causal bound with non-positivity****. Given unbiased evaluation model $\hat{f}$, then $\forall s_a \in S, \forall s_b \in S$,*

$$P(E_{gen}(\Delta \hat{f}) < 2(E_{emp}(\hat{f}) + \sqrt{\frac{1}{2n} \ln(\frac{1}{\sigma})})) \geq 1 - \sigma$$

, where $\Delta \hat{f} = \hat{f}(c, s_a) - \hat{f}(c, s_b)$, n is number of independently, randomly, identically sampled error measurements (IIDE), $0 < 1 - \sigma < 1$ is confidence, s is independently, randomly, identically sampled (IRIS), evaluation error function L is mean square error ranged from 0 to 1.

Proof. According to same proof and denotation in theorem 3,

$$E_{gen}(\Delta\hat{f}) = E(\epsilon_a(c,e)^2) + E(\epsilon_b(c,e)^2) - 2E(\epsilon_a(c,e))E(\epsilon_b(c,e))$$

. Due to IIDE and IRIS assumption, the sub population of errors where $s = a$ and $s = b$ is also independently sampled (IIDE) identically following the whole population (IRIS),

$$E_{gen}(\Delta\hat{f}) = 2E(\epsilon(c,e)^2) - 2E(\epsilon(c,e))^2$$

. Due to the unbiased assumption,

$$E_{gen}(\Delta\hat{f}) = 2E(\epsilon(c,e)^2)$$

. According to theorem 1, let evaluation error function be MSE,

$$P(E(\epsilon(c,e)^2) \leq \sum_{i=1}^{n}(\hat{f}(c_i,s_i) - m_i)^2 + \sqrt{\frac{1}{2n}\ln(\frac{1}{\sigma})}) \geq 1 - \sigma$$

where m_i is true metric. Due to the property of equivalent events,

$$P(2E(\epsilon(c,e)^2) \leq 2(\sum_{i=1}^{n}(\hat{f}(c_i,s_i) - m_i)^2 + \sqrt{\frac{1}{2n}\ln(\frac{1}{\sigma})})) \geq 1 - \sigma$$

, which can be rewrite as

$$P(E_{gen}(\Delta\hat{f}) \leq 2(E_{emp}(\hat{f}) + \sqrt{\frac{1}{2n}\ln(\frac{1}{\sigma})})) \geq 1 - \sigma$$

. □

D Evaluation scenes

We introduce 12 scenes to test the performance of our (conditional) evaluation model: Alert (Wilson et al. [2021]), Withdraw (Wilson et al. [2023]), Grid-S (Schäfer et al. [2016]), Higgs-S (Baldi et al. [2014]), Insurance (Kumar [2023]), Climate-S (Lucas et al. [2013]), TERECO (Li et al. [2022]), Calculus (Kramer et al. [2023]), Ad (Gokagglers [2018]), kin8nm-S (Fischer [2022]), NuScale-S (Huu Tiep [2024]), and Quantum trade. The evaluation subject (agent) and evaluation metric in those applied scenes is listed in the table 4. All the data are collected from real experiments or simulation. As mentioned in the section 6, we use the prediction performance in randomized experiment to approximate the post-experiment performance of randomized deployed mini agents. We also listed the website and license of the mentioned asset in the following table 3.

Scenes	Website	License
Alert	https://datadryad.org/dataset/doi:10.5061/dryad.4f4qrfj95	CC0 1.0
Withdraw	https://datadryad.org/dataset/doi:10.5061/dryad.kh189327p	CC0 1.0
Grid-S	https://www.openml.org/search?type=data&status=active&id=44973	CC BY 4.0
Higgs	https://www.openml.org/search?type=data&status=active&id=42769	Public
Insurance	https://www.kaggle.com/datasets/arun0309/customer	Unknown
Climate	https://www.openml.org/search?type=data&status=active&id=40994	Public
TERECO	https://datadryad.org/dataset/doi:10.5061/dryad.59zw3r27n	CC0 1.0
Calculus	https://datadryad.org/dataset/doi:10.5061/dryad.kkwh70s95	CC0 1.0
Ad	https://www.kaggle.com/datasets/loveall/clicks-conversion-tracking	Other
kin8nm	https://www.openml.org/search?type=data&status=active&id=44980	Public
NuScale	https://archive.ics.uci.edu/dataset/1091	CC BY 4.0

Table 3: Website and license of the mentioned asset.

The heterogeneous mini agent space is configured as a combination of a linear mini agent space (logistic regression or linear regression) and a non-linear mini agent space (MLP Classifier or MLP Regressor). The agent's input dimension is the number of features, and output dimension is the number of the targets. The MLP's hidden layer size limit is set as 8.

Scene	Evaluation Subject (Agent)	Application	Metric
Individual Treatment	input is patients' record, and random pop-up, output is mortality	intelligent Alert, and Withdrawal for AKI	ROC-AUC ACC
	input is patients' record, and random therapy, output is walk distance	Covid-19 recovery (TERECO)	RMSE R^2
Scientific Simulation	input is setting of fission reactor, output is k-inf and PPPF	nuclear power plant (Nuscale)	RMSE R^2
	input is particle features output is whether it is higgs	higgs detection (Higgs)	ROC-AUC ACC
	input is initial parameter of climate, output is whether the climate crash	climate simulation (Climate)	ROC-AUC ACC
	input is joint moving of robotic arm, output is distance to target	robot action (kin8nm)	RMSE R^2
	input is parameter of electric grid, output is whether the grid crash	grid simulation (Grid)	ROC-AUC ACC
Social Experiment	input is students' record, and random class output is students' grading	individual teaching (Calculus)	RMSE R^2
Business	input is customers' record, and random recommendation output is buy-in decision of custom	sell insurance (Insurance)	ROC-AUC ACC
	input is feature of a group, and random recommendation output is buy-in decision of custom	advertisement (Ad)	RMSE R^2
Quantum Trade	input is stock's feature last day, output is decision for the stock	A-share trade	RoI

Table 4: Detailed setting of scene.

The IRIS sampling of the agent is two steps. The first step is the agent type sampling with 0.5 probability to choose linear model and 0.5 probability to choose MLP model. The second step is to choose a specific model in the sub space. For both linear agent and MLP agent, we use Normal distribution $N(0, 1)$ to determine the parameter of the models. The outcome when we calculate Shapley value is set as the negative evaluation errors.

The samples number of scenes Alert, Withdrawal, Higgs, Climate, Grid, Insurance, TERECO, Nuscale, kin8nm, Calculus, Ad, and A-share trade are listed in table 5. The 7 kinds of base learners we used is Linear, MLP, SVM/SVR, RF (Breiman [2001]), LGBM (Aziz et al. [2022]), XGBoost (Chen and Guestrin [2016]), and CatBoost (Prokhorenkova et al. [2018]). All the computation is on the one 13-inch MacBook Pro 2020 with Apple M1 chip and 16GB memory in 2 days.

Scenes	Train Samples	Test Samples
Alert	$1600 * 5911 * 0.2 = 1891520$	$400 * 5911 * 0.2 = 472880$
Withdrawal	$1600 * 4998 * 0.2 = 1599360$	$400 * 4998 * 0.2 = 399840$
Higgs	$1600 * 2000 * 0.2 = 640000$	$400 * 2000 * 0.2 = 160000$
Climate	$1600 * 540 * 0.2 = 172800$	$400 * 540 * 0.2 = 43200$
Grid	$1600 * 10000 * 0.2 = 3200000$	$400 * 10000 * 0.2 = 800000$
Insurance	$1600 * 45211 * 0.2 = 14467520$	$400 * 45211 * 0.2 = 3616880$
TERECO	$1600 * 104 * 0.2 = 33280$	$400 * 104 * 0.2 = 8320$
Nuscale	$1600 * 360 = 576000$	$400 * 360 = 144000$
kin8nm	$1600 * 8192 * 0.2 = 2621440$	$400 * 8192 * 0.2 = 655360$
Calculus	$1600 * 672 * 0.2 = 215040$	$400 * 672 * 0.2 = 53760$
Ad	$1600 * 1143 * 0.2 = 365760$	$400 * 1143 * 0.2 = 91440$
Trade	$160 * 200 * 30 = 960000$	$40 * 200 * 20 = 160000$

Table 5: Evaluation sample numbers in different scenes.

D.1 Scenes to test evaluation model

For 11 scenes to test evaluation model, 2000 agents was randomly sampling from heterogeneous agent space. 20% data was used to build the real systems to generate true evaluation metrics and other 80% data was used to get the proxy metrics of a agent. When the evaluation samples are collected from the built real systems, 20% samples are used for evaluation model testing, and other 80% samples are for evaluation model learning. The experiment is performed 30 times. All the null value was removed from the raw data in preprocessing. The classical features are encoded by the label encoder to reduce the computation. The text classical features are encoded by one-hot encoder.

The proxy metrics we used for Alert, Withdrawal, Higgs, Climate, Grid, Insurance are ROC-AUC, accuracy, recall, precision, F_1 score, and PR-AUC. The proxy metrics we used for TERECO, Nuscale, kin8nm, Calculus, Ad are RMSE, R^2, MAE, MAPE, and MSE.

D.2 Scenes to test conditional evaluation model

We collected the A-share data of China from 2024-8 to 2024-11 (89 days) by an open source tool AkShare. The collected day-wise features of stocks are “Profit Ratio”, “Average Cost”, “90 Cost Low”, “90 Cost High”, “90 Episode Medium”, “70 Cost Low”, “70 Cost High”, “70 Episode Medium”, “Closing”, “Opening”, “Highest”, “Lowest”, “Transaction Volume”, “Transaction Amount”, “Amplitude”. The strategy is to buy 1 hand, sell 1 hand, or hold by the agent following the open price of the market. The slippage and commission is not considered in this work. If the stock was limit up (limit down), then we can only hold even the model decision is to buy (sell). The agent’s input is the feature of the stock last day and output is the decision this day. The agent will be deployed and run for 1 time slot (10 trading days, about 2 weeks), and the evaluation condition is feature of the stock in the last time slot. At the end of deployment in a time slot, we will sell all the hold stocks. When the end of the slot is limit down at open, we assume that it will be sold by 90 percents of open price in future.

In order to simulate the direct effect from the agent itself to the sensitive market, we add a float model which input is the vectorized agents and output is float rate of open price ranged from -0.1% to +0.1%.

We use data of 200 stocks with minimal ID from all 3214 stocks due to our memory limit. The start of train day is from day 10 to day 41, the start of test day is from 41 to 61. The sampling number of agent is 200. The experiment was performed 5 times to reduce the influence of randomness. The proxy metric is the last 10 days’ RoI.

E Assumption assessment

Due to the limitation of computation resources and the difference of addressed problems, we do not use the deep learning models for observation data as base learners of our evaluation models, such as CEVAE (Louizos et al. [2017]), BNR (Li and Fu [2017]), BNN (Johansson et al. [2016]), SITE (Yao et al. [2018]), GANITE (Yoon et al. [2018]), DeepMatch (Kallus [2020]), Dragonnet (Shi et al. [2019]), TARNet/CFR (Shalit et al. [2017]), TabPFN (Hollmann et al. [2025]), deconfounder (Wang and and [2019]), time-series deconfounder (Bica et al. [2020]). Table 6 and table 7 are assumption assessment results of 7 kinds of evaluation models with swift base learners for continuous and discrete target respectively. From the tables, we can see that the IIDE assumption (IID check and ID check) and unbiased evaluation model assumption (Bias Check and GroupBias Check) was not violated obviously with a high probability (>0.8) except the cases that using MLP and SVR as base regression learners of HetEM.

Ratio $p >= 0.05$	TERECO	Calculus	kin8nm	nuscale	ad
Het(Linear)-IID-RMSE	0.93	0.87	0.97	0.9	0.97
Het(Linear)-IID-R^2	0.93	1.0	0.9	0.93	0.83
Het(Linear)-ID-RMSE	1.0	0.47	0.87	1.0	0.93
Het(Linear)-ID-R^2	0.97	0.43	0.93	0.97	0.87
Het(Linear)-Bias-RMSE	0.9	0.43	0.77	0.8	0.77
Het(Linear)-Bias-R^2	0.9	0.33	0.8	0.77	0.77
Het(Linear)-GroupBias-RMSE	0.92	0.85	0.87	0.87	0.85

Het(Linear)-GroupBias-R^2	0.90	0.85	0.86	0.84	0.89
Het(MLP)-IID-RMSE	1.0	0.93	0.83	1.0	0.9
Het(MLP)-IID-R^2	0.9	0.97	0.93	0.97	0.97
Het(MLP)-ID-RMSE	0.73	0.37	0.87	0.73	0.77
Het(MLP)-ID-R^2	0.9	0.3	0.77	0.77	0.67
Het(MLP)-Bias-RMSE	0.5	0.26	0.57	0.57	0.5
Het(MLP)-Bias-R^2	0.57	0.17	0.53	0.57	0.57
Het(MLP)-GroupBias-RMSE	0.68	0.40	0.86	0.70	0.72
Het(MLP)-GroupBias-R^2	0.70	0.31	0.71	0.67	0.68
Het(SVR)-IID-RMSE	0.97	0.97	0.97	1.0	0.87
Het(SVR)-IID-R^2	0.93	0.9	1.0	0.9	0.97
Het(SVR)-ID-RMSE	0.0	0.0	0.27	0.4	0.1
Het(SVR)-ID-R^2	0.0	0.0	0.0	0.2	0.0
Het(SVR)-Bias-RMSE	0.0	0.0	0.1	0.17	0.03
Het(SVR)-Bias-R^2	0.0	0.23	0.0	0.03	0.0
Het(SVR)-GroupBias-RMSE	0.0	0.01	0.19	0.34	0.11
Het(SVR)-GroupBias-R^2	0.0	0.44	0.00	0.19	0.02
Het(RF)-IID-RMSE	0.97	0.93	0.97	0.97	0.93
Het(RF)-IID-R^2	1.0	0.97	0.97	1.0	0.97
Het(RF)-ID-RMSE	0.93	0.87	0.9	0.97	0.83
Het(RF)-ID-R^2	0.93	0.93	0.93	0.9	0.73
Het(RF)-Bias-RMSE	0.93	0.57	0.9	0.9	0.8
Het(RF)-Bias-R^2	0.83	0.63	0.93	0.87	0.8
Het(RF)-GroupBias-RMSE	0.87	0.85	0.91	0.90	0.86
Het(RF)-GroupBias-R^2	0.89	0.86	0.91	0.88	0.86
Het(LGBM)-IID-RMSE	0.87	0.93	0.97	1.0	0.93
Het(LGBM)-IID-R^2	1.0	0.97	0.97	0.93	0.97
Het(LGBM)-ID-RMSE	0.93	0.9	0.9	0.97	0.83
Het(LGBM)-ID-R^2	0.9	0.93	0.83	0.97	0.93
Het(LGBM)-Bias-RMSE	0.9	0.7	0.8	0.9	0.77
Het(LGBM)-Bias-R^2	0.93	0.7	0.77	0.9	0.83
Het(LGBM)-GroupBias-RMSE	0.89	0.85	0.90	0.88	0.85
Het(LGBM)-GroupBias-R^2	0.89	0.86	0.97	0.88	0.87
Het(XGBoost)-IID-RMSE	1.0	0.93	0.93	1.0	0.87
Het(XGBoost)-IID-R^2	0.93	0.93	0.93	0.97	0.93
Het(XGBoost)-ID-RMSE	0.93	0.9	1.0	0.93	0.97
Het(XGBoost)-ID-R^2	0.93	0.9	0.93	0.93	0.87
Het(XGBoost)-Bias-RMSE	0.97	0.53	0.77	0.87	0.87
Het(XGBoost)-Bias-R^2	0.83	0.5	0.9	0.87	0.83
Het(XGBoost)-GroupBias-RMSE	0.92	0.84	0.83	0.86	0.88
Het(XGBoost)-GroupBias-R^2	0.87	0.82	0.90	0.84	0.88
Het(CatBoost)-IID-RMSE	0.87	0.97	0.97	1.0	1.0
Het(CatBoost)-IID-R^2	0.87	0.97	0.93	0.93	1.0
Het(CatBoost)-ID-RMSE	0.93	0.87	0.93	0.97	0.97
Het(CatBoost)-ID-R^2	0.97	0.9	0.93	0.97	1.0
Het(CatBoost)-Bias-RMSE	0.93	0.6	0.97	0.8	0.83
Het(CatBoost)-Bias-R^2	0.93	0.63	0.93	0.8	0.87
Het(CatBoost)-GroupBias-RMSE	0.90	0.90	0.92	0.87	0.87
Het(CatBoost)-GroupBias-R^2	0.90	0.87	0.90	0.86	0.88

Table 6: The ratio that the p-value of test is larger than or equal to 0.05 in the 30 repeated experiments for regression base learners. The name of each row is “EvaluationModelName”-“TestName”-“MetricName”. The name of each column is the scene name.

Ratio $p >= 0.05$	withdraw	higgs	grid	insurance	climate	alert
Het(Linear)-IID-ROCAUC	0.96	0.93	0.93	0.93	1.0	0.97
Het(Linear)-IID-ACC	0.93	0.97	0.9	0.9	0.97	0.97
Het(Linear)-ID-ROCAUC	1.0	0.97	0.87	0.9	1.0	0.97
Het(Linear)-ID-ACC	0.93	0.97	0.87	0.93	0.93	1.0
Het(Linear)-Bias-ROCAUC	0.77	0.83	0.83	0.9	0.9	0.9
Het(Linear)-Bias-ACC	0.83	0.87	0.87	0.8	0.83	0.93
Het(Linear)-GroupBias-ROCAUC	0.83	0.88	0.87	0.88	0.91	0.87
Het(Linear)-GroupBias-ACC	0.87	0.89	0.87	0.88	0.87	0.89
Het(MLP)-IID-ROCAUC	0.97	1.0	0.97	0.97	0.93	0.93
Het(MLP)-IID-ACC	0.93	0.93	0.97	0.9	0.97	0.97
Het(MLP)-ID-ROCAUC	0.8	0.87	0.77	0.87	0.87	0.7
Het(MLP)-ID-ACC	0.8	0.83	0.83	0.8	0.73	0.8
Het(MLP)-Bias-ROCAUC	0.43	0.57	0.53	0.53	0.63	0.47
Het(MLP)-Bias-ACC	0.67	0.6	0.57	0.57	0.43	0.6
Het(MLP)-GroupBias-ROCAUC	0.70	0.76	0.71	0.73	0.77	0.69
Het(MLP)-GroupBias-ACC	0.77	0.73	0.77	0.74	0.63	0.73
Het(SVM)-IID-ROCAUC	0.97	0.97	0.93	0.97	0.97	1.0
Het(SVM)-IID-ACC	0.97	0.97	0.97	0.97	1.0	1.0
Het(SVM)-ID-ROCAUC	0.73	0.5	0.9	0.6	0.83	0.8
Het(SVM)-ID-ACC	0.9	0.67	0.37	0.97	0.97	0.97
Het(SVM)-Bias-ROCAUC	0.4	0.27	0.1	0.37	0.53	0.47
Het(SVM)-Bias-ACC	0.97	0.37	0.43	0.93	0.8	0.93
Het(SVM)-GroupBias-ROCAUC	0.63	0.48	0.36	0.52	0.69	0.66
Het(SVM)-GroupBias-ACC	0.90	0.56	0.62	0.90	0.85	0.86
Het(RF)-IID-ROCAUC	0.97	0.97	0.9	0.97	0.97	0.97
Het(RF)-IID-ACC	0.93	0.97	0.87	0.93	0.9	0.9
Het(RF)-ID-ROCAUC	0.97	0.87	0.9	0.9	0.83	0.97
Het(RF)-ID-ACC	1.0	0.87	0.9	0.97	1.0	0.87
Het(RF)-Bias-ROCAUC	0.73	0.9	0.77	0.87	0.93	0.87
Het(RF)-Bias-ACC	0.9	0.93	0.8	0.97	0.8	0.7
Het(RF)-GroupBias-ROCAUC	0.85	0.87	0.84	0.9	0.91	0.86
Het(RF)-GroupBias-ACC	0.87	0.89	0.85	0.89	0.88	0.85
Het(LGBM)-IID-ROCAUC	1.0	0.97	0.97	0.97	0.87	1.0
Het(LGBM)-IID-ACC	0.93	0.97	0.93	0.9	1.0	0.9
Het(LGBM)-ID-ROCAUC	1.0	0.87	0.93	0.93	0.87	0.93
Het(LGBM)-ID-ACC	0.97	0.83	0.97	1.0	0.97	0.97
Het(LGBM)-Bias-ROCAUC	0.7	0.93	0.83	0.73	0.97	0.9
Het(LGBM)-Bias-ACC	0.83	0.9	0.77	0.83	0.8	0.8
Het(LGBM)-GroupBias-ROCAUC	0.81	0.88	0.85	0.89	0.91	0.86
Het(LGBM)-GroupBias-ACC	0.88	0.88	0.84	0.90	0.86	0.85
Het(XGBoost)-IID-ROCAUC	1.0	0.97	1.0	0.8	0.93	1.0
Het(XGBoost)-IID-ACC	0.93	1.0	0.97	1.0	0.93	0.87
Het(XGBoost)-ID-ROCAUC	0.9	0.83	0.9	0.9	0.87	1.0
Het(XGBoost)-ID-ACC	0.97	0.87	0.93	0.97	0.93	0.97
Het(XGBoost)-Bias-ROCAUC	0.83	0.97	0.83	0.9	0.83	0.83
Het(XGBoost)-Bias-ACC	0.87	0.93	0.9	0.83	0.8	0.87
Het(XGBoost)-GroupBias-ROCAUC	0.88	0.89	0.90	0.90	0.87	0.89
Het(XGBoost)-GroupBias-ACC	0.88	0.89	0.88	0.88	0.89	0.88
Het(CatBoost)-IID-ROCAUC	0.93	0.8	0.9	0.97	0.93	0.9
Het(CatBoost)-IID-ACC	1.0	0.9	1.0	0.87	1.0	0.97
Het(CatBoost)-ID-ROCAUC	0.97	0.93	0.93	0.93	0.87	0.97
Het(CatBoost)-ID-ACC	0.97	0.93	0.83	0.97	0.97	0.97
Het(CatBoost)-Bias-ROCAUC	0.83	0.87	0.8	0.8	0.9	0.87
Het(CatBoost)-Bias-ACC	0.87	0.9	0.93	0.87	0.8	0.8
Het(CatBoost)-GroupBias-ROCAUC	0.86	0.89	0.88	0.86	0.89	0.87

Het(CatBoost)-GroupBias-ACC	0.85	0.89	0.89	0.90	0.86	0.86

Table 7: The ratio that the p-value of test is larger than or equal to 0.05 in the 30 repeated experiments for classification base learners. The name of each row is “EvaluationModelName”-“TestName”-“MetricName”. The name of each column is the scene name.

We also visualization the Q-Q plot of the baseline evaluation model in figure 6 and the assumption assessment result of our evaluation model in table 8. From the visualization, the error distribution of baseline is close to normal distribution. Although the HetEM-Linear and HetEM-CatBoost does not pass our ID check and Bias check, the dramatically empirical error reduction of our evaluation model is still *believed* as valuable if upper bound existed. Of course it can be not used to calculate the upper bound of the two evaluation models by theorem 4 directly in this scene.

Ratio $p \geq 0.05$	A-Share Trade
Het(Linear)-IID	1.0
Het(Linear)-ID	0.0
Het(Linear)-Bias	0.0
Het(CatBoost)-IID	1.0
Het(CatBoost)-ID	0.0
Het(CatBoost)-Bias	0.0

Table 8: The ratio that p-value of test is larger than or equal to 0.05 in the 5 repeated experiments for A-share trade scene. The name of each row is “EvaluationModelName”-“TestName”-“MetricName”. The name of each column is the scene name.

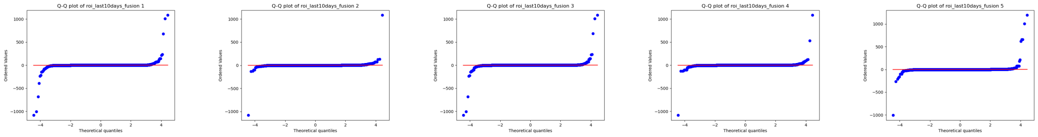


Figure 6: Q-Q plot of evaluation error of Baseline (Last10DayRoI).

F Hyper-parameters of evaluation model

There are 7 kinds of base learners for the evaluation model learning in our experiments. We use the default hyper-parameters for the 7 kinds of base learners in 6 scenes where evaluation metrics are RMSE and R^2 as following.

```
{
  "Linear Regression": {
    "copy_X": true,
    "fit_intercept": true,
    "positive": false
  },
  "MLP": {
    "activation": "relu",
    "alpha": 0.0001,
    "batch_size": "auto",
    "beta_1": 0.9,
    "beta_2": 0.999,
    "early_stopping": false,
    "epsilon": 1e-08,
    "hidden_layer_sizes": [
      100
    ],
    "learning_rate": "constant",
    "learning_rate_init": 0.001,
    "max_fun": 15000,
    "max_iter": 200,
    "momentum": 0.9,
    "n_iter_no_change": 10,
    "nesterovs_momentum": true,
    "power_t": 0.5,
    "shuffle": true,
    "solver": "adam",
    "tol": 0.0001,
    "validation_fraction": 0.1,
    "verbose": false,
    "warm_start": false
  },
  "SVM": {
    "C": 1.0,
    "cache_size": 200,
    "coef0": 0.0,
    "degree": 3,
    "epsilon": 0.1,
    "gamma": "scale",
    "kernel": "rbf",
    "max_iter": -1,
    "shrinking": true,
    "tol": 0.001,
    "verbose": false
  },
  "RF": {
    "bootstrap": true,
    "ccp_alpha": 0.0,
    "criterion": "squared_error",
    "max_features": 1.0,
    "min_impurity_decrease": 0.0,
    "min_samples_leaf": 1,
    "min_samples_split": 2,
    "min_weight_fraction_leaf": 0.0,
    "n_estimators": 100,
    "verbose": 0,
    "warm_start": false
  },
  "XGBoost": {
    "objective": "reg:squarederror",
    "missing": "nan",
    "n_estimators": 100
  },
  "LightGBM": {
    "boosting_type": "gbdt",
    "colsample_bytree": 1.0,
    "importance_type": "split",
    "learning_rate": 0.1,
    "max_depth": -1,
    "min_child_samples": 20,
    "min_child_weight": 0.001,
    "min_split_gain": 0.0,
    "n_estimators": 100,
    "num_leaves": 31,
    "reg_alpha": 0.0,
    "reg_lambda": 0.0,
    "subsample": 1.0,
    "subsample_for_bin": 200000,
    "subsample_freq": 0
  },
  "CatBoost": {
    "loss_function": "RMSE"
  }
}
```

The default hyper-parameters of the 7 kinds of base learners in 5 scenes where evaluation metrics are ROC-AUC and ACC as following.

```
{
  "LogisticRegression": {
    "C": 1.0,
    "dual": false,
    "fit_intercept": true,
    "intercept_scaling": 1,
    "max_iter": 100,
    "multi_class": "deprecated",
    "penalty": "l2",
    "solver": "lbfgs",
    "tol": 0.0001,
    "verbose": 0,
    "warm_start": false
  },
  "MLP": {
    "activation": "relu",
    "alpha": 0.0001,
    "batch_size": "auto",
    "beta_1": 0.9,
    "beta_2": 0.999,
    "early_stopping": false,
    "epsilon": 1e-08,
    "hidden_layer_sizes": [100],
    "learning_rate": "constant",
    "learning_rate_init": 0.001,
    "max_fun": 15000,
    "max_iter": 200,
    "momentum": 0.9,
    "n_iter_no_change": 10,
    "nesterovs_momentum": true,
    "power_t": 0.5,
    "shuffle": true,
    "solver": "adam",
    "tol": 0.0001,
    "validation_fraction": 0.1,
    "verbose": false,
    "warm_start": false
  },
  "SVM": {
    "C": 1.0,
    "break_ties": false,
    "cache_size": 200,
    "coef0": 0.0,
    "decision_function_shape": "ovr",
    "degree": 3,
    "gamma": "scale",
    "kernel": "rbf",
    "max_iter": -1,
    "probability": false,
    "shrinking": true,
    "tol": 0.001,
    "verbose": false
  },
  "RandomForest": {
    "bootstrap": true,
    "ccp_alpha": 0.0,
    "criterion": "gini",
    "max_features": "sqrt",
    "min_impurity_decrease": 0.0,
    "min_samples_leaf": 1,
    "min_samples_split": 2,
    "min_weight_fraction_leaf": 0.0,
    "n_estimators": 100,
    "oob_score": false,
    "verbose": 0,
    "warm_start": false
  },
  "XGBoost": {
    "objective": "binary:logistic",
    "enable_categorical": false,
    "missing": "NaN"
  },
  "LightGBM": {
    "boosting_type": "gbdt",
    "colsample_bytree": 1.0,
    "importance_type": "split",
    "learning_rate": 0.1,
    "max_depth": -1,
    "min_child_samples": 20,
    "min_child_weight": 0.001,
    "min_split_gain": 0.0,
    "n_estimators": 100,
    "num_leaves": 31,
    "reg_alpha": 0.0,
    "reg_lambda": 0.0,
    "subsample": 1.0,
    "subsample_for_bin": 200000,
    "subsample_freq": 0
  },
  "CatBoost": {}
}
```

We also use the Grid search for the hyperparameter searching in quantum trade scene, the searched range is listed as following, the cross validation for the searching is 3-fold cross validation.

```
    "CatBoost" = {
        ‘iterations’: [100, 200],
        ‘depth’: [6, 8],
        ‘learning_rate’: [0.01, 0.1],
        ‘l2_leaf_reg’: [1, 3],
        ‘border_count’: [32, 64],
    }
```

G Original evaluation errors

We listed the original average root mean square evaluation errors for 4 different evaluation metrics: ROC-AUC, ACC, RMSE, R^2 in the 11 scenes for evaluation model learning. The reason we do not use the policy risk, Area Under the Qini Curve (Qini Coefficient, Radcliffe [2007]) and Area

Under the Uplift Curve (AUCC, Radcliffe [2007]) for discrete target is that not all models return a probability for classification. For example, the output continuous value of logistic regression can not be used to represent the probability that it belongs to a class.

RMSE	TERECO	Calculus	kin8nm	NuScale	Ad
Holdout-100	0.45471	0.26197	0.00499	0.01521	0.10611
Holdout-50	0.28538	0.32978	0.00435	0.01328	0.19199
Holdout-20	0.49894	0.29417	0.03387	0.01584	0.21959
Holdout-10	0.39836	0.22317	0.00591	0.01706	0.21805
CV-5	0.43369	0.25795	0.00525	0.01518	0.07848
CV-10	0.39857	0.25568	0.00551	0.01513	0.06187
Bootstrap	0.46825	0.26353	0.00977	0.01544	0.13219
Het-Linear	0.00474	0.00918	0.00027	0.00077	0.00555
Het-MLP	0.03528	0.31249	0.02601	0.05205	0.03862
Het-SVR	0.03881	0.04573	0.03729	0.02167	0.03665
Het-RF	0.01348	0.00513	0.00193	0.00335	0.01378
Het-LGBM	0.00998	0.00530	0.00219	0.00301	0.00985
Het-XGBoost	0.01223	0.00549	0.00235	0.00331	0.01183
Het-CatBoost	0.00756	0.00468	0.00159	0.00241	0.00695

Table 9: Evaluation error when evaluation metric is RMSE.

RMSE	TERECO	Calculus	kin8nm	NuScale	Ad
Holdout-100	0.12591	0.38410	0.01355	0.00946	0.19670
Holdout-50	0.11428	0.53892	0.04429	0.01026	7.52594
Holdout-20	0.20604	0.74010	0.03152	0.01594	6.24597
Holdout-10	0.15633	0.70377	0.04148	0.02123	5.58360
CV-5	0.19360	0.42477	0.01200	0.00950	2.98845
CV-10	0.25662	0.43089	0.01052	0.00961	3.72466
Bootstrap	0.27573	0.39105	0.01911	0.01065	0.24769
Het-Linear	0.01689	0.00061	0.00061	0.00174	0.01208
Het-MLP	0.04252	0.02690	0.02690	0.05219	0.04072
Het-SVR	0.06077	0.04628	0.04628	0.03578	0.05020
Het-RF	0.04748	0.00441	0.00441	0.00770	0.03020
Het-LGBM	0.03539	0.00508	0.00508	0.00702	0.02145
Het-XGBoost	0.04258	0.00521	0.00521	0.00756	0.02537
Het-CatBoost	0.02669	0.00368	0.00368	0.00553	0.01515

Table 10: Evaluation error when evaluation metric is R^2.

RMSE	Withdraw	Higgs	Grid	Insurance	Climate	Alert
Holdout-100	0.02443	0.02132	0.01162	0.04982	0.06820	0.02185
Holdout-50	0.03003	0.02323	0.01125	0.08180	0.07170	0.02442
Holdout-20	0.03636	0.02878	0.01143	0.08537	0.07966	0.02515
Holdout-10	0.04218	0.04217	0.01355	0.09063	0.12213	0.03166
CV-5	0.02445	0.02115	0.01161	0.06179	0.06894	0.02144
CV-10	0.02435	0.02118	0.01164	0.06180	0.06924	0.02159
Bootstrap	0.02634	0.02335	0.01236	0.04996	0.07509	0.02387
Het-Linear	0.01587	0.01438	0.00723	0.01148	0.05123	0.01391
Het-MLP	0.04397	0.04000	0.03974	0.05053	0.06654	0.04581
Het-SVR	0.03433	0.02250	0.05526	0.03135	0.06141	0.07091
Het-RF	0.02107	0.01608	0.01014	0.01645	0.05892	0.02255
Het-LGBM	0.01934	0.01517	0.01056	0.01407	0.05555	0.01890
Het-XGBoost	0.02116	0.01635	0.01016	0.01595	0.05970	0.02229
Het-CatBoost	0.01748	0.01432	0.00888	0.01262	0.05232	0.01839

Table 11: Evaluation error when evaluation metric is ROC-AUC.

RMSE	Withdraw	Higgs	Grid	Insurance	Climate	Alert
Holdout-100	0.05444	0.03244	0.01145	0.17974	0.10162	0.02764
Holdout-50	0.06632	0.02550	0.01119	0.22168	0.13740	0.03271
Holdout-20	0.06731	0.03507	0.01056	0.26858	0.16551	0.03930
Holdout-10	0.06710	0.04374	0.01312	0.28142	0.19290	0.03623
CV-5	0.05443	0.03244	0.01145	0.17974	0.10145	0.02764
CV-10	0.05443	0.03244	0.01145	0.17973	0.10135	0.02765
Bootstrap	0.05489	0.03481	0.01266	0.17976	0.10362	0.02825
Het-Linear	0.01850	0.01438	0.00696	0.01887	0.03357	0.01391
Het-MLP	0.04886	0.03994	0.03924	0.05166	0.06719	0.04581
Het-SVR	0.07261	0.03029	0.03749	0.07211	0.06634	0.07091
Het-RF	0.03444	0.01641	0.00973	0.03068	0.05216	0.02255
Het-LGBM	0.02929	0.01526	0.00972	0.02610	0.04359	0.02255
Het-XGBoost	0.03395	0.01656	0.01000	0.02981	0.05219	0.01890
Het-CatBoost	0.02655	0.01443	0.00879	0.02426	0.04645	0.01839

Table 12: Evaluation error when evaluation metric is ACC.

H Evaluation cost

We list the table to compare the average evaluation time of experimental (simulation-based) evaluation and our computation evaluation as following table 13. We use the open-source data in table 13 to build the benchmark to test the computational evaluation methods.

Scenes	Experiment (Simulation) Time	Computation Time
Alert (Wilson et al. [2021])	1 year, 8 months, and 15 days	0.176s
Withdraw (Wilson et al. [2023])	1 year, 3 months, and 13 days	0.077s
Grid-S (Schäfer et al. [2016])	1500s	0.198s
Higgs-S (Baldi et al. [2014])	(expected) 15 hours	0.151s
Insurance (Kumar [2023])	>1 year	0.500s
Climate-S (Lucas et al. [2013])	unknown	0.132s
TERECO (Li et al. [2022])	34 weeks	0.014s
Calculus (Kramer et al. [2023])	3 semesters	0.035s
Ad (Gokagglers [2018])	3 campaigns	0.034s
kin8nm-S (Fischer [2022])	unknown	0.038s
NuScale-S (Huu Tiep [2024])	1600 hours	0.046s
A-Share Trade	10 days	0.103s

Table 13: Expected evaluation time comparison in 12 scenes per agent.“S” means simulation.

I Upper bounds tables

We list the upper bound of generalized evaluation error and generalized causal effect evaluation error in the table 14 and table 15 for researchers' quick query before data collection and upper bound estimation after learning.

J Paradigm of computational evaluation

In order to help user to better understand the paradigm of computational evaluation, we plot the decision tree to help evaluatology researchers conduct their own research as shown in the figure 7.

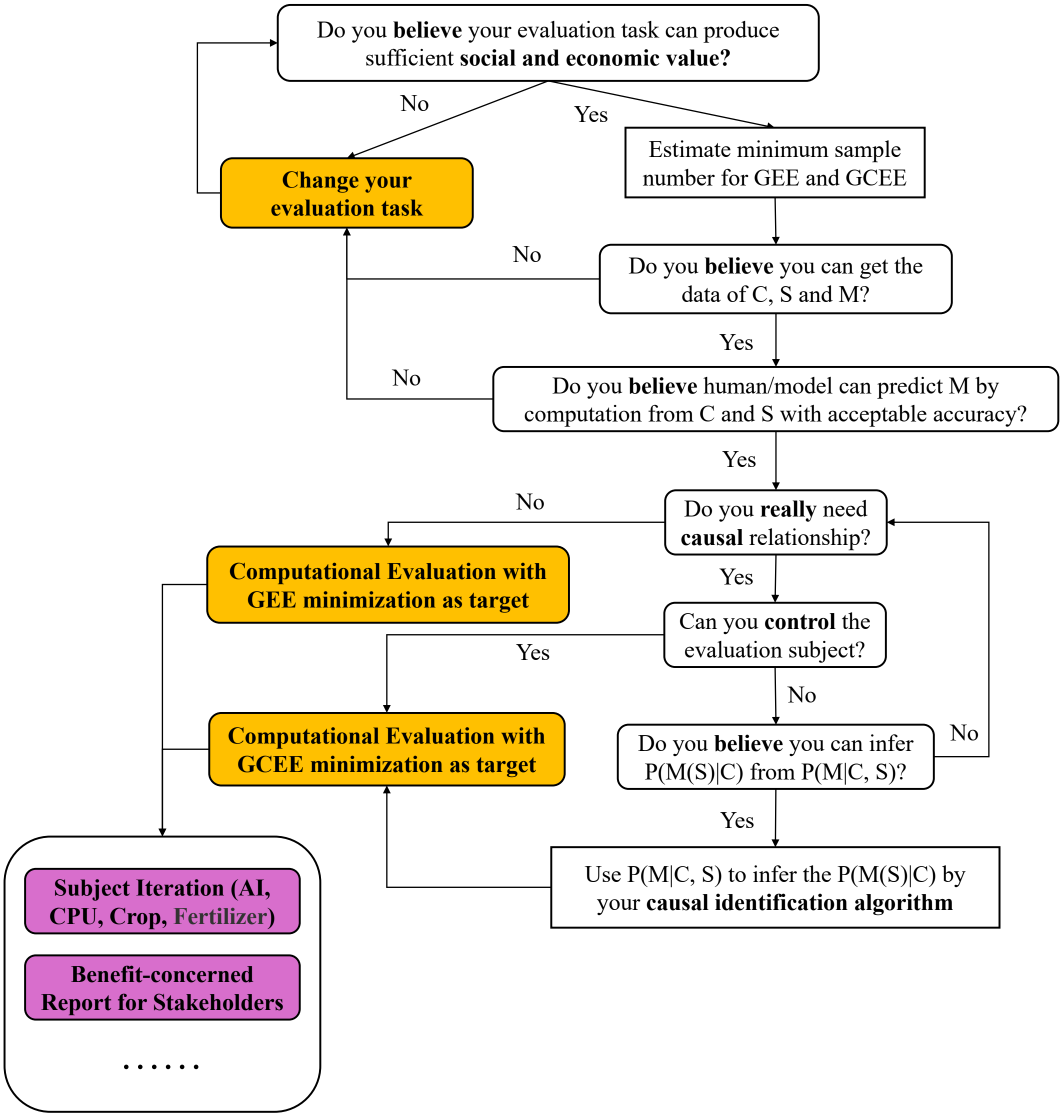

Do you believe your evaluation task can produce sufficient social and economic value?
No
Yes
Change your evaluation task
Estimate minimum sample number for GEE and GCEE
No
Do you believe you can get the data of C, S and M?
Yes
No
Do you believe human/model can predict M by computation from C and S with acceptable accuracy?
Yes
No
Do you really need causal relationship?
Computational Evaluation with GEE minimization as target
Yes
Can you control the evaluation subject?
Yes
No
Computational Evaluation with GCEE minimization as target
Do you believe you can infer P(M(S)|C) from P(M|C, S)?
No
Yes
Use P(M|C, S) to infer the P(M(S)|C) by your causal identification algorithm
Subject Iteration (AI, CPU, Crop, Fertilizer)
Benefit-concerned Report for Stakeholders
……

Figure 7: The decision tree that how to use computational evaluation models in your own evaluation task. The GEE means generalized evaluation error and the GCEE means generalized causal evaluation error. C is evaluation condition, S is evaluation subject, and M is evaluation metric. $P(M(S)|EC)$ is the potential outcome of evaluation metric after deploying the subject given evaluation condition.

Samples/Confidence	0.5	0.95	0.99	0.999
10	$E + 0.186$	$E + 0.387$	$E + 0.480$	$E + 0.588$
20	$E + 0.132$	$E + 0.274$	$E + 0.339$	$E + 0.416$
30	$E + 0.107$	$E + 0.223$	$E + 0.277$	$E + 0.339$
100	$E + 0.059$	$E + 0.122$	$E + 0.152$	$E + 0.186$
1000	$E + 0.0186$	$E + 0.0387$	$E + 0.0480$	$E + 0.0588$
10000	$E + 0.00589$	$E + 0.0122$	$E + 0.0151$	$E + 0.0186$
100000	$E + 0.00186$	$E + 0.00387$	$E + 0.00480$	$E + 0.00588$
1000000	$E + 0.000589$	$E + 0.00122$	$E + 0.00152$	$E + 0.00186$
10000000	$E + 0.000186$	$E + 0.000387$	$E + 0.000480$	$E + 0.000588$
100000000	$E + 0.0000589$	$E + 0.000122$	$E + 0.000152$	$E + 0.000186$
1000000000	$E + 0.0000186$	$E + 0.0000387$	$E + 0.0000480$	$E + 0.0000588$

Table 14: Upper bound of normalized generalized error. E is the empirical prediction error.

Samples/Confidence	0.5	0.95	0.99	0.999
10	$2E + 0.372$	$2E + 0.774$	$2E + 0.960$	$2E + 1.18$
20	$2E + 0.263$	$2E + 0.547$	$2E + 0.678$	$2E + 0.831$
30	$2E + 0.215$	$2E + 0.447$	$2E + 0.554$	$2E + 0.679$
100	$2E + 0.118$	$2E + 0.245$	$2E + 0.303$	$2E + 0.372$
1000	$2E + 0.0372$	$2E + 0.0774$	$2E + 0.0960$	$2E + 0.118$
10000	$2E + 0.0118$	$2E + 0.0245$	$2E + 0.0303$	$2E + 0.0372$
100000	$2E + 0.00372$	$2E + 0.00774$	$2E + 0.00960$	$2E + 0.0118$
1000000	$2E + 0.00118$	$2E + 0.00245$	$2E + 0.00303$	$2E + 0.00372$
10000000	$2E + 0.000372$	$2E + 0.000774$	$2E + 0.000960$	$2E + 0.00118$
100000000	$2E + 0.000118$	$2E + 0.000245$	$2E + 0.000303$	$2E + 0.000372$
1000000000	$2E + 0.0000372$	$2E + 0.0000774$	$2E + 0.0000960$	$2E + 0.000118$

Table 15: Upper bound of normalized generalized causal effect error. E is the empirical prediction error.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Section 1 is introduction.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 7 is limitation and conclusion.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: Section 3 and appendix C are theoretical results and proofs.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: The experimental details were in appendix D, E, F.

Guidelines:

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

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

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

Answer: [Yes]

Justification: The code was in the supplementary material. The data is download or scrawled from website as shown in appendix D. We will push the source code, download link, preprocessed data, intermediate data, generated data to the GitHub in future.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The experimental details were in appendix D, E, and F.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: Section 6 reports 95% confidence interval in figure 3 with empirical distribution.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Appendix D introduces the computer resources.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We checked the NeuralIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [No]

    Justification: It is mainly a theoretical research and it will not increase the gated release of models significantly because the computation requirement is not very high. There are some potential societal consequences of our work, none of which we feel must be specifically highlighted here.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [Yes]

Justification: Despite the open data we used, we also scraped a dataset of A-share stocks in China by the AkShare with MIT License.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: The website and license of used data have been introduced in the appendix D.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: We collected A-share data in China by AKshare. The dataset is too large to upload. We uploaded the code to collect the data.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: [NA]

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.