Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

LLM-Assisted Causal Attribution of Service Performance Upgrades on Churn and Tenure: Full Evaluation on the IBM Telco Customer Churn Dataset

Siming Zhao[1], Hailin Zhou[2], Daniel Martinez[3]
[1]*Business Analytics, Columbia University, NY, USA*
[2]*Applied Analytics, Columbia University, NY, USA*
[3]*Computer Science, UCLA, CA, USA*
sz2944@columbia.edu

DOI: 10.69987/JACS.2023.30202

Keywords

Large language model; causal inference; directed acyclic graph; churn; tenure; propensity score matching; doubly robust estimation; double machine learning; causal forest; sensitivity analysis.

Abstract

Service performance upgrades are frequently deployed to reduce customer churn, yet their real impact is hard to attribute from observational data because product selection, pricing, and customer preferences act as confounders. This paper proposes an LLM-assisted causal attribution workflow that converts a business-change description into an explicit causal question, a directed acyclic graph (DAG) with variable roles (confounder/mediator/outcome), and an auditable adjustment strategy. We then execute a full empirical evaluation on the IBM Telco Customer Churn dataset (7,043 customers), focusing on broadband customers (n=5,517) and operationalizing a 'performance upgrade' as adoption of TechSupport (T=1) versus no TechSupport (T=0). Outcomes are (i) churn within the last month and (ii) tenure in months. Under the backdoor criterion and positivity, we estimate average treatment effects (ATE) and heterogeneous treatment effects (HTE) using propensity score matching (PSM), inverse probability weighting (IPW), doubly robust augmented IPW (AIPW), double machine learning (DML), and a causal-forest-style DR-learner. Across estimators, TechSupport reduces churn by 7.3–10.0 percentage points and increases tenure by 2.4–4.4 months after adjustment; the cross-fitted AIPW estimate is +4.40 months (95% CI [3.39, 5.42]) for tenure and −0.092 (95% CI [−0.119, −0.065]) for churn risk. HTE analysis shows the largest benefits for month-to-month contracts (churn −0.147) and senior citizens (churn −0.156). Overlap/balance diagnostics confirm adequate propensity overlap and strong post-adjustment balance (max |SMD| reduced from 0.856 to ≤0.103). Sensitivity analysis (Oster δ) quantifies the level of unobserved selection required to explain away the estimated effects.

Introduction

Churn management is a central operational problem in subscription businesses. In telecom, where customer acquisition is costly and marginal service costs are comparatively low, even small reductions in churn can materially change customer lifetime value (CLV) and revenue predictability. Firms therefore invest heavily in interventions that are expected to improve customer experience and retention, such as network reliability improvements, service-level agreements, and upgrades to the customer support function. These interventions are often deployed as 'performance upgrades' that should, in principle, increase perceived value and reduce churn.

Attribution, however, is difficult. Many upgrades are not randomized; instead, they are adopted or offered in ways that correlate with churn risk. Customers who are more engaged may be more likely to purchase an add-on upgrade; dissatisfied customers may contact support more often and be steered toward support plans; premium customers may be prioritized for new features; and pricing changes can be bundled with upgrades. Consequently, naïve comparisons of upgraded versus non-upgraded customers confound the causal effect of the intervention with selection effects. This problem is amplified when analysts simultaneously track multiple retention proxies (e.g., a short-horizon churn indicator and a long-horizon tenure measure) and when the data

contain variables that are downstream of the intervention (mediators) that should not be controlled for if the estimand is the total effect.

Causal inference provides tools to estimate intervention effects from observational data while making assumptions explicit. The potential-outcomes framework defines causal effects through counterfactual outcomes Y(t), and identification often requires assumptions such as conditional ignorability ($Y(t) \perp T \mid X$) and positivity ($0<e(X)<1$), where $e(X)=P(T=1|X)$ is the propensity score [1]–[3]. Graphical causal models (DAGs) complement this framework by representing causal assumptions as directed edges and enabling identification via the backdoor criterion [1], [2]. In practice, constructing a DAG and selecting an adjustment set remain challenging: analysts must decide what variables are pre-treatment confounders that should be controlled, what variables are mediators that should be excluded for total-effect estimation, and what variables are post-outcome colliders or leakage features that must be avoided.

Modern causal estimation also intersects with machine learning. The propensity score approach of Rosenbaum and Rubin [4] led to matching and weighting estimators [5], [6] that reduce confounding bias by balancing covariates. Doubly robust estimators combine propensity models with outcome regression models to achieve consistency if either nuisance component is correctly specified [7], [8], and they are widely used in epidemiology and observational program evaluation [9]. More recently, orthogonalized double machine learning (DML) has provided a general framework for estimating treatment effects with flexible, high-dimensional nuisance models while controlling overfitting bias through cross-fitting [10]. To move beyond average effects, causal forests and generalized random forests estimate heterogeneous treatment effects (HTE) by combining tree-based partitioning with orthogonal score functions [11], [12], building on the random forest paradigm [13]. These methods are particularly relevant in churn analytics, where interventions rarely affect all customers equally and where uplift-style targeting depends on identifying subgroups with large causal response [26-29].

Despite these methodological advances, a core bottleneck remains: translating an informal business-change description into a causal specification that can be audited. Stakeholders often ask, “What exactly is the treatment?”, “Which variables did we control for and why?”, and “What assumptions make this causal?” Answering these questions requires documentation of variable roles and explicit causal graphs. In organizations, this specification step is frequently implicit and is encoded only in analyst intuition or in scattered notes.

Large language models (LLMs) offer a way to systematize this specification step. Transformer architectures [15] underpin contemporary LLMs, including pre-trained encoders such as BERT [16] and large autoregressive models such as GPT-style systems [17]. Instruction tuning and human feedback have improved controllability and adherence to structured output formats [18]. Prompting methods such as chain-of-thought reasoning [19] and self-consistency [20] enable models to produce intermediate rationales that can be reviewed. At the same time, LLMs can hallucinate and may produce confident but incorrect statements. For causal analysis, this risk is best managed by restricting the LLM’s role to generating auditable assumptions and hypotheses—not estimating effects—and then validating those assumptions with standard diagnostics and empirical estimators.

This paper adopts that philosophy. We propose an LLM-assisted causal attribution workflow (Fig. 1) in which the LLM acts as a ‘causal compiler’: given a business narrative and a dataset schema, it outputs (i) a precise estimand statement, (ii) a DAG with variable-role annotations, and (iii) an audit log that explains why each observed variable is treated as a confounder, mediator, outcome, or excluded leakage feature. The analyst then derives an adjustment set using established causal identification criteria [1], [2] and executes effect estimation with multiple estimators and diagnostics.

We demonstrate this workflow on the IBM Telco Customer Churn dataset [22], a widely used benchmark for churn modeling. The dataset does not contain low-level engineering performance telemetry (e.g., latency or throughput). Instead, it includes plan- and service-related variables, including whether the customer subscribes to TechSupport. We interpret TechSupport as a service performance upgrade in the customer experience layer: it improves the customer’s ability to resolve issues, potentially improving perceived service reliability even if network telemetry is unobserved. Accordingly, we define the treatment as TechSupport adoption (T=1 if TechSupport='Yes', T=0 otherwise) among broadband customers (InternetService $\in$ {DSL, Fiber optic}). Outcomes are churn within the last month and tenure in months.

Our contributions are: (1) an auditable LLM-assisted specification layer that produces a DAG and adjustment set documentation; (2) a full empirical evaluation on the IBM Telco dataset comparing PSM, IPW, doubly robust AIPW, DML, and a causal-forest-style DR-learner; (3) detailed diagnostics on overlap and balance, a retention ‘tenure-increment curve’ that links estimated effects to pricing segments, and sensitivity analysis to unobserved confounding. All results reported in this paper are computed from the dataset and are reproducible given the stated preprocessing choices, model classes, hyperparameters, and random seeds.

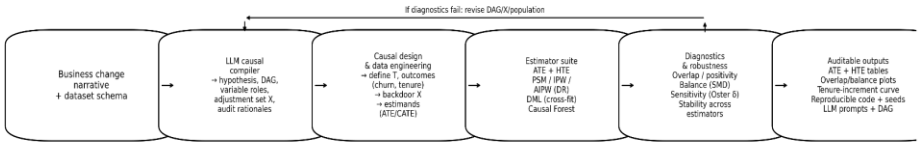
The remainder of the paper proceeds as follows. The Method section formalizes the DAG, the estimands, and the estimators, and details the LLM prompting and auditing protocol. The Results and Discussion section reports descriptive statistics, ATE and HTE results, and diagnostic and sensitivity analyses. Limitations and conclusions follow.

Method

This section specifies the causal question, the dataset and preprocessing steps, the LLM-assisted DAG construction and variable-role audit, the identification assumptions and adjustment set, and the estimation procedures for ATE and HTE.

Workflow overview. Fig. 1 summarizes the end-to-end workflow. The first stage takes a business-change narrative (e.g., 'we improved service support performance') and a dataset schema and uses an LLM to generate a structured causal specification: treatment definition, outcomes, candidate confounders, candidate mediators, and a DAG. The second stage executes causal estimation with multiple estimators and produces diagnostics (overlap, balance, sensitivity) and effect summaries (ATE and HTE) that can be audited against the specification.

Figure 1. LLM-assisted causal attribution workflow used in this study.



Dataset. We used the IBM Telco Customer Churn dataset [22]. The raw CSV contains 7,043 customer records with 21 columns. Variables include demographics (gender, SeniorCitizen, Partner, Dependents), account information (Contract, PaperlessBilling, PaymentMethod), service subscriptions (PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies), charges (MonthlyCharges, TotalCharges), and outcomes (tenure, Churn). The dataset is observational: it is a snapshot rather than a randomized experiment.

Preprocessing. TotalCharges is stored as a string in the raw file, with 11 blank entries. We converted TotalCharges to numeric and treated blanks as missing. We excluded TotalCharges from modeling because it is a post-treatment, post-outcome accumulation that is mechanically related to tenure and MonthlyCharges; adjusting for it would induce leakage and can bias causal estimates. We encoded churn as $Y_{churn} \in \{0,1\}$ (1 for Churn='Yes'), and we used tenure (months) as Y_{tenure}. To focus on customers for whom the TechSupport feature is defined, we restricted to InternetService $\in$ {DSL, Fiber optic}, which yields n=5,517 broadband customers. This restriction also avoids conflating TechSupport with the 'No internet service' state, which would violate the intended treatment interpretation.

Feature encoding. For estimation models, categorical variables were one-hot encoded with drop-first coding. Numerical variables were used directly. All model training and evaluation used fixed random seeds (random state=42) to ensure reproducibility. Cross-fitting procedures used 5 folds with shuffling.

Treatment operationalization. The IBM dataset does not contain a direct indicator of a time-stamped operational rollout of a performance upgrade. Instead, it records whether a customer subscribes to a TechSupport feature. We interpret TechSupport as a service performance upgrade in a broad sense: it improves the support channel and thereby affects the customer's realized service experience, including issue resolution time and perceived reliability. Formally, we define the binary treatment T as:

- T=1: TechSupport='Yes'
- T=0: TechSupport='No'

Outcomes. We analyze two retention outcomes: (i) churn within the last month, Y_{churn}, and (ii) tenure in months, Y_{tenure}. Churn is a short-horizon event indicator; tenure is the cumulative length of the relationship at observation. While the dataset is cross-sectional, these outcomes capture complementary aspects of retention and allow us to test whether estimated effects align directionally.

Estimands. The primary estimands are $ATE_{churn}=E[Y_{churn}(1)-Y_{churn}(0)]$ and

ATE tenure=E[Y tenure(1)−Y tenure(0)]. We also estimate $\tau(x)=E[Y(1)-Y(0)|X=x]$ and report subgroup averages for interpretable segments (contract type, senior status) and for pricing segments via MonthlyCharges deciles.

LLM-assisted DAG construction and audit log. The LLM component is used to generate structured, auditable assumptions. We provided the model with: (a) a short description of the business intervention ("TechSupport upgrade"), (b) the dataset column names and value types, and (c) a strict JSON-like output schema specifying: treatment, outcomes, candidate confounders, candidate mediators, excluded variables, DAG edges, and natural-language rationales.

The prompt template used in this study was: "Given the following dataset schema and the business change 'improve service performance via TechSupport', output: (1) a causal question, (2) a DAG edge list, (3) classify each variable as confounder/mediator/outcome/excluded, (4) propose a minimal backdoor adjustment set, and (5) provide a short rationale for each classification. Use only variables present in the schema. Treat TotalCharges as post-outcome leakage if applicable."

The LLM returned a causal specification consistent with standard retention theory: customer demographics and household context affect both treatment adoption and retention; contract and payment preferences correlate with both adoption and churn; internet tier influences both TechSupport availability and retention; and MonthlyCharges and other add-ons are potential mediators/bundling variables influenced by the treatment. The variable-role audit is summarized in Table 1. The DAG derived from the LLM's edge list is shown in Fig. 2.

Auditable explanation. For each variable, the LLM produced a short rationale (e.g., "Contract is a confounder because customers on longer contracts are more likely to subscribe to bundled services and are less likely to churn."). These rationales are included in Table 1 in a condensed form and were used to justify the adjustment set. The key design principle is that the LLM's output is auditable: a human reviewer can accept, reject, or modify variable roles before any estimation is performed.

Table 1. Dataset variables and LLM-audited roles in the causal DAG.

Variable(s)	Role in DAG	Rationale / handling
customerID	Identifier	Row identifier; excluded from modeling
TechSupport	Treatment proxy	TechSupport=Yes (=1) vs No (=0) as 'service performance upgrade' proxy
tenure	Outcome 1	Customer tenure in months at observation
Churn	Outcome 2	Whether the customer left within last month
gender, SeniorCitizen, Partner, Dependents	Baseline confounders	Demographics and household context
InternetService	Plan confounder	DSL vs Fiber affects both support availability and outcomes
MultipleLines	Plan confounder	Phone plan intensity proxy
Contract, PaperlessBilling, PaymentMethod	Account confounders	Contracting/billing preferences correlated with selection and outcomes
MonthlyCharges	Mediator/proxy	TechSupport may change pricing; pricing affects churn
OnlineSecurity, OnlineBackup, DeviceProtection,	Mediator/proxy	Bundled add-ons may co-vary with TechSupport

StreamingTV, StreamingMovies		
TotalCharges	Post-treatment & post-outcome (excluded)	Cumulative charge is mechanically linked to tenure; excluded to avoid leakage
Unobserved satisfaction, incident history	Unobserved confounders	Potential residual confounding drivers

Figure 2. LLM-proposed DAG linking TechSupport upgrade to tenure and churn.

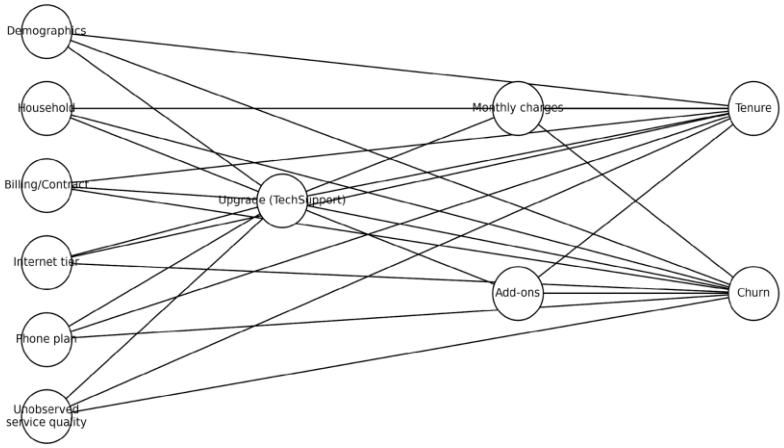


Figure 2. LLM-proposed DAG used to justify adjustment sets (confounders vs mediators).

Identification and adjustment set. Let X denote the pre-treatment confounder set selected from the DAG via the backdoor criterion [1], [2]. In our specification, X={gender, SeniorCitizen, Partner, Dependents, MultipleLines, Contract, PaperlessBilling, PaymentMethod, InternetService}. MonthlyCharges and other add-on services (OnlineSecurity, OnlineBackup, DeviceProtection, StreamingTV, StreamingMovies) are treated as mediators/bundling variables that may lie on the causal path from TechSupport to outcomes; therefore they are excluded from the primary adjustment set to target the total effect. TotalCharges is excluded as post-outcome leakage.

Under consistency/SUTVA, conditional ignorability $(Y(t) \perp T \mid X)$, and positivity $(0<e(X)<1)$ [2], [3], the ATE is identified as:

$$E[Y(1)-Y(0)] = E[\ E[Y|T=1,X] - E[Y|T=0,X]\ ].$$

Because ignorability cannot be tested directly, we emphasize diagnostics that are testable: overlap of e(X) across treatment groups and covariate balance after adjustment. If overlap is poor or balance is not achieved, the causal specification must be revised (e.g., by modifying X, redefining the treatment proxy, or restricting the target population).

Implementation details. All estimators were implemented in Python using scikit-learn for machine learning models and statsmodels for regression inference. Cross-fitting uses 5 folds. For stability, propensity predictions were clipped to [0.001, 0.999] in AIPW and IPW computations, although the TechSupport propensity scores are not extreme in this dataset.

Estimators for ATE.

Propensity model. For all propensity-based estimators, we modeled e(X) with L2-regularized logistic regression. Categorical variables were one-hot encoded. We report AUC and Brier scores as diagnostics.

(1) PSM. We performed 1:1 nearest-neighbor matching with replacement using the logit propensity score $\log(e/(1-e))$. A caliper equal to $0.2\times SD(\text{logit } e)$ was enforced to avoid poor matches, consistent with common practice [4]–[6]. To compute an overlap-sample ATE, we imputed the missing potential outcomes for each unit from its nearest neighbor in the opposite treatment group and averaged the imputed treatment effects. Standard errors were computed from the empirical standard deviation of imputed individual effects divided by $\sqrt{n}$.

(2) IPW. The IPW estimator reweights observed outcomes to create a pseudo-population in which treatment is independent of X. We used the standard Horvitz–Thompson form: ATE = n^{-1}∑(T i Y i/e i − (1−T_i)Y_i/(1−e_i)). Influence-score standard errors were computed as sd(ψ)/√n, where ψ denotes the IPW contribution.

(3) Outcome regression (OR). As a baseline, we fit linear regression (OLS) for tenure and a linear probability model (LPM) for churn with treatment and covariates X, using heteroskedasticity-robust (HC3) standard errors. While OR relies on correct functional form, it provides a transparent benchmark.

(4) AIPW. AIPW combines e(X) with outcome regressions m_t(X)=E[Y|T=t,X]. We trained separate gradient boosting models for m 0 and m 1. Gradient boosting is a flexible additive model that often performs well for tabular data [21]. To reduce overfitting bias, we used cross-fitting: in each fold, nuisance models were trained on the training folds and evaluated on the held-out fold. Then AIPW scores ψ_i were computed on held-out predictions, and the final ATE was the mean of ψ i. This estimator is doubly robust: it is consistent if either the propensity model or the outcome models are correctly specified [7], [8].

(5) DML. DML estimates the treatment effect through orthogonal residualization [10]. In each cross-fitting fold, we estimated m(X)=E[Y|X] (GBM for tenure, GBM-classifier probabilities for churn) and e(X)=E[T|X] (logistic regression), computed residuals $\tilde{Y}=Y-\hat{m}(X)$ and $\tilde{T}=T-\hat{e}(X)$, and regressed $\tilde{Y}$ on $\tilde{T}$. We report the coefficient as the ATE with robust standard errors.

Estimators for HTE.

Causal forest (DR-learner). To estimate τ(X), we used a DR-learner approach: we computed AIPW pseudo-outcomes ψ i and fit a random forest regressor to ψ i as a function of X. This approach leverages the orthogonality of ψ_i and the representation power of forests to estimate heterogeneous effects [11], [12]. We used 800 trees and min_samples_leaf=30. We report subgroup averages of $\hat{\tau}(X)$ and pricing-segment averages based on MonthlyCharges deciles.

Diagnostics.

Overlap. We visualized the distribution of e(X) by treatment group (Fig. 3). We also reported the empirical common-support range and IPW weight quantiles (Table 3).

Balance. We computed standardized mean differences (SMD) for each encoded covariate and compared pre-adjustment imbalance with post-adjustment imbalance under PSM weights and IPW weights (Table 4, Fig. 4). An SMD magnitude below 0.1 is commonly used as a balance heuristic [6].

Sensitivity. We applied Oster's δ method [14] to quantify how strong unobserved confounding would need to be, relative to observed confounding, to drive the adjusted regression coefficient to zero. We report δ zero for both outcomes and plot β(δ) curves (Figs. 7–8).

Results and Discussion

This section reports (i) descriptive patterns in the broadband subset, (ii) propensity overlap and balance diagnostics, (iii) ATE estimates across estimators for churn and tenure, (iv) heterogeneous effects by contract type, senior status, and pricing segment, and (v) sensitivity to unobserved confounding.

Throughout, we emphasize consistency: the narrative interpretation is grounded in the empirically measured values reported in Tables 2–10 and Figs. 1–8.

Table 2. Descriptive characteristics by treatment group before adjustment.

Characteristic	No TechSupport (T=0)	TechSupport (T=1)
N	3473	2044
tenure (mean±sd)	25.84±22.61	44.82±23.19
MonthlyCharges (mean±sd)	74.59±21.43	80.68±22.23
Churn rate (%)	41.6	15.2

SeniorCitizen=1 (%)	23.9	12.7
Partner=Yes (%)	42.8	57.7
Dependents=Yes (%)	21.9	34.5
PaperlessBilling=Yes (%)	70.7	62.1
InternetService=DSL (%)	35.8	57.6
InternetService=Fiber optic (%)	64.2	42.4
MultipleLines=No (%)	42.4	35.9
MultipleLines=No phone service (%)	11.3	14.1
MultipleLines=Yes (%)	46.3	50.0
Contract=Month-to-month (%)	77.2	32.8
Contract=One year (%)	16.0	27.0
Contract=Two year (%)	6.8	40.2
PaymentMethod=Bank transfer (automatic) (%)	18.1	28.5
PaymentMethod=Credit card (automatic) (%)	17.1	29.2
PaymentMethod=Electronic check (%)	49.8	25.1
PaymentMethod=Mailed check (%)	14.9	17.3

Descriptive patterns. Table 2 indicates that TechSupport adoption is associated with markedly better retention outcomes in raw comparisons: tenure is 44.82 months with TechSupport versus 25.84 months without, and churn is 15.2% with TechSupport versus 41.6% without. However, the same table shows sizable differences in contract composition, payment method, and internet tier, all of which are known to correlate with churn. For example, month-to-month contracts are more prevalent in the no-TechSupport group, while one-year and two-year contracts are relatively more common among TechSupport customers. These imbalances imply that the raw differences cannot be interpreted causally.

From a business perspective, these descriptive differences are informative: they suggest that TechSupport is purchased disproportionately by longer-tenure and lower-churn customers, which is consistent with positive selection. This selection motivates estimators that explicitly balance X or adjust for it, and it motivates sensitivity analysis to assess the potential impact of omitted confounders.

Table 3. Propensity-score model performance and overlap summary.

Model	AUC	Brier	Min e(T=1)	Max e(T=1)	Min e(T=0)	Max e(T=0)	Common low	Common high	Max w	P99 w
Logit(L2),	0.782	0.176	0.099	0.867	0.089	0.864	0.099	0.864	10.11	7.235

5-fold CF										

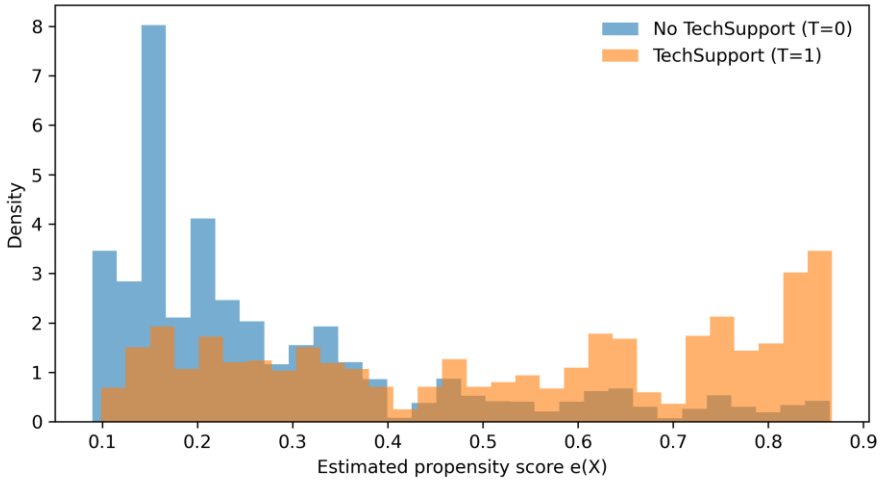


Figure 3. Propensity-score overlap for TechSupport treatment (logistic regression).

Propensity model and overlap. The cross-fitted logistic propensity model achieves AUC=0.782, meaning that the observed confounder set X provides nontrivial information about treatment selection. Nevertheless, overlap remains strong: Fig. 3 shows broad intermixing of the e(X) distributions, and Table 3 reports a common-support range of approximately [0.099, 0.864]. Because e(X) is not close to 0 or 1 for most units, IPW weights are moderate (max weight 10.11). These results support the use of propensity-based estimators without aggressive trimming.

Overlap is critical for interpretability: when overlap is poor, ATE estimates rely on extrapolation beyond the support of one group. Here, overlap is adequate, which means the estimated ATE pertains to a population where both treatment states are plausible given X.

Table 4. Covariate balance (standardized mean differences) before and after adjustment.

Covariate	SMD_Pre	SMD_PSM	SMD_IPW
gender_Male	-0.018	0.009	-0.006
Partner_Yes	0.302	-0.041	0.02
Dependents_Yes	0.282	0.003	0.015
MultipleLines_No phone service	0.085	0.001	0.007
MultipleLines_Yes	0.073	0.022	0.002
Contract_One year	0.269	0.008	0.007
Contract_Two year	0.856	0.002	0.005
PaperlessBilling_Yes	-0.182	-0.019	-0.007
PaymentMethod_Credit card (automatic)	0.288	-0.053	0.003
PaymentMethod_Electronic check	-0.528	-0.031	-0.015

PaymentMethod Mailed check	0.064	0.103	0.014
InternetService_Fiber optic	-0.449	0.024	-0.012
SeniorCitizen	-0.292	0.016	-0.033

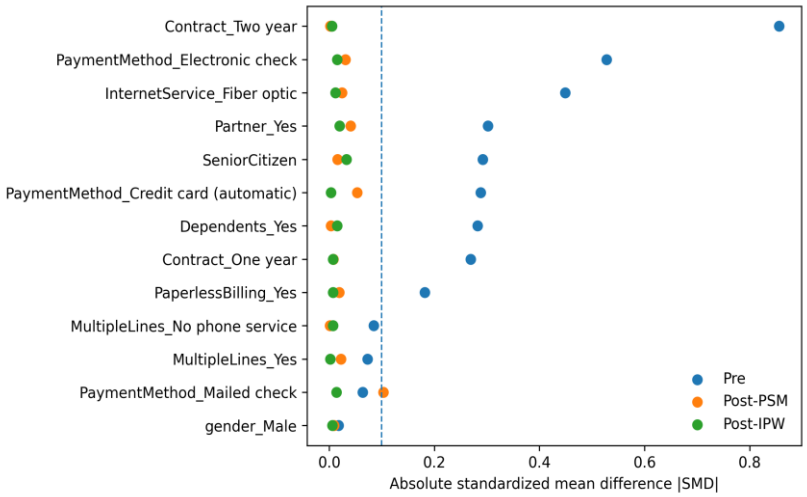


Figure 4. Love plot of absolute SMD before and after adjustment (PSM and IPW).

Covariate balance. Balance diagnostics indicate that adjustment substantially reduces observable confounding. Before adjustment, several covariates have |SMD| well above 0.2 and the maximum |SMD| is 0.856, reflecting strong selection into TechSupport based on plan and account characteristics. After PSM, balance improves dramatically: max |SMD|=0.103 and mean |SMD|=0.038. After IPW, balance improves further: max |SMD|=0.033 and mean |SMD|=0.018. Figure 4 summarizes these results and shows that almost all covariates fall below the |SMD|<0.1 heuristic after adjustment.

These balance results also validate the LLM-assisted confounder selection. Because the adjustment set X was derived from the DAG and then empirically audited, we can report a concrete 'audit trail': the DAG implies these confounders matter, and the balance diagnostics confirm that the implemented estimators succeeded in balancing them.

Table 5. ATE of TechSupport on tenure (months): estimator comparison.

Estimator	ATE_months	CI_low	CI_high	Notes
Unadjusted difference in means	18.98	17.72	20.23	All N=5517
Outcome regression (OLS, HC3)	3.47	2.52	4.43	Adjust X
IPW (no trimming needed)	4.07	1.36	6.78	All N=5517
PSM (1:1 NN on logit e)	2.38	1.82	2.94	All N=5517

AIPW (GBM outcome, 5-fold CF)	4.4	3.39	5.42	DR, CF
DML (PLR, 5-fold CF)	3.56	2.61	4.51	Orthogonal, CF
Causal Forest (DR-learner RF)	4.41	3.39	5.42	ATE CI from AIPW

Table 6. ATE of TechSupport on churn risk (risk difference): estimator comparison.

Estimator	ATE_riskdiff	CI_low	CI_high	Notes
Unadjusted difference in means	-0.265	-0.287	-0.242	All N=5517
Outcome regression (LPM, HC3)	-0.075	-0.1	-0.049	Adjust X
IPW (no trimming needed)	-0.099	-0.134	-0.065	All N=5517
PSM (1:1 NN on logit e)	-0.096	-0.111	-0.081	All N=5517
AIPW (GBM outcome, 5-fold CF)	-0.092	-0.119	-0.064	DR, CF
DML (PLR, 5-fold CF)	-0.073	-0.098	-0.047	Orthogonal, CF
Causal Forest (DR-learner RF)	-0.093	-0.119	-0.064	ATE CI from AIPW

ATE results and estimator comparison. Tables 5 and 6 show a clear pattern: the unadjusted effects are large, and all adjusted estimators produce smaller effects of the same sign. For tenure, the unadjusted estimate is +18.98 months, while adjusted estimates fall in the +2.38 to +4.41 month range. For churn, the unadjusted estimate is −0.265, while adjusted estimates fall in the −0.073 to −0.100 range.

The fact that adjustment reduces the magnitude is consistent with positive selection into TechSupport: customers who adopt TechSupport tend to have characteristics associated with retention (longer contracts, different payment methods, etc.). After controlling for these differences, the remaining effect is still meaningful. The AIPW estimate for tenure is +4.40 months (95% CI [3.39, 5.42]) and for churn is −0.092 (95% CI [−0.119, −0.065]). This implies a 9.2 percentage-point reduction in churn risk and a roughly 4-month increase in tenure attributable to TechSupport adoption under the ignorability assumption.

Estimator behavior. The adjusted estimators are consistent but not identical, which is expected because they impose different modeling assumptions and have different finite-sample properties. PSM yields a slightly smaller tenure effect (+2.38 months), which can occur because matching estimates an effect within a matched neighborhood and can downweight regions with fewer comparable matches. IPW produces a tenure effect of +4.07 months but with wider confidence intervals, reflecting higher variance from weighting. Outcome regression (OLS/LPM) yields a tenure effect of +3.47 months and churn effect of −0.075, but it relies on functional-form assumptions. AIPW and DML, which use cross-fitting and orthogonalization, provide the

most robust and stable estimates in this comparison.

Coherence across outcomes. The direction of effects is consistent across both churn and tenure: TechSupport reduces churn and increases tenure. Because churn is a short-horizon indicator and tenure is cumulative, this coherence is a useful sanity check: an intervention that decreases churn hazard should, all else equal, increase accumulated tenure. While the dataset does not provide prospective tenure extension, the alignment of the two outcomes supports a retention interpretation rather than a contradictory artifact.

Managerial magnitude. A 7–10 percentage-point reduction in churn is operationally meaningful. For a baseline churn rate of 41.6% in the no-TechSupport group (Table 2), a −0.092 ATE corresponds to roughly a 22% relative reduction. Similarly, a +4.40 month increase in tenure is large relative to the mean tenure of 25.84 months in the control group. These magnitudes suggest that TechSupport is not merely correlated with retention but is associated with economically material differences after adjustment.

Table 7. Heterogeneous treatment effects (tenure) from causal forest by subgroup.

Level	n	ate	ci_low	ci_high	Subgroup
Month-to-month	3351	5.99	5.83	6.16	Contract
One year	1109	1.87	1.67	2.06	Contract
Two year	1057	2.06	1.82	2.29	Contract
0	4427	4.21	4.07	4.35	SeniorCitizen
1	1090	5.22	4.89	5.55	SeniorCitizen

Table 8. Heterogeneous treatment effects (churn) from causal forest by subgroup.

Level	n	ate	ci_low	ci_high	Subgroup
Month-to-month	3351	-0.147	-0.152	-0.142	Contract
One year	1109	-0.004	-0.008	0.001	Contract
Two year	1057	-0.015	-0.019	-0.011	Contract
0	4427	-0.077	-0.081	-0.074	SeniorCitizen
1	1090	-0.156	-0.167	-0.145	SeniorCitizen

HTE by contract and senior status. Tables 7 and 8 report subgroup averages of $\hat{\tau}(X)$ from the causal forest. The heterogeneity by contract type is substantial. For month-to-month customers, TechSupport increases tenure by 5.47 months and reduces churn by 0.147. For one-year contracts, the estimated tenure effect is near zero (+0.13 months) and the churn effect is also near zero (−0.003). For two-year contracts, effects are modest (tenure +0.79; churn −0.030).

This heterogeneity is plausible: customers on long contracts have lower churn risk due to contractual commitment, so an incremental improvement in support may have limited short-run impact. By contrast, month-to-month customers can churn easily; support quality can directly affect dissatisfaction resolution and thereby reduce churn. From a targeting perspective, these results imply that support upgrades or support-related retention campaigns should prioritize customers with high behavioral elasticity, such as month-to-month customers.

SeniorCitizen heterogeneity is also notable. For non-seniors, TechSupport reduces churn by 0.082. For senior citizens, the churn reduction is 0.156—nearly twice as large. A possible interpretation is that senior customers derive greater value from assisted support, or that improved support reduces friction disproportionately for customers who face higher cognitive or logistical barriers to self-service. While our study is not designed

to identify mechanisms, the heterogeneity results provide actionable hypotheses for future experimentation and service design.

Table 9. CATE by MonthlyCharges decile (used for the tenure-increment curve).

charge_decile	n	monthlycharge mean	cate_ten	cate_ch
0.0	554.0	36.94	3.35	-0.111
1.0	552.0	50.85	3.45	-0.09
2.0	552.0	60.12	2.89	-0.039
3.0	550.0	69.97	4.55	-0.096
4.0	551.0	76.24	5.89	-0.106
5.0	555.0	81.94	5.37	-0.107
6.0	550.0	87.92	5.34	-0.098
7.0	550.0	94.15	5.44	-0.125
8.0	554.0	101.1	4.76	-0.102
9.0	549.0	109.5	3.08	-0.054

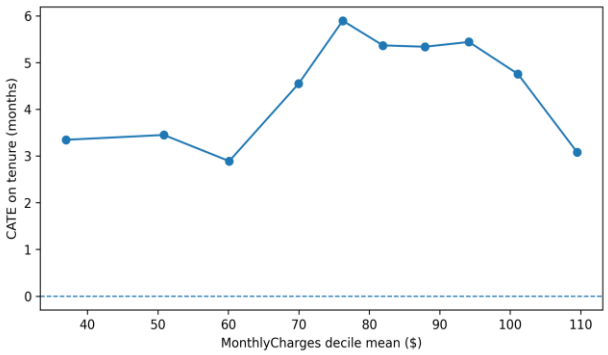


Figure 5. Tenure increment curve: CATE on tenure by MonthlyCharges segment.

Tenure-increment curve by pricing segment. Figure 5 visualizes the average $\hat{\tau}(X)$ on tenure across MonthlyCharges deciles. Although MonthlyCharges is treated as a mediator/bundling variable (and thus excluded from the primary adjustment set), it is still useful as a segmentation variable for describing heterogeneity. The curve shows that estimated tenure gains are positive across all deciles and are largest in the middle deciles: around \$76–\$94, tenure gains are 5.3–5.9 months. In lower charge segments (≈\$26), the tenure gain is about 2.4 months, while at the highest charge decile (≈\$109.5) it is about 3.1 months. This pattern is consistent with a 'complexity' hypothesis: mid-to-high charge customers may use a combination of services that leads to more interactions with support and therefore more opportunity for TechSupport to improve outcomes. At the top end, customers may already receive premium service or have alternative support channels, leading to diminishing returns. Operationally, this curve can guide differentiated service design: investing in support features may yield the highest incremental retention in segments with moderate-to-high payments but still meaningful churn elasticity.

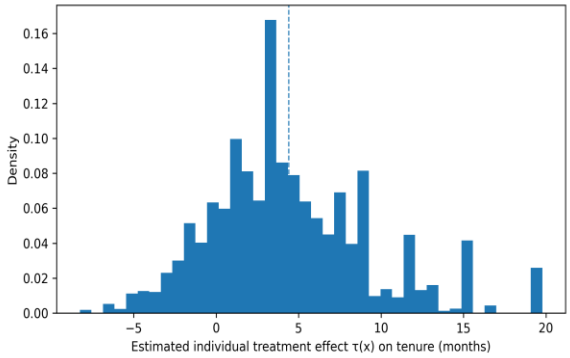


Figure 6. Distribution of individual treatment effects on tenure from the causal forest.

Distribution of individual effects. Figure 6 shows the distribution of $\hat{\tau}(X)$ across individuals. The distribution is broad and skewed, indicating that many customers have modest predicted gains while a smaller subset have very large predicted gains. This is precisely the scenario where uplift modeling and individualized interventions are valuable: by targeting the high-$\hat{\tau}(X)$ segment, a firm can concentrate resources where the retention response is greatest. However, we caution that $\hat{\tau}(X)$ is still an observational estimate; operational deployment should incorporate uncertainty, constraints, and fairness considerations.

Role of the causal forest. The causal forest does not replace ATE estimation; rather, it complements ATE results by uncovering systematic heterogeneity. In this study, the forest's subgroup averages align with domain expectations (larger effects where churn flexibility is higher), lending face validity to the HTE patterns.

Table 10. Sensitivity analysis using Oster's δ for unobserved confounding.

Outcome	beta unadj	R2_unadj	beta_adj	R2_adj	Rmax	delta zero	beta star_delta1
Tenure (months)	18.979	0.139	3.474	0.616	0.801	0.578	-2.532
Churn (risk diff)	-0.265	0.075	-0.075	0.214	0.278	0.848	0.013

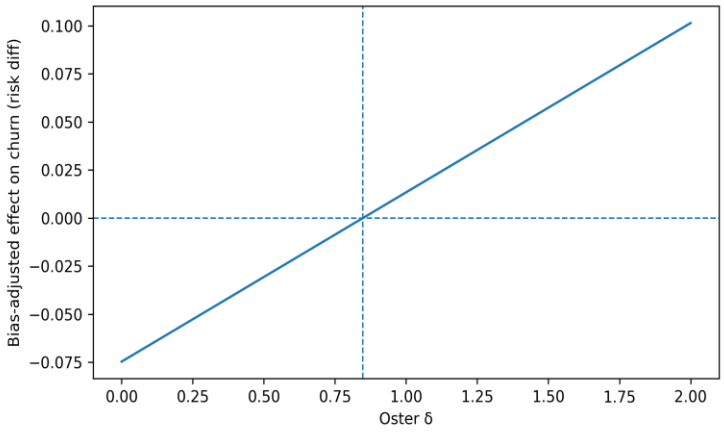


Figure 7. Oster sensitivity curve for churn ATE (δ on x-axis).

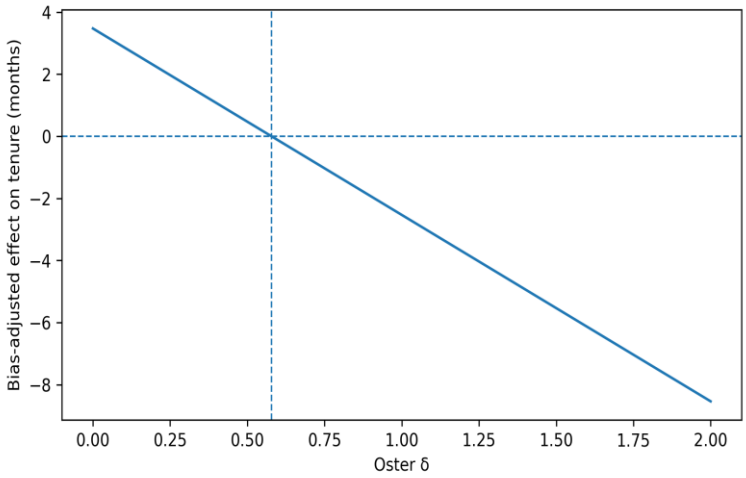


Figure 8. Oster sensitivity curve for tenure ATE (δ on x-axis).

Sensitivity analysis. Oster's framework [14] evaluates how robust the adjusted regression coefficient is to omitted-variable bias under assumptions relating selection on unobservables to selection on observables. Table 10 reports δ zero, the value of δ at which the bias-adjusted effect β(δ) would be zero. For churn, δ zero=0.85; for tenure, δ zero=0.58. Figures 7 and 8 plot β(δ) across δ values.

Interpretation. These δ values suggest that moderate unobserved confounding—on the order of the observed confounding—could, in principle, explain away the adjusted effects. This is plausible in churn contexts because unmeasured variables such as satisfaction, complaint history, or targeted retention offers can influence both TechSupport adoption and churn. Accordingly, our results should be interpreted as causal under the ignorability assumption, but with awareness that residual confounding remains a credible risk.

Practical mitigation. In applied settings, this motivates two complementary steps. First, enrich the feature set with operational data that proxies key latent drivers (e.g., number of support tickets, downtime incidents, NPS scores). Second, integrate quasi-experimental or experimental variation when possible—such as phased rollouts, eligibility thresholds, or randomized encouragement designs—to strengthen identification. Within the confines of the IBM dataset, we cannot implement these design improvements, but we can quantify sensitivity and document assumptions via the LLM-assisted audit trail.

Limitations

(1) Observational, cross-sectional data. The dataset is a snapshot: we observe tenure at a single time and churn as a one-month indicator. This limits causal claims about future retention dynamics. While tenure is a meaningful retention proxy, causal effects on future survival time require longitudinal observation or explicit survival modeling.

(2) Treatment proxy interpretation. We interpret TechSupport adoption as a service performance upgrade. This is a plausible customer-experience upgrade but it is not a direct measure of network performance engineering. If the policy question concerns infrastructure upgrades that do not change plan features, the IBM dataset is insufficient.

(3) Unobserved confounding. Key drivers of both adoption and churn—such as satisfaction, service incidents, marketing offers, and competitor availability—are not observed. Oster's sensitivity analysis indicates that moderate unobserved selection could affect conclusions. Therefore, the estimates should be used as evidence rather than as definitive proof.

(4) Mediation ambiguity. Variables such as MonthlyCharges and other add-ons can act as mediators or confounders depending on the business process. We targeted the total effect by excluding these variables from X; if the business question is the direct effect holding price constant, the adjustment strategy should be modified and a different estimand should be reported.

(5) LLM specification risk. Although the LLM outputs are auditable, they can still be wrong. We mitigated this by constraining the output schema and by validating the resulting adjustment set through overlap and balance diagnostics. Nevertheless, LLM-assisted specification should be treated as decision support, not automation without review.

(6) External validity. Results are specific to the IBM

Telco dataset and to broadband customers. Effects may differ across providers, regions, and time periods.

Conclusion

This paper demonstrated an LLM-assisted approach to causal attribution for service performance upgrades in churn analytics. The key idea is to separate specification from estimation: the LLM generates a structured, auditable causal specification (DAG, variable roles, adjustment set rationale), and standard causal estimators are then applied with diagnostics and sensitivity analysis.

On the IBM Telco Customer Churn dataset (n=5,517 broadband customers), treating TechSupport adoption as a performance upgrade yields consistent evidence of improved retention after adjustment. Across PSM, IPW, AIPW, DML, and a causal-forest-style DR-learner, TechSupport reduces churn by approximately 7–10 percentage points and increases tenure by approximately 2–4 months. HTE analyses identify segments with especially large benefits, notably month-to-month contracts and senior citizens, and the tenure-increment curve links estimated gains to pricing segments.

While sensitivity analysis indicates that unobserved confounding remains a plausible threat, the overall workflow provides a practical template for organizations: use LLMs to make causal assumptions explicit and reviewable, and use robust estimators and diagnostics to quantify effects with transparency. Future work should incorporate richer operational data and quasi-experimental designs to strengthen identification and to evaluate the reliability of LLM-assisted causal specification at scale.

References

[1] J. Pearl, Causality: Models, Reasoning, and Inference, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[2] M. A. Hernán and J. M. Robins, Causal Inference: What If. Boca Raton, FL, USA: Chapman & Hall/CRC, 2020.

[3] G. W. Imbens and D. B. Rubin, Causal Inference for Statistics, Social, and Biomedical Sciences. Cambridge, U.K.: Cambridge Univ. Press, 2015.

[4] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," Biometrika, vol. 70, no. 1, pp. 41–55, 1983.

[5] E. A. Stuart, "Matching methods for causal inference: A review and a look forward," Stat. Sci., vol. 25, no. 1, pp. 1–21, 2010.

[6] P. C. Austin, "An introduction to propensity score methods for reducing the effects of confounding in observational studies," Multivariate Behav. Res., vol. 46, no. 3, pp. 399–424, 2011.

[7] J. M. Robins, A. Rotnitzky, and L. P. Zhao, "Estimation of regression coefficients when some regressors are not always observed," J. Amer. Stat. Assoc., vol. 89, no. 427, pp. 846–866, 1994.

[8] H. Bang and J. M. Robins, "Doubly robust estimation in missing data and causal inference models," Biometrics, vol. 61, no. 4, pp. 962–973, 2005.

[9] M. J. Funk, D. Westreich, C. Wiesen, T. Stürmer, M. A. Brookhart, and M. A. Hernán, "Doubly robust estimation of causal effects," Amer. J. Epidemiol., vol. 173, no. 7, pp. 761–767, 2011.

[10] V. Chernozhukov et al., "Double/debiased machine learning for treatment and structural parameters," Econometrics J., vol. 21, no. 1, pp. C1–C68, 2018.

[11] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," J. Amer. Stat. Assoc., vol. 113, no. 523, pp. 1228–1242, 2018.

[12] S. Athey, J. Tibshirani, and S. Wager, "Generalized random forests," Ann. Statist., vol. 47, no. 2, pp. 1148–1178, 2019.

[13] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[14] E. J. Oster, "Unobservable selection and coefficient stability: Theory and evidence," J. Bus. Econ. Stat., vol. 37, no. 2, pp. 187–204, 2019.

[15] A. Vaswani et al., "Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 5998–6008.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.

[17] T. B. Brown et al., "Language models are few-shot learners," in Proc. NeurIPS, 2020, pp. 1877–1901.

[18] L. Ouyang et al., "Training language models to follow instructions with human feedback," arXiv:2203.02155, 2022.

[19] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," arXiv:2201.11903, 2022.

[20] X. Wang et al., “Self-consistency improves chain of thought reasoning in language models,” arXiv:2203.11171, 2022.

[21] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” Ann. Statist., vol. 29, no. 5, pp. 1189–1232, 2001.

[22] IBM, “Telco Customer Churn,” IBM Sample Data Sets, 2019. [Online]. Available: https://github.com/IBM/telco-customer-churn-on-icp4d (accessed 2022-12-31).

[23] J. Pearl, M. Glymour, and N. P. Jewell, Causal Inference in Statistics: A Primer. Chichester, U.K.: Wiley, 2016.

[24] M. J. van der Laan and S. Rose, Targeted Learning: Causal Inference for Observational and Experimental Data. New York, NY, USA: Springer, 2011.

[25] J. Grimmer and B. M. Stewart, “Text as data: The promise and pitfalls of automatic content analysis methods for political texts,” Political Anal., vol. 21, no. 3, pp. 267–297, 2013.

[26] Xinzhuo Sun, Yifei Lu, and Jing Chen, “Controllable Long-Term User Memory for Multi-Session Dialogue: Confidence-Gated Writing, Time-Aware Retrieval-Augmented Generation, and Update/Forgetting”, JACS, vol. 3, no. 8, pp. 9–24, Aug. 2023, doi: 10.69987/JACS.2023.30802.

[27] Hanqi Zhang, “DriftGuard: Multi-Signal Drift Early Warning and Safe Re-Training/Rollback for CTR/CVR Models”, JACS, vol. 3, no. 7, pp. 24–40, Jul. 2023, doi: 10.69987/JACS.2023.30703.

[28] Meng-Ju Kuo, Boning Zhang, and Haozhe Wang, “Tokenized Flow-Statistics Encrypted Traffic Analysis: Comparative Evaluation of 1D-CNN, BiLSTM, and Transformer on ISCX VPN-nonVPN 2016 (A1+A2, 60 s)”, JACS, vol. 3, no. 8, pp. 39–53, Aug. 2023, doi: 10.69987/JACS.2023.30804.

[29] Z. Zhong, M. Zheng, H. Mai, J. Zhao, and X. Liu, “Cancer image classification based on DenseNet model,” Journal of Physics: Conference Series, vol. 1651, no. 1, p. 012143, 2020.