LARGE LANGUAGE MODELS AS PROXY DESIGNERS FOR PROXIMAL CAUSAL INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Proximal causal inference (PCI) identifies causal effects under unmeasured confounding using two proxies for the latent confounder: a treatment control Z and an outcome control W. Its practical bottleneck is the step before estimation, namely partitioning the covariates into Z, W and observed confounders X so that the identification conditions hold, a judgment that needs fluency in both PCI and the application domain. We show that large language models (LLMs) are increasingly capable of making this judgment. From the PCI conditions, a brief study context, and a description of each covariate, an LLM proposes what the latent confounder represents and assigns every covariate a role; a stratum-targeted outcome-bridge objective then yields conditional effects from one global bridge. We evaluate on three real cohorts with external checks. On a 10M-sample food-delivery corpus with a separate randomized price experiment, three LLM partitions give 12.0–21.2% error (relative to the randomized experiment), two of these are below and one is tied with the best baseline. In the Framingham Offspring study, every LLM partition that admits an estimate lands inside the range randomized trials report, where a third of random partitions and the published proximal split do not. In the SUPPORT study, all eight LLMs give estimates within the published confidence intervals and one reproduces the expert allocation exactly. For the (cardiovascular) Framingham and SUPPORT datasets, we also compare LLM covariate partitions to those from multiple Board-certified cardiologists.

1 INTRODUCTION

The causal effect of a treatment A on an outcome Y is confounded in observational data whenever a common cause U is unmeasured. Outcome regression, inverse probability weighting (IPW) and double machine learning (DML) all assume every such cause has been measured (Glass et al., 2013), which is untestable and frequently violated (Pearl et al., 2016; Rubin, 1974). *Proximal causal inference* (PCI) identifies the effect without such an assumption, given a treatment proxy Z and an outcome proxy W that carry information about U (Kuroki & Pearl, 2014; Miao et al., 2018; Tchetgen Tchetgen et al., 2020). This is done using a *causal bridge function* of the observed proxies that is shown to yield $\mathbb{E}[Y \mid \mathrm{do}(A = a)]$ without directly measuring U (Miao et al., 2018; Cui et al., 2024; Meng et al., 2025).

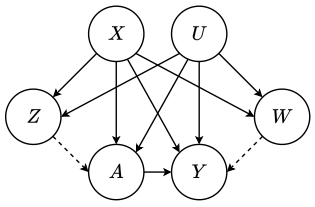

Figure 1: Graphical model for proximal causal inference. U unobserved, X observed; Z, W are observed treatment and outcome proxies, respectively. The edges $Z \to A$ and $W \to Y$ are dashed because they are permitted but not required.

What limits PCI in practice is not estimation – the bridge is learnable by a neural network (Xu et al., 2021; Kompa et al., 2022; Meng et al., 2025) – but the step before it, namely partitioning the covariates into Z, W and X so that the identification conditions hold *and* the domain's causal structure is respected (X are observed confounders).

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

This work We show that state-of-the-art large language models (LLMs) can effectively partition W, Z and X across a range of application domains. Given only the treatment and outcome definitions and each covariate's name and description, the model elucidates a meaning for U from context and classifies every covariate by checking the PCI conditions without manually specifying U nor setting per-variable role labels. Everything downstream is held fixed: an existing neural outcome bridge (Meng et al., 2025), one set of hyperparameters per dataset and no per-model tuning, so only the partition changes between the entries of any table. Because one cannot know in advance which LLM to trust, we ask eight models from four vendors on the clinical cohorts and the three available for the waybill corpus (§5.4), report every one of them, and judge the procedure by how the partitions perform across models and how consistent they are, never by the best arm alone.

Contributions

- **Modern LLMs can do proxy design for PCI.** From a statement of the identification conditions and a description of each variable, we demonstrate that modern LLMs can infer a form for the latent confounder and partitions the covariates into Z, W and X. This is demonstrated on three real-world datasets.
- **PCI validated on a large real-world dataset.** On a 10M-sample commercial corpus with 181 covariates and a separate randomized experiment for comparison, bridges fitted on three LLM partitions recover the population effect with high accuracy.
- **Stratum-based learning of conditional effects.** Conditional effects under PCI have been learned with a dedicated second-stage learner (Sverdrup & Cui, 2023). We instead learn by stratum: one global bridge returns the effect within any stratum of the observed covariates with no further inverse problem (Proposition 1, a corollary of the identification theorem).

2 Related Work

Proximal causal inference PCI originates in the proxy and negative-control literature (Kuroki & Pearl, 2014; Miao et al., 2018; Tchetgen Tchetgen et al., 2020; Miao et al., 2023). The outcome bridge of Miao et al. (2018) has been extended to doubly robust estimation (Cui et al., 2024) and analysed under misspecification (Mastouri et al., 2021). What this literature has in common is that the partition into Z, W and X is *assumed*: it is supplied by a domain expert or derived from marginal associations, and how to obtain it is not their subject of study.

Language models in causal inference Text representations have served causal estimation and testing (Feder et al., 2022; Egami et al., 2022; Knox et al., 2022), typically to *measure* a confounder from text, and Chen et al. (2024) infer proxy *values* from unstructured text. A separate line asks LLMs to reason about causal structure from variable descriptions: pairwise causal direction and counterfactual questions (Kıcıman et al., 2023), and LLM-assisted causal discovery, in which the model proposes or constrains edges of a graph (Long et al., 2023; Ban et al., 2023; Vashishtha et al., 2023). Deciding that a covariate is an observed confounder is a decision of that kind, and on the Framingham Offspring study, it is the decision that carries the result (§5.2). What is new here is the setting of the proximal roles: which covariates can serve as Z, which as W, and the independence between them that the identification conditions require, a decision that matters on SUPPORT (§5.1).

Conditional effects under PCI Sverdrup & Cui (2023) study conditional average treatment effects in the proximal setting: their P-learner targets the fully conditional effect $\tau(X)$ with a dedicated second-stage learner built on estimated bridge functions, and Qi et al. (2024) learn individualized treatment regimes. Our approach is complementary and learns *by stratum*. A single global bridge returns the effect within any stratum $C = \kappa(X)$ of the observed covariates with no additional inverse problem (Proposition 1), and the bridge is trained with an objective that decomposes the Fredholm residual by stratum and weights the strata equally (7), which to our knowledge has not been done for a proximal bridge.

3 Proximal Bridges

Observed and unobserved confounders X and U are assumed to jointly drive A and Y (Pearl et al., 2016). Proxy (control) measurements (Tchetgen Tchetgen et al., 2020) in principle make U tractable: a *treatment* proxy Z, on which A may depend, and an *outcome* proxy W, on which Y may depend; the edges $Z \to A$ and $W \to Y$ are optional (see Figure 1). The treatment takes values

in $\mathcal{A}$: a finite grid of real-valued doses in §5.4, binary in §5.1 and §5.2. Outcome Y is real-valued or binary, and $Y(a)$ denotes the potential outcome given $a \in \mathcal{A}$.

Assumptions As in prior work (Tchetgen Tchetgen et al., 2020; Xu et al., 2021; Cui et al., 2024), we assume consistency ($Y = Y(A)$ almost surely, with X, Z, W non-descendants of A), positivity (every $a \in \mathcal{A}$ is attainable within levels of (U, X)), and:

- $A1$ $Y(a) \perp\!\!\!\perp A \mid U, X$ for $a \in \mathcal{A}$ (latent ignorability);
- $A2$ $W \perp\!\!\!\perp (A, Z) \mid U, X$ (outcome control independence);
- $A3$ $Z \perp\!\!\!\perp Y \mid U, X, A$ (treatment control independence);
- $A4$ for any a, almost every x, and any square-integrable g, $\mathbb{E}[g(U, x) \mid Z, X = x, A = a] = 0$ a.s. if and only if $g(U, x) = 0$ a.s. (completeness).

Outcome bridge The outcome bridge $h(W, a, X)$ is defined by

$$\mathbb{E}[Y \mid X, U, A = a] = \mathbb{E}[h(W, a, X) \mid U, X, A = a]. \tag{1}$$

Under $A2$ and $A3$, averaging (1) over $p(U \mid X, Z, A = a)$ yields a Fredholm integral equation of the first kind in observables (Miao et al., 2018; Tchetgen Tchetgen et al., 2020):

$$\mathbb{E}[Y \mid X, Z, A = a] = \mathbb{E}[h(W, a, X) \mid X, Z, A = a]. \tag{2}$$

Theorem 1 (Miao et al. (2018))**.** *Suppose the assumptions of §3 hold and that some square-integrable $h(W, a, X)$ solves* (2) *almost surely. Then every such solution also satisfies* (1)*, and*

$$\mathbb{E}[Y \mid \mathrm{do}(A = a)] = \mathbb{E}[h(W, a, X)]. \tag{3}$$

Learning The estimator is taken from Meng et al. (2025) and is not a contribution. We learn the kernel of (2) as a generative model $p_\phi(W \mid Z, X, A)$ by conditional maximum likelihood on the n training samples, giving $\hat{\phi}$, which is then held fixed; and we fit h_θ by minimizing the Monte Carlo Fredholm residual, drawing for each sample i a set of J values $\{w_i^{(j)}\}_{j=1}^{J} \sim p_{\hat{\phi}}(W \mid z_i, x_i, a_i)$ from the fitted model,

$$\mathcal{L}^{\mathrm{OB}}(\theta; \hat{\phi}) = \frac{1}{n} \sum_i \Big(Y_i - \tfrac{1}{J} \sum_j h_\theta(w_i^{(j)}, a_i, x_i) \Big)^2. \tag{4}$$

Both clinical cohorts have time-to-event outcomes: on Framingham the squared loss is replaced by the Cox partial likelihood of a bridge linear in W, as in Meng et al. (2025), and the estimate is a hazard ratio; on SUPPORT the outcome is days of survival within 30, treated as continuous, as in the reference analysis.

Stratum-level effects We also report the conditional effect

$$\tau(c; a, a_0) = \mathbb{E}\big[Y \mid \mathrm{do}(A{=}a),\, C = c\big] - \mathbb{E}\big[Y \mid \mathrm{do}(A{=}a_0),\, C = c\big], \tag{5}$$

where $a_0 \in \mathcal{A}$ is a reference treatment level (the control dose $\delta = 0$ in §5.4, no treatment in the clinical cohorts), the stratum variable $C = \kappa(X)$ is a measurable function of the observed covariates taking values in a finite set $\mathcal{C}$, $c \in \mathcal{C}$ is one of its levels, and $n_c = \#\{i : \kappa(x_i) = c\}$; this is a coarser target than the fully conditional $\tau(X)$ of Sverdrup & Cui (2023).

The following is a direct corollary of Theorem 1; we state it because it is what makes stratum effects free once the bridge is fitted.

Proposition 1 (Stratum effects)**.** *Under the conditions of Theorem 1, for any $C = \kappa(X)$ and any c with $P(C = c) > 0$,*

$$\mathbb{E}\big[Y \mid \mathrm{do}(A{=}a), C = c\big] = \mathbb{E}\big[h(W, a, X) \mid C = c\big]. \tag{6}$$

Proof. By consistency and $A1$, $\mathbb{E}[Y(a) \mid U, X] = \mathbb{E}[Y \mid U, X, A = a]$, which by Theorem 1 and (1) equals $\mathbb{E}[h(W, a, X) \mid U, X, A = a]$, and by $W \perp\!\!\!\perp A \mid U, X$ equals $\mathbb{E}[h(W, a, X) \mid U, X]$. Taking $\mathbb{E}[\,\cdot \mid C = c]$ of both ends and applying the tower property gives (6). □

Hence $\hat{m}_c(a) = n_c^{-1} \sum_{i:C_i=c} \hat{h}(w_i, a, x_i)$ gives $\hat{\tau}(c; a, a_0) = \hat{m}_c(a) - \hat{m}_c(a_0)$ with no further inverse problem, for any stratification chosen after fitting; $C = \kappa(X)$ holds whenever strata are drawn from the covariates the LLM assigned to X.

Stratum-balanced training The loss in (4) is a sample-size-weighted sum of per-stratum residuals; we also train with every stratum given equal weight, the objective (7) of Appendix C.1, a group reweighting that holds the bridge to account within each stratum.

4 LLM-DRIVEN PROXY SELECTION

Task formulation Given (i) the treatment A and outcome Y definitions, (ii) a name and description for each of p covariates, and (iii) the PCI DAG and assumptions of §3, the LLM partitions the covariates into Z, W, and X. It is *never* told the latent confounder U (never known precisely), nor any variable's role: inferring what U plausibly drives both A and Y is part of the task.

Prompt structure A *system* prompt casts the LLM as an expert in PCI and the application domain, that reasons step-by-step from the independence assumptions. A *user* prompt supplies the inputs above and a brief study context. The LLM states its inferred U, checks $A2$/$A3$ for each covariate, and outputs a JSON partition (examples in Appendix A).

Decision order A covariate is X if it has direct causal paths to both A and Y, at least one bypassing U; otherwise it is Z if it reflects U, predicts A and satisfies $A3$, or W if it reflects U, predicts Y and satisfies $A2$; what fits neither is discarded (D).

Model arms The clinical experiments use eight models from four vendors, each vendor's current model paired with one of the previous generation so that the effect of capability can be read within vendor: Claude Opus 5 and Claude Haiku 4.5 (Anthropic), Gemini 3.1 Pro and Gemini 3.1 Flash-Lite (Google), GPT-5.6 and GPT-4 Turbo (OpenAI), Grok 4.6 and Grok 4.20 (xAI); we call the first of each pair the *newer* model. API identifiers and request dates are in Appendix B.

Voting protocol Each arm is tried $m = 5$ times, and the partition to which it is credited is the majority vote *per covariate*. Every response is retained, including any with an empty Z or W, and no sampling parameter is set, so each vendor's stochastic default decoding is the only source of variation. Because the vote is per covariate, the voted partition may be one no response proposed, and it may violate completeness ($A4$, an empty Z or W) even when most responses satisfy it.

5 EXPERIMENTS

5.1 RIGHT HEART CATHETERISATION (SUPPORT)

SUPPORT (Connors et al., 1996) is the standard example of the PCI literature: Tchetgen Tchetgen et al. (2020) named four proxies among ten candidate variables and split them, and later work has reused such allocation unchanged (Cui et al., 2024; Qi et al., 2024; Sverdrup & Cui, 2023).

Setting $n = 5{,}735$ intensive-care admissions; A is right heart catheterisation within the first 24 hours (38.1% treated), Y days from admission to death, censored at 30. Of 53 clinical covariates, ten are blood chemistry and blood gases from a single first-day draw, error-prone snapshots of one physiological state; the models saw those ten, told that the other 43 are conditioned on as X. The published allocation is $Z = (\texttt{pafi1}, \texttt{paco21})$ and $W = (\texttt{ph1}, \texttt{hema1})$. The arms are the eight of §4, five replicates each. The published allocation is a reference point, not a ground truth: its authors call it a practical strategy, and the yardstick is whether a model reasons as an expert does.

Agreement with the expert All 40 replicates returned an admissible partition: parseable, every variable assigned, Z and W non-empty. Claude Opus 5 reproduces the published allocation on all ten variables, in four of its five replicates individually. The match is not recognition of a well-known dataset: with variable names replaced by shuffled neutral codes and the study's name and counts removed, all five replicates return it. Across arms the partitions are close in structure: every arm keeps the two blood gases `pafi1` and `paco21` as treatment proxies and haematocrit as an outcome proxy, six of eight place pH beside the blood gases, and the arms differ mainly in whether the five chemistry values are conditioned on as X or used as further outcome proxies (every vote is in Table 11, Appendix C.3).

Where the other arms depart, they argue from the identification conditions. Six of eight arms move `ph1` from W to Z, beside `paco21`: arterial pH and $PaCO_2$ are tied by the Henderson–Hasselbalch relation, so the published allocation asks $Z \perp\!\!\!\perp W \mid U, X$ to hold across a chemical identity, and

allocation and estimator	$\lvert Z\rvert/\lvert W\rvert$	effect (days)	interval
Tchetgen Tchetgen et al. (2020), least squares	—	-1.25	$[-1.80, -0.70]$
Tchetgen Tchetgen et al. (2020), proximal 2SLS	2/2	-1.80	$[-2.64, -0.96]$
Sverdrup & Cui (2023), proximal DR kernel	2/2	-1.66	$[-2.19, -1.13]$
Claude Opus 5 (= Tchetgen Tchetgen et al., 2020), ours	2/2	-1.65	$[-1.88, -1.57]$
Gemini 3.1 Flash-Lite	3/3	-1.59	$[-1.64, -1.52]$
GPT-4 Turbo	3/5	-1.55	$[-1.80, -1.20]$
GPT-5.6	2/7	-1.46	$[-2.23, -0.79]$
Claude Haiku 4.5	3/6	-1.34	$[-1.95, -0.30]$
Gemini 3.1 Pro	3/7	-1.31	$[-1.89, -0.83]$
Grok 4.6	2/6	-1.24	$[-2.04, -0.82]$
Grok 4.20	4/5	-1.24	$[-1.75, -0.76]$
all LLM arms, runs pooled	varies	-1.51	$[-1.95, -0.80]$
random allocations, runs pooled	varies	-1.52	$[-3.00, -0.40]$

Table 1: SUPPORT: effect of right heart catheterisation in days of 30-day survival. Upper block: published estimates for the allocation of Tchetgen Tchetgen et al. (2020) (least squares without proxies, proximal two-stage least squares, and the doubly robust kernel estimator of Sverdrup & Cui, 2023), with the 95% confidence interval implied by the published standard error. Lower block: one estimator, only the allocation changing, median [min, max] over 20 runs. Last two rows: every run of the eight LLM arms (160) and of thirty random allocations (600) pooled, the interval holding the central 95% of runs.

the data show the binding (`paco21` predicts `ph1` with a first-stage $F \approx 1{,}100$). A rule that ranks marginal associations cannot register this; the models, given that condition in words, raise it unprompted (Appendix A).

Effect estimates The published two-stage least squares (Tchetgen Tchetgen et al., 2020), for the effect in days of 30-day survival, requires $|Z| \geq |W|$ and returns nothing for six of the eight voted allocations, so we keep the estimand and learn the first stage $\mathbb{E}[W \mid A, Z, X]$ with the mixture-density network of §3. The second stage is the linear bridge $h(a, x, w) = \eta_0 + \eta_a a + \eta_x' x + \eta_w' w$ solved from $\mathbb{E}[Y \mid A, Z, X] = \eta_0 + \eta_a A + \eta_x' X + \eta_w' \hat{\mathbb{E}}[W \mid A, Z, X]$ with the learned $\hat{\mathbb{E}}[W \mid A, Z, X]$ as instrument, so that a linear first stage recovers the published estimator exactly; η_a is the effect in days. On the published allocation we obtain -1.65 days, beside the -1.66 (0.27) of Sverdrup & Cui (2023) from a doubly robust kernel estimator: estimators sharing neither function class nor fitting method agree to a hundredth of a day (Table 1). This agreement validates the estimator on the published allocation, not the LLM; Claude Opus 5 inherits it because its vote is that allocation. What speaks for the models as a group is that all eight arms' medians lie inside the confidence interval of each published estimate, from -1.24 to -1.65 days, a spread of 0.41 days.

Random allocations are not confined like this. With ten candidate variables, all passed to the estimator in some role, a random allocation is often not far from a voted one and the pooled medians agree; even so, random allocations are far less dependable. Their thirty estimates run from -2.23 to -0.81 days, a spread of 1.41 days against the LLM arms' 0.41; run by run, the central 95% of their runs spans 2.60 days against 1.15 for the LLM arms, and five of their 600 runs even give a beneficial effect, which none of the 160 LLM runs does. Eight of the thirty give an effect more harmful than any published estimate, which no LLM arm does. Further controls, including the exchange of Z and W, which on this cohort moves six of the eight allocations by 0.3 to 1.7 days, are in Appendix C.3.

5.2 Framingham Offspring

Setting The second experiment is for an observational study, not a randomized clinical trial (RCT), however, there is external evidence about the treatment effect. A is statin use at Exam 6 (9.7% treated); Y is all-cause mortality (617 events among $2{,}404$ training subjects). Observational analysis is confounded by indication—a Cox fit on treatment alone returns $\mathrm{HR} = 1.70$ in these data, the wrong direction—because statins go to patients already at cardiovascular risk. That risk is the latent U, and the LLM is never told what it is. There are 16 clinical variables (32 covariates once spline bases are included), and the eight model arms of §4.

Reference, estimator and controls Framingham Offspring is observational, but the effect estimated from it, that of statin therapy on all-cause mortality, has been quantified from RCTs in other

partition	$\lvert Z\rvert/\lvert W\rvert/\lvert X\rvert$	HR	runs	in range
HOPE-3 trial (Yusuf et al., 2016)	—	≈ 0.75	—	—
Claude Opus 5	9/12/11	0.76	[0.73, 0.83]	yes
Grok 4.6	6/9/17	0.78	[0.73, 0.86]	yes
Gemini 3.1 Pro, Grok 4.20	9/9/14	0.81	[0.77, 0.88]	yes
GPT-4 Turbo	9/4/19	0.86	[0.80, 0.93]	yes
Gemini 3.1 Flash-Lite	4/12/15	0.89	[0.81, 0.98]	yes
Meng et al. (2025) split, same pipeline	16/16/0	0.91	[0.79, 1.00]	no
uniform random (30)	varies	0.84 [0.44, 1.06]	—	20 of 30
size-matched random (30)	9/12/11	0.75 [0.43, 1.13]	—	18 of 30
all LLM arms, runs pooled	varies	0.82	[0.68, 1.03]	68%
uniform random, runs pooled	varies	0.82	[0.49, 1.14]	53%
size-matched random, runs pooled	9/12/11	0.75	[0.43, 1.14]	44%

Table 2: Framingham: statin→all-cause-mortality hazard ratio, one estimator. LLM and Meng et al. (2025) rows: median and interquartile range over 20 runs. Random rows: median [min, max] over the 30 partitions' medians, and how many of the 30 fall in the range $[0.70, 0.90]$. Pooled rows: median and central 95% over every run (120 LLM runs; 600 per control), and the share of runs in the range. Gemini 3.1 Pro and Grok 4.20 voted identically.

populations, which bracket the effect. At the high-risk end, the Scandinavian Simvastatin Survival Study, a placebo-controlled trial in patients with coronary heart disease, reports a relative risk of death of 0.70 (Scandinavian Simvastatin Survival Study Group, 1994). At the general-population end, the Cholesterol Treatment Trialists' meta-analysis of 26 trials reports 0.90 per mmol/L reduction in LDL cholesterol (Cholesterol Treatment Trialists' (CTT) Collaboration, 2010), and since standard statin therapy lowers LDL by about 1 mmol/L, 0.90 is the effect of a typical course in a broad trial population. The source analysis of this cohort (Meng et al., 2025) takes HR ≈ 0.75 as reference from the HOPE-3 RCT (Yusuf et al., 2016). None of these is the Offspring population or its treatment contrast, so we do not score distance to a point but ask whether an estimate falls in $[0.70, 0.90]$, the interval these trial values bracket, taken as an external plausibility range. The estimator is that of Meng et al. (2025), held fixed across arms: a bridge linear in W fitted by Cox partial likelihood, the hazard ratio reported as the median over 20 runs. There are two controls of thirty random partitions each: *uniform*, each variable's role drawn uniformly from Z, W and X, and *size-matched*, which holds the role counts of Claude Opus 5 and randomises only which variable receives which role. We also run the split of Meng et al. (2025), which divides all 32 covariates $16/16$ into Z and W by association.

Results Every arm declares between 11 and 29 of the 32 covariates to be observed confounders, where the split of Meng et al. (2025) declares none. The partitions agree on the confounders and differ on the proxies: all eight arms place age, sex, smoking, diabetes and the disease-history indicators in X, six of eight use the exam-5 lipid panel (total cholesterol, HDL, triglycerides) as treatment proxies, and blood pressure, body mass index and fasting glucose are outcome proxies for some arms and confounders for others (every vote is in Table 9, Appendix C.2). Two of the eight votes violate completeness—GPT-5.6 with $W = \emptyset$, Claude Haiku 4.5 with $Z = \emptyset$—and carry no estimate. Gemini 3.1 Pro and Grok 4.20 voted identically, so the six remaining arms are five distinct partitions. Their hazard ratios lie between 0.76 and 0.89, all inside the range (Table 2); the three newer models give 0.76, 0.78 and 0.81 and the older models 0.81, 0.86 and 0.89. The split of Meng et al. (2025) run through the same pipeline gives 0.91, outside the range; the conventional Cox fits reported by the source analysis, unweighted, inverse-probability weighted and overlap weighted, are 1.72, 1.17 and 0.90.

Random partitions are not confined like this. With sixteen candidate variables, all handed to the estimator in some role, a random partition is often not far from a voted one and the control medians, 0.84 and 0.75, are unremarkable; what the LLM partitions buy is not central accuracy but dispersion: the uniform draws run from 0.44 to 1.06, a factor of 2.4, and ten of the thirty leave the range; with the role counts held at those of Claude Opus 5, random partitions run from 0.43 to 1.13 and twelve of thirty leave it, ten of them by over-correcting below 0.70, where the LLM partition of that size gives 0.76. Run-to-run variation is not small on this estimator and is the same for LLM and random partitions (Table 2); the difference is where the partitions sit: pooling every run, 68% of LLM runs

fall in the range against 53% and 44% for the two controls, and no LLM run falls below 0.60 where 8% and 17% of random runs do.

Which decision matters here Keeping each LLM partition's X and reassigning its proxies to Z and W at random, thirty draws, gives hazard ratios from 0.75 to 0.94 with 27 of 30 in range, and exchanging Z and W outright moves the estimate by factors of 0.95 to 1.07, within run-to-run noise (Appendix C.2). On this cohort, what the LLM contributes is the choice of observed confounders, and proxy assignment matters little. That is confounder selection, a task LLMs are known to do (Kıcıman et al., 2023; Vashishtha et al., 2023); proxy assignment is consequential on SUPPORT.

5.3 Domain experts without PCI training

The premise of this work is that the partition step needs someone fluent in both the identification conditions and the domain, and that such people are rare. The two clinical cohorts let us test the premise directly: we asked six board-certified cardiologists at an academic medical center to carry out the same task the models were given. This proved harder than expected. In a briefing we presented the setting of Figure 1 and the conditions of §3, and the clinicians could not apply them: the conditions are statements of conditional independence given a latent quantity U, and there is no clinical vocabulary in which such a statement can be checked variable by variable. What the clinicians could answer was a simpler pair of questions, and we built the survey around it.

Instrument Each cardiologist, working independently on a web form, was given the treatment, the outcome, the study context and the name and description of each covariate, and placed every covariate as a point $(z, w) \in [0, 1]^2$, where z is the covariate's influence on the treatment decision and w its influence on the outcome. The natural reading of the grid into roles is the one earlier association-based splits use (Meng et al., 2025), now supplied by clinicians rather than by correlations: a covariate that influences both is an observed confounder X; one that influences the outcome but not the treatment is an outcome proxy W; one that influences the treatment but not the outcome is a treatment proxy Z. Figure 2 shows the mean placements on SUPPORT.

The empty quadrant No covariate was placed in the treatment-proxy quadrant, high in z and low in w; the clinicians were unwilling to say that a variable guides the decision to catheterise while not bearing on survival. The same quadrant is empty on Framingham. This is not an error on their part. It is what a treatment proxy is: Z guides treatment because it carries information about U, and U affects the outcome, so a treatment proxy does influence the outcome, through U. Condition $A3$ asks for no *direct* path from Z to Y once U is fixed, and a question about influence cannot separate the direct path from the one through a quantity the clinician has no name for. The role that PCI needs most, and the one that the published allocation and the models both fill, is therefore the one that clinical judgment on its own does not produce. The same gap is why association-based data splits assign every variable to Z or W and none to X (§5.2): influence, whether scored by a correlation or by an expert, is not the quantity the identification conditions are about.

Agreement with the models Table 3 asks a narrower question: whether the covariates each model places in Z are ones the clinicians see as influencing treatment, and whether its W are ones they see as influencing the outcome. For each cardiologist and model, the ten covariates are scored by z with the model's Z as the positive class (all others negative), and by w with the model's W as positive; the area under the ROC curve is averaged over the six cardiologists, and the table reports the two areas and their harmonic mean. The score is one axis at a time, so a covariate placed high on both axes counts towards agreement on Z and on W alike; it measures how far a partition resembles the influence heuristic, not how far it satisfies the identification conditions, and with ten covariates and two to seven positives a single covariate can move an area by more than a tenth. Read that way, the pattern is the expected one. The published allocation and Claude Opus 5, whose vote reproduces it, have the lowest agreement of any admissible partition, and they part from the clinicians on exactly the variables where the conditions bite: the blood gases and pH, which the clinicians rate as influencing both treatment and outcome and which the PCI reasoning of §5.1 places as proxies. Agreement is not monotone in capability: Claude Haiku 4.5 and the majority vote across the eight arms, which coincide, agree most with the clinicians, while GPT-4 Turbo agrees least and is also the least self-consistent arm (Appendix B). The majority-vote row is shown only for this comparison and is not used elsewhere in the paper.

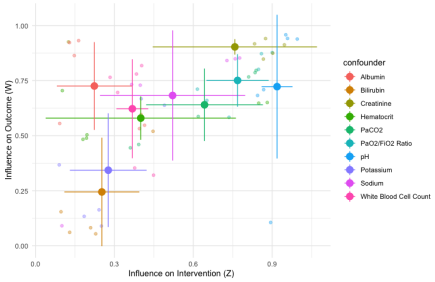

Figure 2: Placement of the ten SUPPORT blood-panel covariates by six cardiologists: horizontal axis, influence on the intervention (z); vertical axis, influence on the outcome (w). Error bars are standard deviations and small dots are individual placements.

allocation	Z AUROC	W AUROC	H_{AUROC}
Claude Opus 5	0.563	0.813	0.665
Gemini 3.1 Pro	0.714	0.952	0.816
GPT-5.6	0.714	1.000	0.833
Grok 4.6	0.875	0.875	0.875
Claude Haiku 4.5	0.875	0.952	0.912
Gemini 3.1 Flash-Lite	0.619	0.952	0.750
GPT-4 Turbo	0.560	0.571	0.566
Grok 4.20	0.800	0.833	0.816
majority vote, eight arms	0.875	0.952	0.912
Tchetgen Tchetgen et al.	0.563	0.813	0.665

Table 3: Agreement between the cardiologists' placements and each allocation on SUPPORT: area under the ROC curve, averaged over the six cardiologists, scoring the ten covariates by z against the allocation's Z and by w against its W; H_{AUROC} is the harmonic mean.

What the survey shows is the premise the paper starts from. Cardiologists can say which variables matter and to what; whether a variable can serve as a proxy under the identification conditions is a question they could not be brought to answer, and the association-shaped answer they could give omits the treatment proxies entirely.

5.4 Food-Delivery Waybill (10M)

A food-delivery platform's waybill corpus of about 10 million observational samples with 181 covariates is paired with a separate randomized price experiment of 3.6M orders. The observational corpus is the training data; the randomized experiment is used only to assess the estimates, and none of its orders is in the training data.

Setting and data The treatment A is the platform's real-time price adjustment and the outcome Y is whether a rider accepts the waybill; all three LLMs independently read the latent U as zone-level supply-demand tension. The experiment perturbs the price over 18 levels δ, decreases as well as increases of well under one yuan on a fixed grid, against a shared $\delta{=}0$ control, and $\tau(\delta)$ is the difference in mean acceptance from that control: a dose–response curve, not a single contrast. At the data owner's request the dose values and the absolute effects are withheld, and effects are reported only relative to the experimental effect. The corpus is accessible only within the company's network, and the three models available there were Claude Opus 4.8, GPT-4.1 and Gemini 2.5 Flash, called through their vendors' APIs, each run at ten seeds.

Error measure We report the relative error $\mathrm{err} = \dfrac{\sum_{\delta\in\mathcal{D}}\left|\tau(\delta)-\hat{\tau}(\delta)\right|}{\sum_{\delta\in\mathcal{D}}\left|\tau(\delta)\right|}$, over the dose grid $\mathcal{D}$, with $\hat{\tau}(\delta)$ the estimate. The average treatment effect (*ATE*) applies err to the population effects whereas the average conditional average treatment effect (*avg-CATE*) applies it within each of the five city tiers and averages over tiers, unweighted.

Partitions, estimator and baselines Of the 181 covariates, the voted partitions place 27, 66 and 66 in Z, W and X (Opus 4.8), 50, 37 and 81 (GPT-4.1), and 5, 7 and 160 (Gemini 2.5 Flash), discarding the rest as mediators, treatment components or identifiers. Each feeds the two-stage bridge pipeline of Meng et al. (2025), trained with the stratum-balanced loss (7) over the five city tiers; the effect at dose δ is the fitted bridge's contrast against $\delta = 0$, as in the trial. Hyperparameters are identical across arms and untuned, and model selection never touches the experiment. The baselines, all assuming strong ignorability, are an *S-Learner*, *IPW/GPS* (Hirano & Imbens, 2004) and cross-fitted partialling-out *DML* (Robinson, 1988), each given 174 covariates, more than any bridge is fitted on. The three partitions agree on what to condition on and disagree on what to use as a treatment proxy: 45 covariates are in X for all three, every outcome proxy of the smaller partitions is an outcome proxy of Opus 4.8, whereas the three treatment-proxy sets are disjoint (Appendix C.1). Model settings, hyperparameters and the ten seeds are in Appendix C.1.

Method	ATE (↓)	avg-CATE (↓)
IPW/GPS	32.1	41.8
S-Learner (NN)	21.1	39.3
DML (partialling-out)	23.0	**33.9**
PCI, GPT-4.1 partition	21.2 [14.6, 53.0]	40.1 [37.2, 58.3]
PCI, Gemini partition	19.8 [16.3, 86.3]	38.8 [35.9, 87.3]
PCI, Opus partition	**12.0** [11.8, 17.3]	37.1 [35.5, 39.8]
PCI, Opus partition, global loss	16.0 [11.7, 45.8]	39.1 [37.7, 52.9]

Table 4: Waybill: relative error err against the randomised experiment as % (lower is better). PCI rows: median [range] over ten seeds. The last row is the Opus partition trained with the global loss (4) instead of the stratum-balanced loss, five runs. Every LLM partition we ran is shown.

Population effect Table 4 reports every method. Across the three LLM partitions the median error runs from 12.0% to 21.2%: two are below the best baseline's 21.1%, the third is within 0.1 points of it, and all are below the other two baselines (23.0% and 32.1%). The partitions differ in how far they go: the Opus partition's error is 12.0% and is below the best baseline in all ten runs, whereas the Gemini and GPT-4.1 partitions, at 19.8% and 21.2%, are level with it. The experimental effects are themselves estimates: given their sampling noise, which we estimate from the experiment alone, an estimator that recovered the smooth dose–response exactly would still score about 12%. The Opus partition is at that floor, so no method could do visibly better against this experiment. All six methods recover the shape of the dose–response; they differ in magnitude, which the ignorability baselines understate (Appendix C.1). Bridges occasionally diverge (two GPT-4.1 runs and one Gemini run), and the ranges include them.

Stratum effects Proposition 1 makes the effect within each city tier available from the same bridge. The avg-CATE errors of all methods lie between 34% and 42% (Table 4); the experiment's noise floor at the stratum level is 21–51% by tier, 31% on average, so at this level the methods are within its resolution of one another.

Stratum-balanced training To isolate the objective, we retrain the Opus bridge under the global loss (4) and under the balanced loss (7), five runs each (last row of Table 4); this is a separate five-run pool, which is why its balanced-loss figures differ slightly from the ten-run Opus row. Balanced training lowers the median ATE error from 16.0% to 12.0% and the median avg-CATE error from 39.1% to 36.8%. The gain holds under two other stratifications (Table 8, Appendix C.1), and balanced training is the more stable: the ATE error exceeds 20% in two of the five global runs and in one of the fifteen balanced runs across the three stratifications. The comparison with the baselines does not rest on the objective: under the global loss the Opus partition's 16.0% is still below the best baseline's 21.1%. The baselines are not stratum-reweighted; whether an S-Learner would gain from the same reweighting we could not test on this corpus.

6 Discussion and Conclusion

Language models can do the step of proximal inference that has needed a person fluent in both the identification conditions and the domain; since one cannot know in advance which model to trust, we read the evidence across models. **Waybill:** PCI with LLM partitions recovers the effect a randomised experiment measures, three partitions giving 12.0–21.2% error, two below the best ignorability baseline and the third tied with it. **Framingham Offspring:** the partitions yield plausible estimates where chance does not (all six in the trial-informed range, a third or more of random partitions outside it), through the choice of confounders. **SUPPORT:** the partitions agree with expert reasoning, all eight estimates inside the published confidence intervals and one allocation matching the expert's exactly, where random allocations are twice as dispersed. Training the bridge by stratum improves the population estimate as well, and the newer models are the more self-consistent as shown in Appendix B. Six cardiologists asked to carry out the task on the Framingham and SUPPORT cohorts could not be brought to the identification conditions, and the influence-based placements they could give leave the treatment-proxy role empty (§5.3). **Limitations:** the clinical cohorts have no ground truth, so a partition is judged by the estimate it yields; completeness (A4) is tested only by non-emptiness of Z and W; the Waybill corpus is proprietary; and the partitions are tied to the model versions queried, so the ordering of vendors should not be expected to persist.

AI USE STATEMENT

In this work, we used generative AI tools to edit the text, to create and edit analysis code, and to implement the control experiments and supplementary analyses of Appendix C and to provide feedback on them; large language models are also the object of study, and their partitions are obtained as described in §4. We have not used generative AI tools to generate data, to design the main experiments, or to formulate or prove mathematical claims; the remaining tasks with required disclosure are not applicable to this work. All AI-assisted code and analyses were verified by the authors, and all AI-assisted text was reviewed and revised by the authors. We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

For the two public cohorts we release every prompt, every model response, the voted partitions, the random, size-matched, X-fixed, role-swap and perturbation control partitions, and the estimation code; the model identifiers and request dates are in Appendix B and excerpted responses in Appendix A. The proof of Proposition 1 is in §3. The waybill corpus and its randomised experiment are proprietary and cannot be released; the estimator, objective, baselines and error measures are specified in §5.4 and Appendix C.1, and all effects are reported relative to the experimental effect at the data owner's request.

REFERENCES

Taiyu Ban, Lyuzhou Chen, Derui Lyu, Xiangyu Wang, Qinrui Zhu, Qiang Tu, and Huanhuan Chen. From query tools to causal architects: Harnessing large language models for advanced causal discovery from data. *arXiv preprint arXiv:2306.16902*, 2023.

Jacob M. Chen, Rohit Bhattacharya, and Katherine A. Keith. Proximal causal inference with text data. *arXiv preprint arXiv:2401.06687*, 2024.

Cholesterol Treatment Trialists' (CTT) Collaboration. Efficacy and safety of more intensive lowering of LDL cholesterol: A meta-analysis of data from 170 000 participants in 26 randomised trials. *The Lancet*, 376(9753):1670–1681, 2010.

Alfred F. Connors, Theodore Speroff, Neal V. Dawson, Charles Thomas, Frank E. Harrell, Douglas Wagner, Norman Desbiens, Lee Goldman, Albert W. Wu, Robert M. Califf, William J. Fulkerson, et al. The effectiveness of right heart catheterization in the initial care of critically ill patients. *JAMA*, 276(11):889–897, 1996.

Yifan Cui, Hongming Pu, Xu Shi, Wang Miao, and Eric J. Tchetgen Tchetgen. Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 119(545):1348–1359, 2024.

Naoki Egami, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. How to make causal inferences using texts. *Science Advances*, 8(42):eabg2652, 2022.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zachary Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, et al. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics (TACL)*, 10:1138–1158, 2022.

Thomas A. Glass, Sander N. Goodman, Miguel A. Hernán, and Jonathan M. Samet. Causal inference in public health. *Annual Review of Public Health*, 34:61–75, 2013.

Keisuke Hirano and Guido W. Imbens. The propensity score with continuous treatments. In Andrew Gelman and Xiao-Li Meng (eds.), *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, pp. 73–84. Wiley, 2004.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*, 2023.

Dean Knox, Christopher Lucas, and Wendy K. Tam Cho. Testing causal theories with learned proxies. *Annual Review of Political Science*, 25:419–441, 2022.

Benjamin Kompa, David R. Bellamy, Thomas Kolokotrones, James M. Robins, and Andrew L. Beam. Deep learning methods for proximal inference via maximum moment restriction. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Manabu Kuroki and Judea Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.

Stephanie Long, Tibor Schuster, and Alexandre Piché. Can large language models build causal graphs? *arXiv preprint arXiv:2303.05279*, 2023.

Afsaneh Mastouri, Yuchen Zhu, Limor Gultchin, Anna Korba, Ricardo Silva, Matt Kusner, Arthur Gretton, and Krikamol Muandet. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7512–7523, 2021.

Ruolin Meng, Ming-Yu Chung, Dhanajit Brahma, Ricardo Henao, and Lawrence Carin. Coupling generative modeling and an autoencoder with the causal bridge. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Wang Miao, Zhi Geng, and Eric J. Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

Wang Miao, Wenjie Hu, Elizabeth L. Ogburn, and Xiao-Hua Zhou. Identifying effects of multiple treatments in the presence of unmeasured confounding. *Journal of the American Statistical Association*, 118(543):1953–1967, 2023.

Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. *Causal Inference in Statistics: A Primer*. Wiley, Chichester, 2016.

Zhengling Qi, Rui Miao, and Xiaoke Zhang. Proximal learning for individualized treatment regimes under unmeasured confounding. *Journal of the American Statistical Association*, 119(546):915–928, 2024.

Peter M. Robinson. Root-N-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

Scandinavian Simvastatin Survival Study Group. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: The Scandinavian Simvastatin Survival Study (4S). *The Lancet*, 344(8934):1383–1389, 1994.

Erik Sverdrup and Yifan Cui. Proximal causal learning of conditional average treatment effects. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 33285–33298, 2023.

Eric J. Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N. Balasubramanian, and Amit Sharma. Causal inference using LLM-guided discovery. *arXiv preprint arXiv:2310.15117*, 2023.

Liyuan Xu, Heishiro Kanagawa, and Arthur Gretton. Deep proxy causal learning and its application to confounded bandit policy evaluation. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Salim Yusuf, Jackie Bosch, Gilles Dagenais, Jun Zhu, Denis Xavier, Lisheng Liu, Prem Pais, Patricio López-Jaramillo, Lawrence A. Leiter, Antonio Dans, et al. Cholesterol lowering in intermediate-risk persons without cardiovascular disease. *New England Journal of Medicine*, 374(21):2021–2031, 2016.

A EXAMPLE RESPONSES

Three of the forty responses on each clinical cohort, excerpted verbatim: the model's statement of the latent confounder, its reasoning on two variables, and the partition it returned. `[...]` marks an omission; the two variables are ones on which the arms differ. On each cohort the first two responses are from newer models and the third from an older one. Each is a single replicate; the partitions of §5.2 and §5.1 are the majority vote over five.

A.1 FRAMINGHAM OFFSPRING

Claude Opus 5 (`claude-opus-5`), replicate 1 of 5

```
# Part 0 -- Identifying the latent confounder U

**U = the clinician's (and the patient's own) global, unrecorded assessment of atherosclerotic
     /cardiometabolic risk and overall frailty at Exam 6** -- the "confounding-by-indication
     gestalt": accumulated vascular burden, subclinical disease, comorbidity load, functional
     decline, and intensity of medical surveillance that a physician integrates when deciding
     to start a statin, but which no single measured variable captures. U raises the
     probability of statin prescription (higher perceived risk → treat) and simultaneously
     raises all-cause mortality (sicker, frailer, more atherosclerotic patients die sooner).
     [...]

# Part 1 -- Per-variable reasoning

[...]

**chol5** -- **Z**. Exam-5 total cholesterol is the canonical trigger of the statin decision (a
      Z→A edge is explicitly permitted), it tracks U (dyslipidemic/atherosclerotic burden),
     and in a cohort of this age total cholesterol has no robust direct effect on all-cause
     mortality once vascular risk U is fixed → A3a holds; it is a lab value, not a determinant
      of BP/BMI, so A3b is plausible.

**sbp6** -- **W**. Systolic BP strongly predicts mortality (a W→Y edge is permitted, dashed)
     but is not itself a statin indication -- it triggers antihypertensives, not lipid-lowering
      therapy -- so its association with statin use runs through the global risk gestalt U (and
      through bpmeds6/X), satisfying A2.

[...]

# Part 2 -- Classification JSON

```json
{
  "Z": ["chol5", "chol51", "chol52", "hdl5", "hdl51", "hdl52", "trigly5", "trigly51", "
       trigly52", "gluc5", "gluc51", "gluc52", "bpmeds6"],
  "W": ["sbp6", "sbp61", "sbp62", "dbp6", "dbp61", "dbp62", "bmi6", "bmi61", "bmi62"],
  "X_obs": ["age6", "age61", "age62", "female", "smoke6", "stk_hx6", "pad_hx6", "mi_hx6", "
       ascvd_hx6", "diab6"],
  "D": []
}
```
```

Grok 4.6 (`grok-4.6`), replicate 1 of 5

```
**Part 0 -- Identify the latent confounder U.**

U is the patient's residual statin-indication severity and cardiovascular (plus general) risk
     burden that remains after the measured covariates are taken as given: subclinical
     atherosclerotic mass, unmeasured lipid fractions (LDL-C, apoB, Lp(a)), inflammatory tone
     (e.g. CRP), family history, frailty / physiologic reserve, physician gestalt, and health-
     care engagement. [...]

**Part 1 -- Per-variable reasoning**

[...]

**chol5.** Role = Z. Total cholesterol at the previous exam is the lab that ATP-II / early-
     statin-era physicians actually treat; the dashed Z → A edge is therefore present and
     permitted. A3a holds: given U (the atherosclerotic / lipid-risk state that cholesterol
     produces), the Exam-5 reading is measurement-plus-transient noise and has no residual
     path to death. A2 fails, so chol5 cannot be W; there is no Y-channel independent of U, so
      it is not X_obs.

**sbp6.** Role = X_obs. SBP is an explicit FRS / ATP-II input (direct path to A) and
     mechanically produces stroke, heart-failure and death via hypertensive target-organ
     damage that is not identical to the atherosclerotic indication U (direct path to Y not
     fully through U). W would violate A2; Z would violate A3a.
```

````
[...]

**Part 2 -- Classification JSON**

```json
{
  "Z": ["chol5", "chol51", "chol52", "hdl5", "hdl51", "hdl52"],
  "W": ["trigly5", "trigly51", "trigly52", "dbp6", "dbp61", "dbp62", "bmi6", "bmi61", "bmi62",
       "gluc5", "gluc51", "gluc52"],
  "X_obs": ["age6", "age61", "age62", "sbp6", "sbp61", "sbp62", "female", "smoke6", "stk_hx6",
       "pad_hx6", "mi_hx6", "ascvd_hx6", "diab6", "bpmeds6"],
  "D": []
}
```
````

GPT-4 Turbo (`gpt-4-turbo`), replicate 1 of 5

````
## Part 0 -- Identification of the Latent Confounder U

In the context of statin use and its impact on all-cause mortality in the Framingham Offspring
     Cohort, the latent confounder U likely represents the underlying cardiovascular risk
    profile of each individual. [...]

## Part 1 -- Per-variable reasoning

[...]

1. **chol5, chol51, chol52** - **W**: Cholesterol levels influence statin prescription and
    cardiovascular outcomes, but mostly reflect cardiovascular risk (U) without directly
    driving treatment decisions independently of this risk. **A2** holds: These variables
    likely predict the outcome of cardiovascular risk, not treatment assigment when
    conditioned on U and X_obs.

5. **sbp6, sbp61, sbp62** - **X_obs**: Systolic blood pressure directly affects both the
    decision to prescribe statins and directly impacts cardiovascular outcomes, acting
    independently of U.

[...]

## Part 2 -- Classification JSON

```json
{
  "Z": ["smoke6"],
  "W": ["chol5", "chol51", "chol52", "hdl5", "hdl51", "hdl52", "trigly5", "trigly51", "
       trigly52", "age6", "age61", "age62", "stk_hx6", "pad_hx6", "mi_hx6", "ascvd_hx6"],
  "X_obs": ["sbp6", "sbp61", "sbp62", "dbp6", "dbp61", "dbp62", "bmi6", "bmi61", "bmi62", "
       gluc5", "gluc51", "gluc52", "female", "diab6", "bpmeds6"],
  "D": []
}
```
````

A.2 SUPPORT

Claude Opus 5 (`claude-opus-5`), replicate 1 of 5

````
## Part 0 -- The latent confounder U

In this cohort U is best understood as the clinician's *gestalt of acute cardiopulmonary/
    haemodynamic instability and short-term prognosis* -- the degree of circulatory shock,
    occult hypoperfusion, and overall "sickness" that the treating physician perceives at the
     bedside (fluid status, adequacy of cardiac output, trajectory over the first hours) but
    that is not captured by any single recorded number. [...]

## Part 1 -- Per-variable reasoning

[...]

**paco21 (arterial PaCO2).** Z. PaCO2 is a ventilation index that directly shapes the acute
    management decision (respiratory failure prompting invasive monitoring and support),
    giving the optional Z→A edge, while its link to death is mediated by the underlying
    respiratory/haemodynamic severity U rather than by a separate causal channel -- A3a is
    satisfied. It is measured on the same gas but is not a near-deterministic function of pH
    given U (pH also depends on the metabolic/bicarbonate component), so A3b is tenable.

**ph1 (arterial pH).** W. Acidaemia is a powerful *prognostic* marker -- it tracks lactate/
    hypoperfusion, i.e. U itself -- and it is legitimate for W to have a direct effect on Y (
    the dashed W→Y edge is permitted). The key requirement is A2: physicians do not order a
````

````
     right heart catheter *because* of the pH number itself once the haemodynamic picture and
     diagnosis are fixed; pH's association with A runs through U (and X), so A2 holds.

[...]

**Part 2 -- Classification JSON**

```json
{
  "Z": ["pafi1", "paco21"],
  "W": ["ph1", "hema1"],
  "X_obs": ["sod1", "pot1", "crea1", "bili1", "alb1", "wblc1"],
  "D": []
}
```
````

Grok 4.6 (`grok-4.6`), replicate 3 of 5

````
**Part 0 -- Latent confounder U**

U is the unmeasured residual of acute hemodynamic and physiologic instability -- the shock
     phenotype, cardiac function, tissue hypoperfusion, and "how sick this patient looks" that
      are not captured by recorded diagnosis, vital signs, comorbidities, or the SUPPORT
     physiology score. [...]

**Part 1 -- Per-variable reasoning**

[...]

**paco21.** Role **Z**. PaCO2 does not independently kill at these values (A3a holds) and is
     obtained on the same ABG that is in front of the physician at the moment of the RHC
     decision, so it tracks the ventilatory face of U. Putting it on the Z side with pH avoids
      an A3b violation (Henderson-Hasselbalch would otherwise make pH a near-deterministic
     function of a W variable). It is not X_obs: there is no PaCO2 ↛ A indication.

**ph1.** Role **Z**. Arterial pH in this IQR is not independently lethal (A3a holds), but
     documented metabolic acidosis is part of the shock work-up that leads to invasive
     monitoring (Z ↛ A dashed edge permitted). It is therefore a treatment proxy for U. A2 fails
     , so it cannot be W; the missing independent Y channel means it is not X_obs.

[...]

**Part 2 -- Classification JSON**

```json
{
  "Z": ["paco21", "ph1"],
  "W": ["sod1", "pot1", "bili1", "alb1", "wblc1", "hema1"],
  "X_obs": ["crea1", "pafi1"],
  "D": []
}
```
````

GPT-4 Turbo (`gpt-4-turbo`), replicate 1 of 5

````
### Part 0 -- Identification of the Latent Confounder U.
In this critical care setting, the latent confounder U likely represents the underlying
     severity and complexity of a patient's medical condition that both influences the
     decision to perform right heart catheterization (RHC) and affects the patient's overall
     survival outcome. [...]

7. **paco21 (PaCO2)** - Role: **X_obs**.
   - Justification: As a measure of respiratory function, directly influences both treatment
        decisions (such as the need for RHC) and survival outcomes. Must be treated as X_obs
        to accurately adjust for its influences on both A and Y.

8. **ph1 (Arterial blood pH)** - Role: **X_obs**.
   - Justification: Blood pH values directly bear on both the necessity of interventions like
        RHC and critical care outcomes like organ health state and thus survival,
        necessitating its role as X_obs.

```json
{
  "Z": ["sod1", "pot1"],
  "W": ["crea1", "bili1", "alb1", "wblc1", "hema1"],
  "X_obs": ["pafi1", "paco21", "ph1"],
  "D": []
}
```
````

B SELF-CONSISTENCY OF THE PARTITIONS

Neither clinical cohort can score a partition for correctness, but both can measure whether an arm reaches the same partition when asked again, which needs no ground truth. Table 5 reports, for each arm and cohort, the voted partition's sizes and its estimate, and two statistics over the arm's five replicates: *unan.* counts the variables on which all five replicates agreed, and *pair.* averages, over the ten pairs of replicates, the fraction of variables given the same label (chance is 0.25).

		Framingham				SUPPORT			
model	identifier	sizes	HR	unan.	pair.	sizes	days	unan.	pair.
Claude Opus 5	`claude-opus-5`	9/12/11	0.76	12	0.88	2/2/6	−1.65	3	0.72
Gemini 3.1 Pro	`gemini-3.1-pro-preview`	9/9/14	0.81	12	0.88	3/7/0	−1.31	4	0.76
GPT-5.6	`gpt-5.6-luna`	3/0/29†	—	9	0.78	2/7/1	−1.46	3	0.63
Grok 4.6	`grok-4.6`	6/9/17	0.78	14	0.93	2/6/2	−1.24	6	0.84
Claude Haiku 4.5	`claude-haiku-4-5`	0/15/17†	—	2	0.61	3/6/1	−1.34	7	0.85
Gemini 3.1 Flash-Lite	`gemini-3.1-flash-lite`	4/12/15	0.89	2	0.46	3/3/3	−1.59	0	0.60
GPT-4 Turbo	`gpt-4-turbo`	9/4/19	0.86	4	0.57	3/5/2	−1.55	0	0.35
Grok 4.20	`grok-4.20-0309-reasoning`	9/9/14	0.81	14	0.93	4/5/1	−1.24	4	0.64

Table 5: Every arm on both clinical cohorts: voted sizes $|Z|/|W|/|X|$, the estimate (hazard ratio on Framingham, days of 30-day survival on SUPPORT; median over 20 runs), and self-consistency over five replicates. Framingham sizes count its 32 covariates and its self-consistency its 16 clinical variables; SUPPORT has 10 variables. † the vote violates completeness and carries no estimate. Gemini 3.1 Flash-Lite discards one covariate on Framingham and one variable on SUPPORT.

Three observations follow. First, every one of the 80 replicates is a parseable, assumption-grounded partition: a weaker arm does not return malformed output, it returns legal partitions that it then fails to reproduce, which is the argument for drawing replicates and voting. Second, more capable models are more reliable. The four newer models average 0.86 pairwise agreement on Framingham against 0.64 for the four older ones, and 0.74 against 0.61 on SUPPORT; within vendor, the newer model is the more self-consistent in six of the eight vendor–cohort pairs. On Framingham 19 of the 20 newer-model replicates return non-empty Z and W, against 16 of 20 for the older models, and all 40 do on SUPPORT. Third, models agree with themselves more than with each other (pairwise agreement across arms is 0.65 on Framingham and 0.57 on SUPPORT), so we report each arm's own vote and not a cross-model consensus.

Model identifiers and dates The *identifier* column of Table 5 gives the identifier each request was sent with, through the vendor's API with no sampling parameter set. The Framingham requests to the Gemini, OpenAI and xAI models were made on 22 August 2026 and to the Anthropic models on 4 September 2026; all SUPPORT requests were made on 4 September 2026. Vendors deprecate models, so these responses are the reproducible object, and all of them are released with the partitions and prompts. The waybill models, Claude Opus 4.8, GPT-4.1 and Gemini 2.5 Flash, were likewise called through their vendors' APIs; Opus 4.8 and Opus 5 are different models, and no statement in this paper joins a waybill result to a clinical result through a model's name.

C FURTHER ANALYSIS OF EACH EXPERIMENT

C.1 WAYBILL

The three partitions The 181 covariates describe the order (distance, package, destination attributes), the calendar and weather, the pickup point's and the zone's recent history (order and waybill volumes, price levels, acceptance and completion rates, rider supply and speed), and the platform's pricing parameters; at the data owner's request they are described by what they measure and not named. Seven of them are components of the offered price and are excluded from every model. Each partition was obtained as in §4, five responses per model voted per covariate. Table 6 sets the three voted partitions side by side. The three models agree on the observed confounders and on the outcome proxies: 45 covariates are in X for all three, the seven outcome proxies of Gemini 2.5 Flash are all among the 37 of GPT-4.1, and those 37 are all among the 66 of Opus 4.8, so the outcome-proxy sets are nested (historical acceptance, completion and delivery-performance

rates, which each model reads as consequences of the latent supply–demand state). They disagree on the treatment proxies, where the three sets are disjoint. Opus 4.8 uses the pickup point's and zone's historical demand volumes and price levels (27 covariates), which GPT-4.1 conditions on as X; GPT-4.1 uses rider supply, speed and distance statistics (50), of which Opus 4.8 treats 29 as outcome proxies and 21 as confounders; and Gemini 2.5 Flash uses the platform's own pricing parameters (5), which Opus 4.8 discards as treatment components and GPT-4.1 conditions on. Pairwise, the three partitions give the same role to 43–51% of the covariates. The disagreement therefore sits on the side of the partition that decides how much of the confounding is corrected, and §5.4 finds the three arms differ exactly there: in the magnitude of the correction, not in the shape of the dose–response.

	$\lvert Z\rvert/\lvert W\rvert/\lvert X\rvert/\lvert D\rvert$	GPT-4.1	Gemini 2.5 Flash
Opus 4.8	27/66/66/22	88 (51%)	75 (43%)
GPT-4.1	50/37/81/13		81 (47%)
Gemini 2.5 Flash	5/7/160/9		
same role in all three		52: 45 in X, 7 in W, none in Z or D	

Table 6: The three voted waybill partitions of the 181 covariates, and their overlap. Diagonal: the role counts of each partition. Off-diagonal: how many covariates the pair assigns the same role, and the share of the 174 non-treatment covariates on which the pair agrees.

Seeds and settings Every proximal arm is run at ten seeds, which set the bridge's initialisation, mini-batch order and Monte Carlo draws of W; the first-stage sampler is trained once per arm and shared by its runs. Stage 1 is a five-component mixture-density network for $p(W \mid T, Z, X)$ with a 256–512–256 ReLU body, trained with Adam (learning rate 10^{-4}, batch size 8192, 80 epochs, gradient clipping at 1.0, learning rate halved after five epochs without improvement, best validation state restored, validation on the last 20% of the day shards). Stage 2 is a bridge with hidden widths 256, 256, 128 and a sigmoid output, trained with AdamW (learning rate 2×10^{-4}, weight decay 10^{-5}, batch size 8192, 80 epochs, the same schedule), $J = 10$ draws of W per sample per step, checkpoints selected on the global validation loss under both objectives. All hyperparameters are shared across arms and were not tuned on the experiment.

Magnitude Write $s = \langle \hat{\tau}, \tau \rangle / \langle \tau, \tau \rangle$ for the slope of the estimated on the trial effects over the dose grid; $s = 1$ is a correctly scaled curve. The three baselines give 0.80, 0.79 and 0.68 (Table 7): they understate the effect of price on acceptance by a fifth to a third. That is the direction the confounder predicts, since the platform raises prices where riders are scarce and acceptance is low, and adjusting for 174 measured covariates does not remove it. The proximal arms act on exactly that attenuation. At the median run all three are better calibrated than every baseline, $|s - 1|$ being 0.05, 0.12 and 0.18 against 0.20 for the best baseline, and the Opus partition is closer to 1 than every baseline in all ten runs. The partitions differ in interpretable ways: the Gemini partition, with five treatment and seven outcome proxies, makes almost no correction, the GPT-4.1 partition over-corrects, and the Opus partition gives 0.95. Within city tiers, the Opus partition's slope is closer to 1 than every baseline in four of the five tiers, GPT-4.1 in two and Gemini in one.

Method	slope s	wins
IPW/GPS	0.68	—
S-Learner (NN)	0.80	—
DML (partialling-out)	0.79	—
PCI, GPT-4.1 partition	1.12 [0.47, 1.42]	5/10
PCI, Gemini partition	0.82 [0.14, 0.86]	6/10
PCI, Opus partition	0.95 [0.84, 1.01]	10/10

Table 7: Waybill, further columns for Table 4, over the same ten seeds: median [range]. *wins* counts the runs in which the arm's ATE error is below the best baseline's 21.1%.

The trial's noise floor On each side of $\delta = 0$ a straight line is fitted to the trial effects; the pooled residual variance estimates the sampling variance σ^2 of one trial effect, and an estimator returning

the fitted lines exactly would score $18\,\sigma\sqrt{2/\pi}\big/\sum_\delta |\tau(\delta)|$ under the relative error, about 12%. A quadratic on each side changes this by 0.3 points. The same calculation within a city tier gives floors of 21% to 51%, 31% on average.

The stratum-balanced objective The loss (4) decomposes exactly by stratum, $\mathcal{L}^{\mathrm{OB}} = \sum_c (n_c/n)\,\mathcal{L}_c$, with $\mathcal{L}_c$ the mean squared Fredholm residual within stratum c, so a large stratum can buy a slight improvement with a large residual in a small one. We give every stratum the same weight instead,

$$\mathcal{L}^{\mathrm{OB}}_{\mathrm{strata}} = \sum_{c\in\mathcal{C}} \lambda_c\,\mathcal{L}_c, \qquad \lambda_c = \frac{1}{|\mathcal{C}|}, \tag{7}$$

which still pools all n samples through the shared θ; $\lambda_c = n_c/n$ recovers (4).

Stratum-balanced training under three stratifications Table 8 repeats the comparison of §5.4 under three stratifications taken from the variables of X. The global arm is a single stratum-agnostic model evaluated under each, so its ATE error is the same in the three rows. Balanced training lowers the median error in every cell, and the interquartile range of the ATE error, 16.0 points wide under the global objective, is at most 1.7 points wide under the balanced one.

stratification		global	balanced
city tier (5)	ATE	16.0 [13.8, 29.8]	**12.0 [11.8, 12.9]**
	avg-CATE	39.1 [37.8, 43.9]	**36.8 [36.4, 37.2]**
destination access mode (6)	ATE	16.0 [13.8, 29.8]	**12.2 [12.0, 13.7]**
	avg-CATE	40.2 [39.3, 47.1]	**38.2 [37.9, 38.7]**
destination floor access (5)	ATE	16.0 [13.8, 29.8]	**12.3 [12.2, 12.4]**
	avg-CATE	41.2 [40.2, 49.5]	**39.2 [38.0, 39.3]**

Table 8: Stratum-balanced (7) against global (4) training of the Opus bridge: relative error in %, median [interquartile range] over five runs; the number of strata is in parentheses.

C.2 FRAMINGHAM

The voted partitions Table 9 gives every arm's vote on each of the sixteen clinical variables. Each continuous variable's raw column and its two spline bases were voted identically by every arm, so the 32 covariates appear as 16 rows with nothing omitted.

variable	Opus 5	Gem. 3.1 Pro	GPT-5.6	Grok 4.6	Haiku 4.5	Gem. Flash-Lite	GPT-4 T.	Grok 4.20	cross	Meng et al. (2025)
total cholesterol	Z	Z	Z	Z	W	W	Z	Z	Z	Z
HDL cholesterol	Z	Z	X	Z	W	W	Z	Z	Z	Z/W
triglycerides	Z	Z	X	W	W	W	Z	Z	Z	Z/W
age	X	X	X	X	X	X	X	X	X	W
systolic blood pressure	W	X	X	X	X	X	X	X	X	Z/W
diastolic blood pressure	W	W	X	X	X	X	X	W	X	Z
body mass index	W	W	X	W	W	X	X	W	W	Z
fasting glucose	W	W	X	W	W	W	X	W	W	W
sex	X	X	X	X	X	X	X	X	X	W
current smoker	X	X	X	X	X	X	X	X	X	W
stroke history	X	X	X	X	X	Z	W	X	X	W
peripheral artery disease	X	X	X	X	X	Z	W	X	X	W
myocardial infarction	X	X	X	X	X	Z	W	X	X	Z
ASCVD history	X	X	X	X	X	Z	W	X	X	W
diabetes	X	X	X	X	X	X	X	X	X	W
antihypertensive medication	X	X	X	X	X	D	X	X	X	Z

Table 9: Framingham: voted role of each clinical variable per model, the majority over the eight models (*cross*), and the split of Meng et al. (2025), which assigns every column to Z or W and none to X; where that split puts a variable's spline bases on both sides it reads Z/W. Z treatment proxy, W outcome proxy, X observed confounder, D discarded. Models are ordered as in Table 5.

How often would chance agree this closely? The five LLM partitions span a factor of 1.17 in the hazard ratio. Of all $142{,}506$ sets of five among the thirty uniform random partitions, 6.9% agree with one another that closely, a typical set spanning a factor of 1.36; against the size-matched control the

figure is 3.5%. The models agree because they share a reading of the problem: all eight arms place age, sex, smoking and diabetes in X, which a uniform draw does for a given variable one time in three, and they disagree mainly over which lipid and blood-pressure measurements are proxies.

Why random partitions are not far behind on this cohort The sixteen measured clinical variables capture much of the confounding by indication: the overlap-weighted Cox fit of the source analysis, which adjusts for them and uses no proxies, already reaches 0.90. Any partition that keeps the main confounders in X therefore lands in a similar place, whichever variables it names as proxies, and a random partition does so two times in three. What separates partitions is whether the confounders stay in X: that is what the uniform draws that leave the range have lost, and what the perturbation control below removes deliberately.

Which decision matters Keeping each LLM partition's X and reassigning its proxies to Z and W at random (thirty draws, twenty runs each) gives hazard ratios from 0.75 to 0.94, a factor of 1.26 against 2.4 for fully random partitions, with 27 of 30 in the trial range. Exchanging Z and W outright moves the estimate by factors of 0.95 to 1.07, within run-to-run noise. On this cohort what the models contribute is chiefly the choice of X, the decision an association-based rule cannot make.

Perturbing the LLM's choice of confounders The complement of the control above corrupts X instead of the proxies. For each of the five distinct admissible partitions, $k \in \{1, 2, 4\}$ of the clinical variables the arm placed in X are moved out, either to a proxy role (Z or W, drawn per variable) or to D, so that they leave the model altogether; four draws per arm and k, twenty runs each, the identical pipeline throughout, and each perturbed partition is compared with its own arm's estimate in Table 2. Moving confounders into a proxy role changes little: the estimate moves by a median factor of 1.02–1.05 and 15 to 18 of the 20 perturbed partitions stay in the trial range, consistent with the X-fixed control above, since a variable moved to Z or W still enters the model. This is also what separates the uniform random partitions that leave the range from these perturbations: a uniform draw keeps a given variable in X one time in three, so of the eight variables every LLM placed in X a random partition keeps 3.0 on average, 2.0 among the ten that leave the range against 3.5 among the twenty that stay, where even the $k = 4$ perturbation removes at most four of an arm's confounders. Dropping confounders degrades the estimate with k: the share of perturbed partitions in the range falls from 16 of 20 at $k = 1$ to 9 of 20 at $k = 4$, the median hazard ratio drifts from the arms' 0.76–0.89 to 0.94, the direction of the confounded fit, and the worst partitions reach 1.20. The LLM's choice of which variables to keep as confounders is therefore load-bearing, and the error it would be costliest to make, discarding a confounder, is the one that shows.

perturbation	k	HR	in range	factor	runs in range
to proxy	1	0.84 [0.76, 0.95]	18 of 20	1.05	69%
	2	0.82 [0.73, 1.06]	18 of 20	1.02	62%
	4	0.82 [0.67, 0.99]	15 of 20	1.05	61%
to D	1	0.83 [0.74, 1.06]	16 of 20	1.04	68%
	2	0.83 [0.75, 1.20]	14 of 20	1.04	58%
	4	0.94 [0.77, 1.20]	9 of 20	1.11	43%

Table 10: Perturbing the LLM partitions on Framingham: k of the variables an arm placed in X are moved to Z or W (*to proxy*) or dropped from the model (*to D*), four draws for each of the five admissible arms, twenty runs each. Columns: median and [min, max] over the 20 perturbed partitions' medians; how many of the 20 lie in the trial range $[0.70, 0.90]$; the median factor by which a perturbed partition's estimate differs from its own arm's in Table 2; and the share of all 400 runs in the range, against 68% for the unperturbed arms.

C.3 SUPPORT

The voted allocations Table 11 gives every arm's vote on each of the ten blood-panel variables, beside the allocation of Tchetgen Tchetgen et al. (2020), which the models were never shown.

		Opus 5	Gem. 3.1 Pro	GPT-5.6	Grok 4.6	Haiku 4.5	Gem. Flash-Lite	GPT-4 T.	Grok 4.20	cross	Tchetgen Tchetgen et al. (2020)
sod1	sodium	X	W	W	W	W	X	Z	W	W	X
pot1	potassium	X	W	W	W	W	X	Z	W	W	X
crea1	creatinine	X	W	W	X	X	W	W	X	X	X
bili1	bilirubin	X	W	W	W	W	W	W	W	W	X
alb1	albumin	X	W	W	W	W	W	W	W	W	X
pafi1	PaO_2/FiO_2	Z	Z	Z	X	Z	Z	Z	Z	Z	Z
paco21	$PaCO_2$	Z	Z	X	Z	Z	Z	X	Z	Z	Z
ph1	pH	W	Z	Z	Z	Z	Z	X	Z	Z	W
wblc1	white cell count	X	W	W	W	W	D	W	Z	W	X
hema1	haematocrit	W	W	W	W	W	X	W	W	W	W

Table 11: SUPPORT: voted role of each of the ten blood-panel variables per model, the majority over the eight models (*cross*), and the allocation of Tchetgen Tchetgen et al. (2020). Claude Opus 5 matches the published column on every row; six models place pH in Z.

Direction of the correction Least squares on every non-discarded covariate gives -1.43 days under every allocation, since an allocation never changes what is adjusted for. Against that baseline five allocations move the estimate away from zero and four toward it, and their 0.41-day spread is of the order of the source analysis's 0.55-day gap between proximal and ignorability estimates: the partition decides the size of the unmeasured-confounding correction.

How often would chance agree this closely? A typical set of eight among the thirty uniform random allocations spans 0.90 days, and of all $5{,}852{,}925$ such sets 1.8% agree as closely as the eight LLM arms do. That the random medians are not far from the LLM medians is expected here: the models allocate only the ten blood-panel variables, while the other 43 covariates are conditioned on as X under every allocation, so a random allocation changes a small part of what the estimator sees, and the difference it can make appears in the dispersion and the tails rather than in the centre.

The side a proxy sits on Exchanging Z and W, with X untouched, moves six of the eight distinct allocations by 0.3 to 1.7 days, each by more than 2.8 standard errors, and seven of the eight shifts exceed the LLM arms' entire spread. Here, unlike on Framingham, the side a proxy sits on matters.

De-identification With variable names replaced by shuffled neutral codes and the study's name and counts removed, all five Claude Opus 5 replicates return the published allocation; with the treatment also paraphrased and all statistics dropped, three of five do and the vote is unchanged.