CAUSALLY FAIR GENERATION WITH LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly used to generate, complete, and transform information in settings where their outputs can shape consequential decisions, raising concerns about their impact on demographic disparities. In this context, causal inference provides a principled basis for assessing fairness, because it attributes observed disparities to the mechanisms that generated them, which a purely statistical approach cannot do even with infinite data. In LLM generation, a query may request several causally related variables, each of which is both an outcome of interest and a possible cause of other outputs, and the information supplied in the prompt need not follow a topological or a temporal order. This calls for methods that can analyze and selectively remove disparities from such a flexible generation process. In this paper we introduce Causally Fair Generation with LLMs (CFG, for short). CFG extracts relevant concepts, grounds generation in a reference population and causal diagram, and removes user-selected causal effects. CFG also allows pathways deemed justifiable for the task's utility to be retained, which is known in legal literature as business necessity. Further, we provide formal guarantees for our method when eliminating all discriminatory causal effects in the adapted population model, under appropriate causal assumptions. We evaluate CFG with four LLMs in three real-world settings based on population data and on a synthetic dataset with a known causal ground truth.

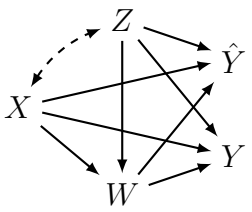
1 INTRODUCTION

Large language models (LLMs) have become general-purpose tools for producing, retrieving, and transforming information (Brown et al., 2020). Their ability to turn partial instructions into fluent text has led to their use in education, health care, hiring, and other settings where generated content can inform consequential decisions (Kasneci et al., 2023; Singhal et al., 2023; An et al., 2024). At scale, an LLM need not make a final decision itself to have an effect: its summaries, recommendations, synthetic profiles, and inferred attributes can shape the information presented to a human or a downstream system. These uses make the behavior of LLMs across demographic groups an important object of study. Since these models learn from data shaped by social and historical processes, their outputs may reproduce existing disparities or introduce new ones.

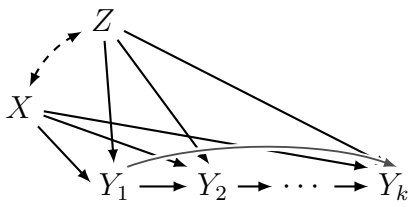
Most evaluations of social bias begin with statistical comparisons. They test, for example, whether prompts referring to different demographic groups produce different occupations, sentiments, or decision rates (Gallegos et al., 2024). Such comparisons are a useful starting point for detecting disparities, but they do not explain how those disparities arose. A difference may be introduced when the model uses a protected attribute directly, inherited from another generated attribute, or induced by background characteristics associated with the protected attribute. The same observed disparity can therefore reflect different underlying causal mechanisms.

The growing literature on causal fairness (Kusner et al., 2017; Kilbertus et al., 2017; Chiappa, 2019; Plecko & Bareinboim, 2022) offers two important benefits in this context. First, it yields human-understandable and interpretable notions of fairness, tied to the causal mechanisms that transmit the disparity between groups. Second, the causal approach offers a language aligned with legal notions of discrimination, such as the disparate impact doctrine. This includes the notion of business necessity, which specifies which covariates a decision-maker may justifiably use, even if their use results in a disparity between groups. In other words, the causal approach allows granular removal

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.



(a) Standard Fairness Model (SFM) (Plecko & Bareinboim, 2022).



(b) Standard Fairness Model for LLMs.

Figure 1: Causal diagrams for the standard fairness model and its LLM adaptation. Panel (b) orders outputs topologically from left to right. Its template includes $Y_i \to Y_j$ for every $i < j$; only selected forward edges are drawn for readability. The general formulation permits any subset of these edges.

or retention of causal effects in line with user requirements, something purely associational methods cannot achieve.

Within causal fairness, it is helpful to distinguish three fairness tasks: (1) bias detection, which assesses discrimination in existing outcomes or decisions; (2) fair prediction, which constructs predictors satisfying desired fairness constraints; and (3) fair decision-making, which designs actions to improve real-world outcomes under fairness constraints. In this work, we contribute to a new fairness task, namely that of *fair generation*: constructing outputs that remove selected sources of disparity while retaining pathways deemed permissible. Fair generation is related to fair prediction, but is markedly more complex. To understand this difference, we compare our setting with the classical machine learning setting, shown in Fig. 1(a), where X denotes the protected attribute, $\mathbf{Z}$ background covariates, $\mathbf{W}$ mediators, and Y the outcome. In classical ML, the task of fair prediction aims to learn a single mechanism for producing $\widehat{Y}$ from these covariates, with $\widehat{Y}$ mimicking Y in an appropriate way. In contrast, instead of a single prediction, an LLM can generate multiple causally related attributes from a partially specified profile, implicitly supplying a separate generation mechanism for each desired covariate. This is illustrated in the following example:

Example 1 (Applicant Profiles)**.** *For safety and privacy reasons, testing of public-sector data-processing systems is increasingly based on synthetic data (Office for National Statistics, 2023). Consider an analyst creating synthetic applicant stories to test an employment-support service, using the American Community Survey (ACS) as the population data source (U.S. Census Bureau, 2023). In each prompt, the analyst supplies sex (X) and age ($\mathbf{Z}$) taken from real data, and requests the model to complete the information on education (Y_1), employment status (Y_2), and salary (Y_3):*

> Write a story about a 40-year-old person recorded as female, including their education, employment status, and annual salary.

To avoid encoding stereotypical patterns in the resulting outputs, the analyst also wants to remove direct effects of sex on employment and salary, $X \to Y_2$ and $X \to Y_3$, while permitting education-mediated pathways for this scenario, such as $X \to Y_1 \to Y_2$ and $X \to Y_1 \to Y_3$. Importantly, education and employment are both generation targets and mediators of the effect of sex on salary, so fair generation must account for the causal relationships among these variables. □

In short, the distribution resulting from the above example can be represented as

$$P_{\text{rw}}(x, \mathbf{z}) P_{\text{LLM}}(y_1, y_2, y_3 \mid x, \mathbf{z}), \tag{1}$$

indicating that $x, \mathbf{z}$ are drawn from real data, while y_1, y_2, y_3 are drawn from the LLM's implied distribution. Naturally, the LLM's generative distribution may differ from the real-world one, $P_{\text{LLM}}(y_1, y_2, y_3 \mid x, \mathbf{z}) \neq P_{\text{rw}}(y_1, y_2, y_3 \mid x, \mathbf{z})$. Moreover, even if the LLM did generate data according to the real-world distribution, this would simply reproduce existing disparities – not satisfying the analyst's requirement for removing the direct effects of sex on employment and salary. On the other hand, methods for statistical output filtering (Gallegos et al., 2024) equalize group distributions and erase the education-mediated differences the analyst chose to retain. Fair generation, therefore, requires careful modification of the selected mechanisms to ensure generation satisfies the target constraints.

More generally, we arrange the generated attributes $Y_1, \ldots, Y_k$ in a topological order, with each attribute potentially affecting any later one (Fig. 1(b)). Interestingly, a prompt may also supply later attributes and request earlier ones, so generation need not follow a topological or a temporal order (e.g., the analyst may request generation of the sex variable based on age and salary). Fairness constraints must account for these dependencies across the complete generation. In this context, our contributions are:

(i) We formulate fair generation under the standard fairness model for topological and non-topological queries (Sec. 2.1). We characterize business-necessity removal choices and provide formal guarantees for generation under full effect removal (Thm. 1).

(ii) We introduce CFG, an interactive method for causally fair generation with LLMs (Sec. 2.2), which extracts concepts from a query, grounds them in population data and a causal diagram, removes selected effects while allowing business-necessity exceptions, and guides the LLM to express the resulting attributes in text.

(iii) We evaluate bias removal with CFG using four LLMs in three real-world settings. We measure reductions in selected effects at intermediate and final attributes, test non-topological queries, and examine the retention of permitted pathways (Sec. 3.2).

Related Work. Bias benchmarks measure stereotypes and representational harms (Nangia et al., 2020; Nadeem et al., 2021; Parrish et al., 2022; Smith et al., 2022); broader audits assess trustworthiness, implicit bias, and discrimination (Wang et al., 2024; Tamkin et al., 2023; Bai et al., 2025); see Gallegos et al. (2024) for a survey. These works measure disparities statistically, whereas CFG focuses on the selective removal of undesirable causal effects from generated outputs. The second line of related work elicits causal knowledge from LLMs or evaluates their causal reasoning (Kıcıman et al., 2024; Zečević et al., 2023; Jin et al., 2024a;b; Joshi et al., 2024), or uses them to propose causal graphs (Abdulaal et al., 2024; Liu et al., 2024); CFG uses an LLM to elicit a standard fairness model over the query's concepts, but deploys the diagram for the downstream purpose of bias removal. Next, causal approaches to LLM bias use mediation, invariant learning, controlled generation, and prompting (Vig et al., 2020; Zhou et al., 2023; Madhavan et al., 2023; Li et al., 2025), but a causal model is embedded in the LLM's internal computation, rather than imposed over population variables as in CFG. Finally, causal fairness methods address *fair prediction*, constraining a single predicted outcome in the classical ML setting (Kusner et al., 2017; Kilbertus et al., 2017; Nabi & Shpitser, 2018; Chiappa, 2019; Plecko & Bareinboim, 2022; Plecko & Meinshausen, 2020), including for images (Joo & Kärkkäinen, 2020; Jung et al., 2024), facial expressions (Cheong et al., 2022), and text (Garg et al., 2019). Plecko (2026) addresses *bias detection* in generative AI, focusing on just quantifying causal effects in generated outputs. CFG, on the other hand, is the first method for *fair generation*: constructing multiple causally related attributes under user-selected fairness constraints, including for non-topological queries, while retaining pathways deemed permissible.

1.1 PRELIMINARIES

We use structural causal models (SCMs) as our semantic formalism (Pearl, 2009). An SCM is a tuple $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{U}) \rangle$, where $\mathbf{V}$ and $\mathbf{U}$ are the endogenous (observed) and exogenous (latent) variables, respectively, and $P(\mathbf{U})$ is a strictly positive distribution over $\mathbf{U}$. The set $\mathcal{F}$ contains one structural assignment $V_i \leftarrow f_i(\text{pa}(V_i), \mathbf{U}_{V_i})$ for each $V_i \in \mathbf{V}$, where $\text{pa}(V_i) \subseteq \mathbf{V}$ and $\mathbf{U}_{V_i} \subseteq \mathbf{U}$. The corresponding causal diagram $\mathcal{G}$ contains a directed edge $V_i \to V_j$ when V_i is an argument of f_j, and a dashed bidirected edge $V_i \dashleftarrow\!\!\dashrightarrow V_j$ when their exogenous causes are dependent. A realization $\mathbf{U} = \mathbf{u}$ is called a *unit*. The potential outcome $Y_x(\mathbf{u})$ is the value of Y for unit $\mathbf{u}$ in the submodel $\mathcal{M}_x$, obtained by replacing the structural assignment for X with $X \leftarrow x$.

We adopt the standard fairness model in Fig. 1(a). Here, $X \in \{x_0, x_1\}$ is a protected attribute, $\mathbf{Z}$ contains pre-exposure covariates that may induce spurious association between X and Y, $\mathbf{W}$ contains mediators, and Y is the outcome. We use lowercase events as shorthand; for example, $P(y_{x,\mathbf{w}} \mid x') := P(Y_{x,\mathbf{w}} = y \mid X = x')$. The nested potential outcome $Y_{x,\mathbf{W}_{x'}}$ sets $X = x$ while assigning $\mathbf{W}$ the value it would attain under $X = x'$. Following the causal-fairness literature (Zhang & Bareinboim, 2018; Plecko & Bareinboim, 2022), the x-specific direct, indirect, and spurious effects are x-$\text{DE}_{x_0,x_1}(y \mid x) := \mathbb{E}(Y_{x_1,\mathbf{W}_{x_0}} \mid x) - \mathbb{E}(Y_{x_0} \mid x)$, x-$\text{IE}_{x_1,x_0}(y \mid x) := \mathbb{E}(Y_{x_1,\mathbf{W}_{x_0}} \mid x) - \mathbb{E}(Y_{x_1} \mid x)$, and x-$\text{SE}_{x_1,x_0}(y) := \mathbb{E}(Y_{x_1} \mid x_0) - \mathbb{E}(Y_{x_1} \mid x_1)$. The direct effect changes

X from x_0 to x_1 along the direct path while holding the mediators at their values under x_0. The indirect effect measures the reverse change from x_1 to x_0 along paths through $\mathbf{W}$, with X fixed at x_1 along the direct path. The spurious effect compares the two observed groups after fixing $X = x_1$ interventionally. Based on these notions, the total variation of outcome Y, defined as $\mathrm{TV}_{x_0,x_1}(y) := \mathbb{E}(Y \mid x_1) - \mathbb{E}(Y \mid x_0)$, can be decomposed as (Zhang & Bareinboim, 2018):

$$\mathrm{TV}_{x_0,x_1}(y) = x\text{-DE}_{x_0,x_1}(y \mid x_0) - x\text{-IE}_{x_1,x_0}(y \mid x_0) - x\text{-SE}_{x_1,x_0}(y). \tag{2}$$

In words, total variation decomposes into direct, indirect, and spurious components.

2 CFG: CAUSALLY FAIR GENERATION WITH LLMS

We assume the data-generating process is compatible with the SFM in Fig. 1(a), and we additionally assume a Markovian structure within the set of mediators $\mathbf{W}$, yielding the graph in Fig. 1(b).

2.1 TOPOLOGICAL & NON-TOPOLOGICAL QUERIES

Example 1 (Applicant Profiles – continued)**.** *An analyst supplies sex (X) and age ($\mathbf{Z}$) variables, and uses an LLM to generate applicant stories, while removing direct effects of sex on employment (Y_2) and salary (Y_3) and permitting pathways mediated by education (Y_1). She samples from*

$$P_{\mathrm{rw}}(x, \mathbf{z}) P_{\mathrm{LLM}}(y_1, y_2, y_3 \mid x, \mathbf{z}), \tag{3}$$

with a specific goal for the P_{LLM} to satisfy a causal fairness constraint described above. □

The query in the above example is *topological*, since its supplied variables $X, \mathbf{Z}$ causally precede the requested variables Y_1, Y_2, Y_3. In such a setting, each requested attribute is both an outcome and a potential mediator for later outputs. Since part of the data comes from the user (real world), and part of the data from the LLM, we need to represent real-world and generative mechanisms within a common SCM. Specifically, we introduce one selector indicator for each structural mechanism, $\mathbf{S} = (S_0, \ldots, S_k)$, where $S_j \in \{\mathrm{rw}, \mathrm{gm}\}$. The selector S_0 determines how the root variables $(X, \mathbf{Z})$ are generated, while S_i, $i \geq 1$, determines the mechanism generating Y_i. Each selector chooses both the corresponding structural function and its exogenous distribution. Thus, for a fixed selection $\mathbf{s} = (s_0, \ldots, s_k)$, the induced SCM $\mathcal{M}_{\mathbf{s}}$ is

$$(X, \mathbf{Z}) \leftarrow f_0^{s_0}(U_0), \qquad U_0 \sim P_0^{s_0}, \tag{4}$$

$$Y_i \leftarrow f_i^{s_i}(\mathrm{pa}_G(Y_i), U_i), \qquad U_i \sim P_i^{s_i}, \qquad i = 1, \ldots, k, \tag{5}$$

where $U_0, \ldots, U_k$ are mutually independent. Hence, different choices of $\mathbf{s}$ correspond to replacing individual real-world mechanisms by their generative counterparts while leaving the remaining mechanisms unchanged. For the applicant-profile example, $X, \mathbf{Z}$ come from the real world ($s_0 = \mathrm{rw}$), while (Y_1, Y_2, Y_3) are generated from the model, so that $s_i = \mathrm{gm}$ for $i = 1, 2, 3$. More generally, other generative combinations are possible by varying $\mathbf{s}$. In particular, the selection $\mathbf{s}^{\mathrm{rw}} = (\mathrm{rw}, \ldots, \mathrm{rw})$ recovers the real-world population SCM and induces the population distribution

$$P_{\mathrm{rw}}(x, \mathbf{z}, \mathbf{y}) = P_{\mathcal{M}_{\mathbf{s}^{\mathrm{rw}}}}(x, \mathbf{z}, \mathbf{y}),$$

which simply corresponds to the ground-truth data in ACS. For a mixed selection $\mathbf{s}$, and a fixed LLM, prompting, and decoding parameters, we assume that the distribution of extracted generated attributes is induced by the corresponding selected SCM,

$$P_{\mathrm{LLM}}(\mathbf{y} \mid x, \mathbf{z}) = P_{\mathcal{M}_{\mathbf{s}}}(\mathbf{y} \mid x, \mathbf{z}).$$

With this in mind, we can introduce the concept of *population grounding* – where the generative distribution of the model, e.g., $P_{\mathrm{LLM}}(\mathbf{y} \mid x, \mathbf{z})$, is steered towards the real world $P_{\mathrm{rw}}(\mathbf{y} \mid x, \mathbf{z})$. Grounding alone, however, reproduces the population's disparities, and thus we may be interested in first removing undesirable causal effects, as described in the sequel.

Outcome Effects & Business Necessity. For each generated target Y_i, we set $\mathbf{W}^{(i)} = \mathbf{Y}_{<i}$ and adopt the DE, IE, and SE definitions from the preliminaries. Evaluating these quantities in $\mathcal{M}_{\mathbf{s}}$ gives

$$\mathrm{TV}_{x_0,x_1}(y_i) = x\text{-DE}_{x_0,x_1}(y_i \mid x_0) - x\text{-IE}_{x_1,x_0}(y_i \mid x_0) - x\text{-SE}_{x_1,x_0}(y_i). \tag{6}$$

The direct effect bypasses earlier generated variables, the indirect effect passes through them, and the spurious effect captures association induced by $\mathbf{Z}$. In our example, education (Y_1) has no earlier mediator and hence no indirect effect; education can mediate effects on employment (Y_2), while both education and employment can mediate effects on salary (Y_3). The spurious component reflects the association between sex and age that subsequently influences Y. For each target Y_i, the analyst specifies whether the direct, indirect, and spurious components should be removed.

Fully removing an indirect effect, however, may be too restrictive when some mediated pathways are substantively justified for the task. Such exceptions may be captured through the notion of *business necessity*, a common consideration in the legal literature (Bernhardt, 1971). Hence, for each target Y_i, the analyst may specify a BN set $\mathbf{B}_i$ of generated variables on directed paths from X to Y_i, possibly including the target itself. A directed path is permitted whenever it contains at least one variable in $\mathbf{B}_i$, and is therefore exempt from the corresponding DE or IE removal request. In our example, choosing $\mathbf{B}_3 = \{Y_1\} = \{\text{education}\}$ for salary permits both $X \to Y_1 \to Y_3$ and $X \to Y_1 \to Y_2 \to Y_3$, while not exempting $X \to Y_3$ or $X \to Y_2 \to Y_3$, which bypass education. Likewise, $\mathbf{B}_2 = \{Y_1\}$ permits the education-mediated pathway to employment $X \to Y_1 \to Y_2$. BN exceptions concern directed pathways; spurious component removal is specified separately.

The choice of $\mathbf{B}_i$ also raises an identification question. A sufficient condition for identifying (and thus correctly removing) the permitted path effects is that $\mathbf{B}_i$ is *backward closed*:

$$\mathrm{pa}_G(V) \cap \mathrm{De}_G(X) \subseteq \mathbf{B}_i \qquad \text{for every } V \in \mathbf{B}_i,$$

where $\mathrm{pa}_G(V)$ denotes the parents of V and $\mathrm{De}_G(X)$ the strict descendants of X. Thus, whenever a selected BN variable has a parent that is itself affected by X, that parent must also be included. The education-only set used above satisfies this condition because education has no generated parents. When backward closure does not hold, the relevant path effects may still be identified under stronger assumptions. App. A.2 offers a broader discussion on business necessity.

We collect the analyst's target-specific DE, IE, and SE removal choices together with the BN exceptions into a policy ρ. The full notation allows these choices to vary across targets, but the interface can be simpler in practice; for example, the analyst may specify a single BN set to be applied wherever relevant, or use the same removal rule across multiple generated attributes.

Generation for topological queries. The policy ρ defines the *fair world*: starting from the real-world population distribution, we modify the selected causal pathways according to ρ and the causal diagram G, obtaining an adapted distribution $P^{\rho}_{\mathrm{adapt}}$. Generation then amounts to fitting and sampling from the conditional distribution required by the query. In our running example, sex and age are supplied, while education, employment, and salary are requested:

$$P_{\mathrm{rw}}(X, \mathbf{Z}, \mathbf{Y}) \xrightarrow{\mathsf{Adapt}(G,\rho)} P^{\rho}_{\mathrm{adapt}}(X, \mathbf{Z}, \mathbf{Y}) \xrightarrow{\mathsf{Fit}} \widehat{P}^{\rho}_{\mathrm{adapt}}(\mathbf{Y} \mid X, \mathbf{Z}). \tag{7}$$

The first step moves from the observed population into the fair world specified by ρ; the second extracts the conditional needed for the particular query. Because this query is topological, with X and $\mathbf{Z}$ supplied before their generated descendants, the resulting conditional also admits the usual interventional interpretation in the adapted SCM, i.e., $P^{\rho}_{\mathrm{adapt}}(\mathbf{Y} \mid X = x, \mathbf{Z} = \mathbf{z}) = P^{\rho}_{\mathrm{adapt}}(\mathbf{Y} \mid do(X = x), \mathbf{Z} = \mathbf{z})$ under the SFM assumptions.

Non-topological queries. The flexible nature of LLMs also allows for non-topological queries:

Example 1 (Applicant Profiles – continued)**.** *The analyst provides the following query to the model:*

> A 40-year-old applicant has a bachelor's degree, is employed full-time, and earns $52,000 annually. Complete the profile by providing the applicant's sex.

Here, $\mathbf{Z}$ *and* $\mathbf{Y}$ *are supplied while* X *is requested, to be sampled from* $P_{\mathrm{LLM}}(X \mid \mathbf{Z}, \mathbf{Y})$. □

Since the supplied covariates include descendants of X, the query is called *non-topological*, and it is a unique feature of generative models. Interestingly, our previous recipe for fair generation can be adapted appropriately: we first move into the fair world $P^{\rho}_{\mathrm{adapt}}$ and only then condition on the attributes supplied by the query. Thus,

$$P_{\mathrm{rw}}(X, \mathbf{Z}, \mathbf{Y}) \xrightarrow{\mathsf{Adapt}(G,\rho)} P^{\rho}_{\mathrm{adapt}}(X, \mathbf{Z}, \mathbf{Y}) \xrightarrow{\mathsf{Fit}} \widehat{P}^{\rho}_{\mathrm{adapt}}(X \mid \mathbf{Z}, \mathbf{Y}). \tag{8}$$

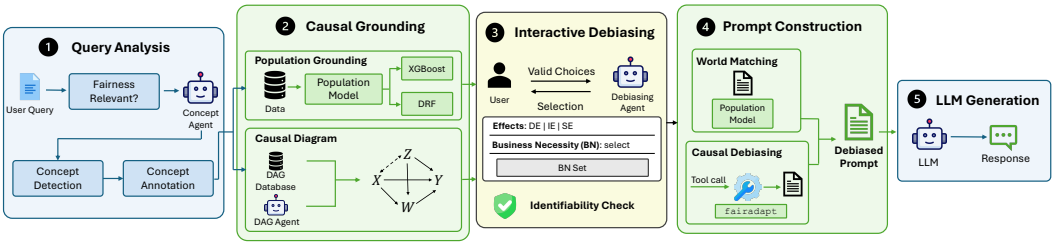


Figure 2: Overview of CFG for causally fair LLM generation. A concept agent extracts and annotates concepts from a fairness-relevant query. CFG then estimates a population model and obtains a causal diagram. A debiasing agent presents the user with supported debiasing choices, checks their compatibility across outputs, and records the selected effects and business-necessity assumptions. Finally, CFG constructs a causally fair output.

The conceptual difference is therefore not which causal pathways are removed, but which conditional of the fair-world distribution is sampled. Unlike the topological case above, $P^{\rho}_{\text{adapt}}(X \mid \mathbf{Z}, \mathbf{Y})$ is purely predictive and has no causal interpretation: conditioning on education, employment, and salary does not treat them as causes of sex.

2.2 Harness for Causally Fair Generation

Our next goal is to turn the selected effect removals and BN exceptions into LLM-generated stories. CFG implements this process as an inference-time harness around the LLM, coordinating population grounding, causal debiasing, and generation from sampled attribute values. Fig. 2 summarizes the five steps of CFG: query analysis, population grounding, causal-diagram acquisition, interactive debiasing, and prompt construction.

(S1) Query analysis. A routing classifier first determines whether an input query raises a potential fairness concern. Selected queries are passed to an LLM-based concept agent, which uses the task instructions and few-shot examples provided in App. D to identify the variables $V_i \in \mathbf{V}$ and normalize their explicitly supplied values. Write $\mathbf{C} = \mathbf{c}$ for the variables supplied in context and $\mathbf{T}$ for those requested for generation. In Ex. 1, $\mathbf{C} = (X, \mathbf{Z})$ and $\mathbf{T} = \mathbf{Y}$. For the non-topological query that generates sex from the remaining attributes, $\mathbf{C} = (\mathbf{Z}, \mathbf{Y})$ and $\mathbf{T} = X$. We also support generation from the full joint distribution, with no supplied attributes: $\mathbf{C} = \varnothing$ and $\mathbf{T} = (X, \mathbf{Z}, \mathbf{Y})$. For partial completion, the query supplies an initial sequence of outcomes, $\mathbf{C} = (X, \mathbf{Z}, \mathbf{Y}_{\leq i})$, and requests the remaining attributes, $\mathbf{T} = \mathbf{Y}_{>i}$. In our example, supplying sex, age, and education leaves employment and salary to be generated.

(S2) Population grounding. We estimate $P_{\text{rw}}(\mathbf{T} \mid \mathbf{C} = \mathbf{c})$ from real-world data of the relevant population, which is assumed to be available. In Ex. 1, this is the joint distribution of education, employment, and salary ($\mathbf{Y}$) in ACS, conditional on the supplied sex and age ($(X, \mathbf{Z})$). The fitted population model can be reused across queries. Practically, conditional distributions are estimated either via survey-weighted XGBoost classifiers, or distributional random forests (DRF) (Cevid et al., 2022). Joint generation requires composing conditional classifiers in topological order or sampling whole records with forest weights. Details are described in App. E.

(S3) Causal-diagram acquisition. We obtain a causal diagram G over $\mathbf{V}$ either by matching the variables to a model in a curated repository or by querying a second LLM-based agent. The latter returns a candidate SFM assigning the variables their roles as X, $\mathbf{Z}$, and $\mathbf{Y}$. In both cases, the diagram is treated as an explicit structural assumption for the subsequent analysis. App. F evaluates the agent's ability to propose a causal diagram from the variables in a query.

(S4) Interactive debiasing. Our goal is to remove the user-selected direct, indirect, and spurious components while retaining pathways exempted by BN, as specified in ρ. We therefore pre-process the data so as to achieve the causal fairness constraint, as described in Eq. 7. In practice, we use the `fairadapt` method (Plecko & Meinshausen, 2020), which adapts attributes in causal order using

conditional quantile models, preserving their conditional ranks as parent values change. Further details are provided in Apps. A.1 and A.2. In summary, the user records the removal choices in ρ, and CFG returns an adapted population model:

$$\boxed{\mathsf{Adapt}(G, \rho)\colon\ \widehat{P}_{\mathrm{rw}} \longmapsto \widehat{P}^{\rho}_{\mathrm{adapt}}} \tag{9}$$

The interface additionally checks that the selected removals admit an identifiable, jointly consistent adaptation across all target variables. When no effects are selected for removal, generation is calibrated to the real-world reference distribution $\widehat{P}_{\mathrm{rw}}$, preserving its observed disparities.

(S5) Prompt construction. For each query, we jointly sample the requested values $\mathbf{t} \sim \widehat{P}^{\rho}_{\mathrm{adapt}}(\mathbf{T} \mid \mathbf{C} = \mathbf{c})$ and insert them into the prompt as fixed constraints alongside the supplied values $\mathbf{c}$. The LLM expresses this record in a story. We then verify whether the sampled covariates appear in the story correctly, and if needed, re-prompt the model.

Caching and reuse. For a fixed population and set of removal choices, we cache population and adaptation models, causal diagrams, compatibility checks, and retrieval conditionals. Only query analysis, conditioning, record sampling, and LLM generation need be repeated for each request.

2.3 Theoretical Guarantees

Let $\rho^\star$ denote the debiasing policy that removes all directed effects of X on $\mathbf{Y}$, such as `fairadapt` used by CFG. At the population level, this corresponds to recursive substitution (RS):

$$Y_i^{\rho^\star} := f_i^{\mathrm{rw}}\Big(x_0, \mathbf{Z}, \mathbf{Y}^{\rho^\star}_{<i}, U_i\Big), \qquad i = 1, \ldots, k, \tag{10}$$

so that $\mathbf{Y}^{\rho^\star} = \mathbf{Y}_{x_0}$. In practice, CFG implements this operation using `fairadapt`.

Theorem 1 (Debiasing under recursive substitution)**.** *Under $\rho^\star$, the following guarantees hold.*

(i) Topological queries. *For every $\mathbf{z}$, and every x, we have*

$$P^{\rho^\star}_{\mathrm{adapt}}(\mathbf{y} \mid x, \mathbf{z}) = P(\mathbf{Y}_{x_0} = \mathbf{y} \mid \mathbf{Z} = \mathbf{z}) = P_{\mathrm{rw}}(\mathbf{y} \mid x_0, \mathbf{z}), \tag{11}$$

implying $\mathbf{z}$-conditional parity. Further, if the user supplies $(X, \mathbf{Z}) \sim P^{\mathrm{user}}$ with $P^{\mathrm{user}}(\mathbf{z} \mid x) = P^{\mathrm{user}}(\mathbf{z})$, then the induced joint distribution $P^{\mathrm{user}}(x, \mathbf{z}) P^{\rho^\star}_{\mathrm{adapt}}(\mathbf{y} \mid x, \mathbf{z})$ satisfies $X \perp\!\!\!\perp (\mathbf{Z}, \mathbf{Y})$.

(ii) Non-topological queries. *For every $(\mathbf{z}, \mathbf{y})$,*

$$P^{\rho^\star}_{\mathrm{adapt}}(x \mid \mathbf{z}, \mathbf{y}) = P^{\rho^\star}_{\mathrm{adapt}}(x \mid \mathbf{z}) = P_{\mathrm{rw}}(x \mid \mathbf{z}). \tag{12}$$

Thus the generated X is insensitive to the supplied $\mathbf{Y}$ once $\mathbf{Z}$ is fixed. If the user supplies $(\mathbf{Z}, \mathbf{Y}) \sim P^{\mathrm{user}}$ with $P^{\mathrm{user}}(\mathbf{z}, \mathbf{y}) = P^{\rho^\star}_{\mathrm{adapt}}(\mathbf{z}, \mathbf{y})$, then $P^{\mathrm{user}}(\mathbf{z}, \mathbf{y}) P^{\rho^\star}_{\mathrm{adapt}}(x \mid \mathbf{z}, \mathbf{y}) = P^{\rho^\star}_{\mathrm{adapt}}(x, \mathbf{z}, \mathbf{y})$, so the complete generated population coincides with the debiased joint.

The theorem considers two cases. For topological queries, Eq. 11 states that after recursive substitution the distribution of the generated record $\mathbf{Y}$ is the same for every supplied value of X, conditional on $\mathbf{Z}$. The guarantee is therefore $\mathbf{z}$-conditional. Additionally, if the user supplies $(X, \mathbf{Z})$ from a distribution in which X and $\mathbf{Z}$ are independent, this strengthens the conditional guarantee to the independence of X from the complete record $(\mathbf{Z}, \mathbf{Y})$. For non-topological queries, Eq. 12 gives the corresponding reverse-direction guarantee: for any input $(\mathbf{z}, \mathbf{y})$, the generated X does not depend on $\mathbf{y}$ once $\mathbf{z}$ is fixed. Importantly, this pointwise statement does not require the user-supplied $(\mathbf{z}, \mathbf{y})$ to have been drawn from the debiased population. Additionally, however, if the user's inputs $\mathbf{Z}, \mathbf{Y}$ follow the debiased marginal $P^{\rho^\star}_{\mathrm{adapt}}(\mathbf{z}, \mathbf{y})$, then the resulting records follow the complete debiased joint distribution $P^{\rho^\star}_{\mathrm{adapt}}(x, \mathbf{z}, \mathbf{y})$.

More generally, the user controls the distribution of the supplied variables $\mathbf{C}$, while CFG controls only the conditional distribution of the generated variables $\mathbf{T}$. Hence the resulting joint distribution is $P^{\mathrm{user}}(\mathbf{C}) P^{\rho^\star}_{\mathrm{adapt}}(\mathbf{T} \mid \mathbf{C})$, and coincides with the fully debiased joint only when $P^{\mathrm{user}}(\mathbf{C}) = P^{\rho^\star}_{\mathrm{adapt}}(\mathbf{C})$. This observation also covers hybrid queries such as $\mathbf{C} = (X, \mathbf{Z}, \mathbf{Y}_{\le i})$: supplied descendants must either already follow the debiased world or be overwritten by their recursively substituted values. We discuss these alternatives in App. A.3.

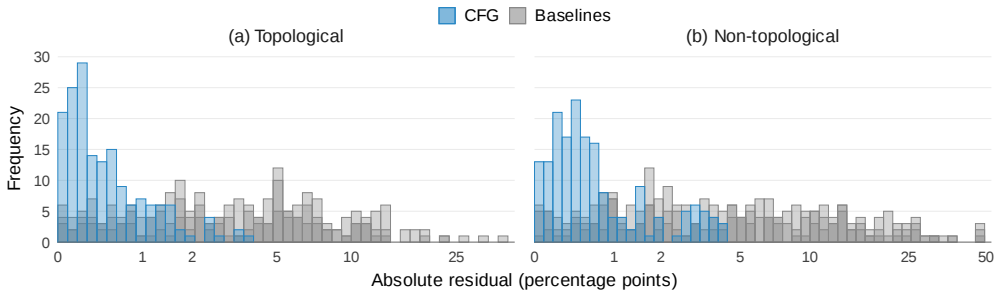


Figure 3: Absolute recovery residuals under topological and non-topological queries.

3 EXPERIMENTS

We evaluate whether CFG can reproduce the causal effects of a reference population in LLM-generated stories and whether its causal debiasing procedure removes selected sources of disparity from those stories. Both evaluations examine intermediate (Y_i, $i < k$) and final outcomes (Y_k) to assess causal fairness at these stages of the generated sequence.

Evaluation settings. We evaluate generation in three real-world settings: (i) income and substance use, studying racial disparities using the National Survey on Drug Use and Health (NSDUH) (Substance Abuse and Mental Health Services Administration, 2023); (ii) obesity and diabetes, studying racial disparities using the Behavioral Risk Factor Surveillance System (BRFSS) (Centers for Disease Control and Prevention, 2023); and (iii) employment and salary, studying sex disparities using the American Community Survey (ACS) (U.S. Census Bureau, 2023). In NSDUH and BRFSS, X is race and $\mathbf{Z}$ contains age and sex; the generated attributes $\mathbf{Y}$ are education, income, and marijuana use for NSDUH, and education, income, BMI, exercise, and diabetes for BRFSS. In ACS, X is sex, $\mathbf{Z}$ contains age, race, and region, and $\mathbf{Y}$ comprises education, working hours, employer type, and salary. We also study sex disparities in employment and salary using Census-Synth, a synthetic population with known ground-truth causal effects, where X is sex, $\mathbf{Z}$ is age, and $\mathbf{Y}$ comprises education, employment, and salary. Fig. 6 of App. C presents all the causal diagrams. App. B provides synthetic data details, while App. C provides extended experimental results.

We evaluate four generators: Qwen3-4B (Yang et al., 2025), Llama 3-8B (Grattafiori et al., 2024), Qwen3.8-27B (Qwen Team, 2026), and Gemma 4-31B (Gemma Team et al., 2026). For each model, CFG is compared against three prompting baselines. *Vanilla* provides only the task and supplied attributes, with no external data or retrieval; its generations therefore reflect the distribution induced by the model's learned weights and the prompt. *Chain-of-thought (CoT)* adds a request for intermediate reasoning to Vanilla, with the same information available. *Retrieval-augmented generation (RAG)* supplies the relevant estimated conditional probabilities directly in the prompt. RAG and CFG use the same fitted population model. App. C.3 details the baseline construction and comparison.

The main topological experiments generate all attributes, $\mathbf{C} = \varnothing$ and $\mathbf{T} = (X, \mathbf{Z}, \mathbf{Y})$; non-topological queries supply $\mathbf{C} = (\mathbf{Z}, \mathbf{Y})$ and request $\mathbf{T} = X$. Completion with supplied $(X, \mathbf{Z})$, as in Ex. 1, is evaluated in App. H. For all generators and methods, a fixed Qwen3.8-27B annotator extracts the requested attributes $\mathbf{T}$ from generated stories as categorical values, which then form the data used to estimate causal effects. App. C.2 details the annotation procedure.

3.1 MATCHING REFERENCE CAUSAL EFFECTS

We first test whether CFG can reproduce the reference population's causal effects before applying any bias removal. We compare TV and its DE, IE, and SE components in generated stories with their reference values for one intermediate and one final target per dataset. We use a cross-fitted one-step estimator (Chernozhukov et al., 2018) for effects in generated stories and survey reference effects; Census-Synth references are computed from the known SCM by Monte Carlo simulation. For each method, we compute absolute errors of the fairness measures relative to the reference values. App. C provides outcome definitions, estimation and aggregation details, and mean errors by method. Fig. 3 pools absolute errors in TV, DE, IE, and SE across generators, datasets, and targets. Under both query types, CFG matches reference effects better than baselines, with errors concentrating near

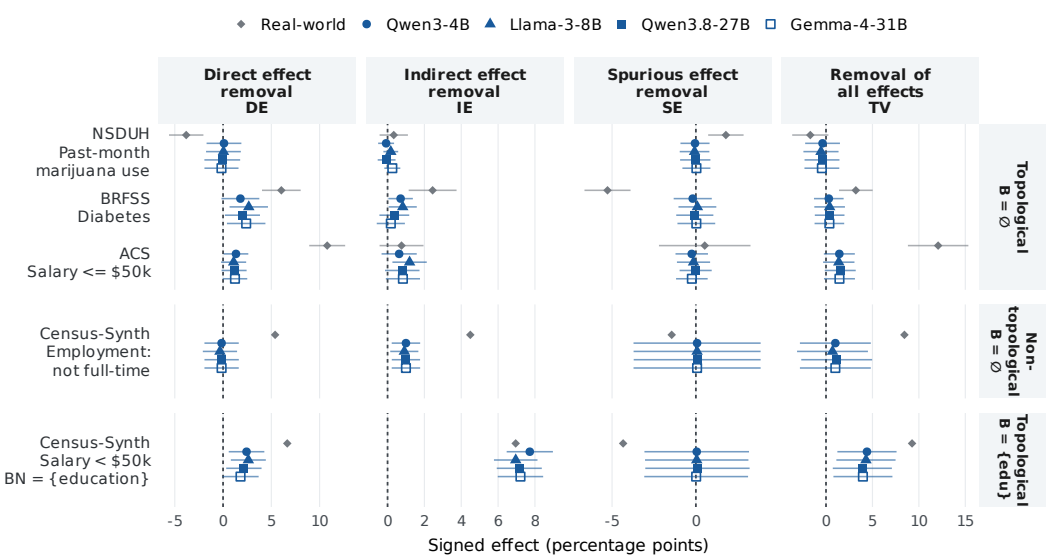


Figure 4: Rows 1–3: topological queries with empty BN set; Row 4: non-topological query; Row 5: topological query with BN requirement. Intervals indicate 95% confidence intervals.

zero, while baselines show broader distributions with larger errors. Thus, even in the absence of fairness considerations, CFG can align generation with real-world distributions.

3.2 Removing Direct, Indirect, and Spurious Effects

We next evaluate CFG by removing DE, IE, or SE individually, or all three together, for the same outcomes. We first use an empty BN set, so no pathways are exempted from removal. Generated stories are evaluated using the same annotator and effect estimator as above. We study topological queries across all four datasets and non-topological queries for Census-Synth employment. We additionally study Census-Synth salary with a BN set $\mathbf{B} = \{\text{education}\}$.

CFG aligns fairness behavior across LLM architectures. The first four rows of Fig. 4 show that CFG brings the selected effects toward zero for both topological queries and non-topological queries that generate sex from age, education, and employment. The closely aligned estimates across Qwen, Llama, and Gemma show that CFG induces similar causal fairness behavior, as measured by the targeted effects, across the tested model families and sizes. When all effects are selected for removal, the 95% confidence intervals for TV include zero in all 20 empty-BN comparisons, corresponding to five dataset-outcome pairs across four generators (see App. C.6 for details).

Effect removal with a non-empty BN set. Returning to the applicant example in Ex. 1, the analyst permits education-mediated sex effects on salary. We examine this choice using Census-Synth as a controlled analogue, with $\mathbf{B} = \{\text{education}\}$. BN permits these pathways and retains larger IE and TV than full effect removal, as shown for topological queries in the bottom row of Fig. 4. A remaining salary gap can therefore reflect the analyst's chosen exceptions. The same pattern holds under non-topological queries; App. C.6 provides extended results for BN experiments.

Limitations, Strengths, and Future Work. An important limitation is that CFG assumes access to relevant real-world data through its population-grounding component; such data may be unavailable or insufficiently representative of the target population. Causal effect removal also requires a credible causal diagram, and LLM-proposed diagrams may require domain-expert correction. In addition, CFG treats the pre-exposure covariates $\mathbf{Z}$ as a single block, retaining or removing their association with X jointly. CFG, however, requires no access to the generator's weights or gradients, making it applicable to black-box LLMs, and supports both topological and non-topological queries. Future work should address limited population data, sensitivity to causal-diagram uncertainty, and selective adjustment of individual background covariates while accounting for their dependencies.

AI USE STATEMENT

We used AI tools to improve grammar and correct typographical errors in the manuscript, and to assist with code for plotting experimental results. The original implementation of our method was developed without generative AI assistance. We also used AI tools to help create a subset of the evaluation examples used to test concept detection in our ablation studies. We reviewed and verified all AI-assisted edits, plotting code, and resulting figures for correctness and consistency with the underlying results.

ETHICS STATEMENT

This work aims to reduce specified sources of disparity in LLM-generated content. Our experiments use publicly available survey data and a synthetic population. Generated stories are intended as fictional examples, and their statistical patterns should not be interpreted as inherent characteristics of demographic groups. The choice of causal diagram and business-necessity exceptions involves substantive judgments about which pathways are justifiable. These choices require domain expertise and consideration of affected groups. The method's fairness conclusions depend on these assumptions and on the representativeness of the population data. Reducing the evaluated causal effects does not guarantee that generated content is free of stereotypes or other harms. Applications in consequential settings therefore require additional validation and human oversight.

REPRODUCIBILITY STATEMENT

The source code, computational environment, setup instructions, and instructions for reproducing all experiments are available in our anonymized repository: `https://anonymous.4open.science/r/cfg/`. App. A provides the theoretical assumptions and proofs. Apps. B, C and H describe the synthetic data-generating process, data preparation, and experimental protocols, while Apps. D to F detail the component evaluations.

REFERENCES

Ahmed Abdulaal, adamos hadjivasiliou, Nina Montana-Brown, Tiantian He, Ayodeji Ijishakin, Ivana Drobnjak, Daniel C. Castro, and Daniel C. Alexander. Causal modelling agents: Causal graph discovery through synergising metadata- and data-driven reasoning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=pAoqRlTBtY.

Haozhe An, Christabel Acquaye, Colin Wang, Zongxia Li, and Rachel Rudinger. Do large language models discriminate in hiring decisions on the basis of race, ethnicity, and gender? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 386–397, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.37. URL https://aclanthology.org/2024.acl-short.37/.

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, 122(8):e2416228122, 2025. doi: 10.1073/pnas.2416228122. URL https://www.pnas.org/doi/abs/10.1073/pnas.2416228122.

Herbert N Bernhardt. Griggs v. duke power co.: The implications for private and public employers. *Tex. L. Rev.*, 50:901, 1971.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc.,

2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System: 2023 Survey Data and Documentation, 2023. URL https://www.cdc.gov/brfss/annual_data/annual_2023.html. Accessed: 2026-09-09.

Domagoj Cevid, Loris Michel, Jeffrey Näf, Peter Bühlmann, and Nicolai Meinshausen. Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *Journal of Machine Learning Research*, 23(333):1–79, 2022.

Jiaee Cheong, Sinan Kalkan, and Hatice Gunes. Counterfactual fairness for facial expression recognition. In *Computer Vision – ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pp. 245–261, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-25071-2. doi: 10.1007/978-3-031-25072-9_16. URL https://doi.org/10.1007/978-3-031-25072-9_16.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33 (01), pp. 7801–7808, 2019.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, September 2024. doi: 10.1162/coli_a_00524. URL https://aclanthology.org/2024.cl-3.8/.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, pp. 219–226, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3317950. URL https://doi.org/10.1145/3306618.3317950.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

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

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. Cladder: Assessing causal reasoning in language models, 2024a. URL https://arxiv.org/abs/2312.04350.

Zhijing Jin, Jiarui Liu, Zhiheng LYU, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona T. Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation? In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=vqIH0ObdqL.

Jungseock Joo and Kimmo Kärkkäinen. Gender slopes: Counterfactual fairness for computer vision models by attribute manipulation. In *Proceedings of the 2nd International Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia*, FATE/MM '20, pp. 1–5, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450381482. doi: 10.1145/3422841.3423533. URL https://doi.org/10.1145/3422841.3423533.

Nitish Joshi, Abulhair Saparov, Yixin Wang, and He He. LLMs are prone to fallacies in causal inference. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10553–10569, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.590. URL https://aclanthology.org/2024.emnlp-main.590/.

Sangwon Jung, Sumin Yu, Sanghyuk Chun, and Taesup Moon. Do counterfactually fair image classifiers satisfy group fairness? - a theoretical and empirical study. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023. ISSN 1041-6080. doi: https://doi.org/10.1016/j.lindif.2023.102274. URL https://www.sciencedirect.com/science/article/pii/S1041608023000195.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/f5f8590cd58a54e94377e6ae2eded4d9-Paper.pdf.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality, 2024. URL https://arxiv.org/abs/2305.00050.

Jingling Li, Zeyu Tang, Xiaoyu Liu, Peter Spirtes, Kun Zhang, Liu Leqi, and Yang Liu. Prompting fairness: Integrating causality to debias large language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 63606–63629, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/9f26a2d143a227376dff99a279f93f99-Paper-Conference.pdf.

Chenxi Liu, Yongqiang Chen, Tongliang Liu, Mingming Gong, James Cheng, Bo Han, and Kun Zhang. Discovery of the hidden world with large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 102307–102365. Curran Associates, Inc., 2024. doi: 10.52202/079017-3249. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/b99a07486702417d3b1bd64ec2cf74ad-Paper-Conference.pdf.

Rahul Madhavan, Rishabh Garg, Kahini Wadhawan, and Sameep Mehta. CFL: Causally fair language models through token-level attribute controlled generation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 11344–11358, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.720. URL https://aclanthology.org/2023.findings-acl.720/.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32 (1), 2018.

Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5356–5371, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.416. URL https://aclanthology.org/2021.acl-long.416/.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1953–1967, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.154. URL https://aclanthology.org/2020.emnlp-main.154/.

Office for National Statistics. Enabling data access through privacy preserving synthetic data, November 2023. URL https://datasciencecampus.ons.gov.uk/enabling-data-access-through-privacy-preserving-synthetic-data/. Data Science Campus.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. BBQ: A hand-built bias benchmark for question answering. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.165. URL https://aclanthology.org/2022.findings-acl.165/.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

Drago Plecko. Causal bias detection in generative artificial intelligence, 2026. URL https://arxiv.org/abs/2605.11365.

Drago Plecko and Elias Bareinboim. Causal fairness analysis, 2022. URL https://arxiv.org/abs/2207.11385.

Drago Plecko and Nicolai Meinshausen. Fair data adaptation with quantile preservation. *Journal of Machine Learning Research*, 21(242):1–44, 2020. URL https://www.jmlr.org/papers/v21/19-966.html.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9180–9211, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.625. URL https://aclanthology.org/2022.emnlp-main.625/.

Substance Abuse and Mental Health Services Administration. National Survey on Drug Use and Health (NSDUH), 2023. URL https://www.samhsa.gov/data/data-we-collect/nsduh-national-survey-drug-use-and-health. Accessed: 2026-09-09.

Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. Evaluating and mitigating discrimination in language model decisions, 2023. URL https://arxiv.org/abs/2312.03689.

U.S. Census Bureau. 2023 American Community Survey 1-Year Estimates, 2023. URL https://www.census.gov/data/developers/data-sets/acs-1year/2023.html. Accessed: 2026-09-09.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models, 2024. URL https://arxiv.org/abs/2306.11698.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. Causal parrots: Large language models may talk causality but are not causal. *arXiv preprint arXiv:2308.13067*, 2023.

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32 (1), 2018.

Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4227–4241, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.232. URL https://aclanthology.org/2023.acl-long.232/.

SUPPLEMENTARY MATERIAL: *Causally Fair Generation with Large Language Models*

The source code for reproducing the results can be found in the anonymized code repository: `https://anonymous.4open.science/r/cfg/`. The repository includes a README file explaining the setup steps. All experiments were run on an NVIDIA GH200 GPU, and required 2,000 GPU hours.

A ADDITIONAL THEORETICAL DETAILS

A.1 PROOF OF THM. 1

Recall that under $\rho^\star$, recursive substitution fixes the protected attribute input for every downstream mechanism to the value x_0:

$$Y_i^{\rho^\star} := f_i^{\mathrm{rw}}\Big(x_0, \mathbf{Z}, \mathbf{Y}_{<i}^{\rho^\star}, U_i\Big), \qquad i = 1, \ldots, k. \tag{13}$$

By recursive substitution, $\mathbf{Y}^{\rho^\star} = \mathbf{Y}_{x_0}$.

Proof of Thm. 1. Because the downstream exogenous variables are independent of the exogenous variables generating $(X, \mathbf{Z})$ (according to the SFM in Fig. 1(b)), $\mathbf{Y}_{x_0}$ is a function only of $\mathbf{Z}$ and the downstream exogenous variables. Hence, the counterfactual independence $\mathbf{Y}_{x_0} \perp\!\!\!\perp X \mid \mathbf{Z}$ holds. Using consistency,

$$\begin{aligned} P_{\mathrm{adapt}}^{\rho^\star}(\mathbf{y} \mid x, \mathbf{z}) &= P(\mathbf{Y}_{x_0} = \mathbf{y} \mid X = x, \mathbf{Z} = \mathbf{z}) \\ &= P(\mathbf{Y}_{x_0} = \mathbf{y} \mid \mathbf{Z} = \mathbf{z}) = P_{\mathrm{rw}}(\mathbf{y} \mid x_0, \mathbf{z}), \end{aligned} \tag{14}$$

which proves Eq. 11. The debiased joint can therefore be written as

$$P_{\mathrm{adapt}}^{\rho^\star}(x, \mathbf{z}, \mathbf{y}) = P_{\mathrm{rw}}(x, \mathbf{z})\, P_{\mathrm{rw}}(\mathbf{y} \mid x_0, \mathbf{z}). \tag{15}$$

Conditioning this joint in the reverse direction gives

$$P_{\mathrm{adapt}}^{\rho^\star}(x \mid \mathbf{z}, \mathbf{y}) = \frac{P_{\mathrm{rw}}(x \mid \mathbf{z}) P_{\mathrm{rw}}(\mathbf{y} \mid x_0, \mathbf{z})}{\sum_{x'} P_{\mathrm{rw}}(x' \mid \mathbf{z}) P_{\mathrm{rw}}(\mathbf{y} \mid x_0, \mathbf{z})} = P_{\mathrm{rw}}(x \mid \mathbf{z}), \tag{16}$$

which also equals $P_{\mathrm{adapt}}^{\rho^\star}(x \mid \mathbf{z})$ and proves Eq. 12.

For the topological population statement, if the supplied variables follow $P^{\mathrm{user}}(x, \mathbf{z}) = P^{\mathrm{user}}(x) P^{\mathrm{user}}(\mathbf{z})$, then the resulting joint is

$$P^{\mathrm{user}}(x) P^{\mathrm{user}}(\mathbf{z}) P_{\mathrm{adapt}}^{\rho^\star}(\mathbf{y} \mid \mathbf{z}),$$

which factorizes into a distribution of X and a distribution of $(\mathbf{Z}, \mathbf{Y})$, giving $X \perp\!\!\!\perp (\mathbf{Z}, \mathbf{Y})$. For the non-topological population statement, if $P^{\mathrm{user}}(\mathbf{z}, \mathbf{y}) = P_{\mathrm{adapt}}^{\rho^\star}(\mathbf{z}, \mathbf{y})$, then

$$P^{\mathrm{user}}(\mathbf{z}, \mathbf{y}) P_{\mathrm{adapt}}^{\rho^\star}(x \mid \mathbf{z}, \mathbf{y}) = P_{\mathrm{adapt}}^{\rho^\star}(x, \mathbf{z}, \mathbf{y})$$

by the definition of conditional probability. □

A.2 BUSINESS-NECESSITY POLICIES

Thm. 1 concerns $\rho^\star$, which removes all directed effects of X on the generated outcomes. More general policies may retain selected pathways through business-necessity (BN) exceptions. Identification of such path-specific effects is not always possible. However, backward closure of the BN set gives a sufficient condition for identification (Plecko & Bareinboim, 2022, Thm. 6.11).

In practice, we also allow BN specifications not covered by the above identification conditions. In these cases, stronger assumptions are needed. Specifically, we adopt the quantile-preservation assumption used by `fairadapt` (Plecko & Meinshausen, 2020): conditional latent ranks are preserved when parent values are changed, thereby specifying the joint distribution over counterfactual outcomes. Under this joint coupling, all the resulting quantities are identified.

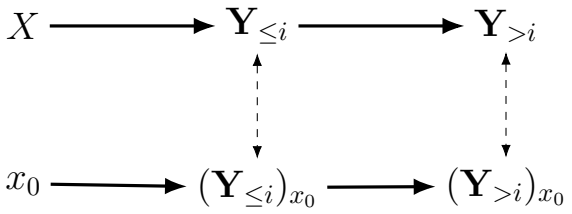


Figure 5: Twin-network over counterfactual outcomes, related to the hybrid completion setting where the user supplies descendants of X. The upper path represents the factual values of X and the supplied prefix $\mathbf{Y}_{\leq i}$, and the generated suffix $\mathbf{Y}_{>i}$; the lower path represents the recursively substituted world obtained by setting $X = x_0$.

A.3 Hybrid Completion Queries

Thm. 1 considers two canonical settings, where either $(X, \mathbf{Z})$ are supplied and the entire outcome vector $\mathbf{Y}$ is generated, or $(\mathbf{Z}, \mathbf{Y})$ are supplied and X is generated. A more general query may instead supply $(X, \mathbf{Z}, \mathbf{Y}_{\leq i})$ and request the remaining attributes $\mathbf{Y}_{>i}$. Since the user supplies the values of X and $\mathbf{Y}_{\leq i}$, the input distribution generated by the user may encode arbitrary relationships of $X \to \mathbf{Y}_{\leq i}$, which are not under the control of our method. If $\mathbf{C}$ denotes the user-supplied variables and $\mathbf{T}$ the variables generated by CFG, the resulting distribution is

$$P^{\text{user}}(\mathbf{C})\, P^{\rho^\star}_{\text{adapt}}(\mathbf{T} \mid \mathbf{C}).$$

Thus, recovering the complete debiased joint distribution depends on the inputs of the user, and we discuss two possible cases below.

Fair-world user inputs. The first possibility is that the user supplies $(X, \mathbf{Z}, \mathbf{Y}_{\leq i})$ according to their joint marginal distribution in the debiased world. In particular, the supplied descendants are then realizations of $(\mathbf{Y}_{\leq i})_{x_0}$, as shown in Fig. 5. CFG then samples the remaining attributes from the corresponding debiased conditional,

$$P^{\rho^\star}_{\text{adapt}}(\mathbf{Y}_{>i} \mid X, \mathbf{Z}, \mathbf{Y}_{\leq i})\,.$$

Under this interpretation, the supplied and generated attributes belong to the same debiased population, and the completed record follows the adapted distribution $P^{\rho^\star}_{\text{adapt}}(x, \mathbf{z}, \mathbf{y})$.

Real-world user inputs. Alternatively, the supplied $\mathbf{Y}_{\leq i}$ may be factual values from the real world, corresponding to the upper path in Fig. 5. These values already encode part of the factual $X \to \mathbf{Y}$ relationship. Consequently, simply appending a recursively substituted suffix $(\mathbf{Y}_{>i})_{x_0}$ combines factual and counterfactual components in the generated record, and should not be interpreted as a draw from the fully debiased population. In this case, the natural way forward is to overwrite the supplied descendants with their recursively substituted counterparts $(\mathbf{Y}_{\leq i})_{x_0}$ and then generate $(\mathbf{Y}_{>i})_{x_0}$, thereby moving the entire outcome vector to the lower path in Fig. 5. In this case, some of the user-provided covariates are overwritten, but the generated records can again be shown to follow the adapted distribution $P^{\rho^\star}_{\text{adapt}}(x, \mathbf{z}, \mathbf{y})$.

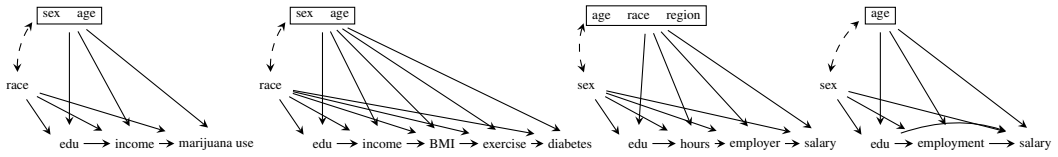
B CENSUS-SYNTH STRUCTURAL CAUSAL MODEL

We generate $n = 8{,}192$ observations from the following ordinal SCM. Let $U, \epsilon_A, \epsilon_E, \epsilon_M, \epsilon_S \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, let $\text{cut}_{\boldsymbol{\tau}}(r)$ map a latent score r to the ordered category determined by the cut points $\boldsymbol{\tau}$, and let $s(V)$ denote the fixed centered category code of V. The structural assignments are

$$\begin{aligned}
\text{sex} &\leftarrow \text{Bernoulli}\big(\text{logit}^{-1}(0.9U)\big)\,, \\
\text{age} &\leftarrow \text{cut}_{(-0.8,0,0.8)}(0.9U + \epsilon_A), \\
\text{education} &\leftarrow \text{cut}_{(-0.8,0.1,1.0)}(-0.15 - 0.35\,\text{sex} + 0.35s(\text{age}) + \epsilon_E), \\
\text{employment} &\leftarrow \text{cut}_{(-0.45,0.65)}(-0.10 - 0.25\,\text{sex} - 0.10s(\text{age}) + 0.75s(\text{education}) + \epsilon_M), \\
\text{salary} &\leftarrow \text{cut}_{(-0.65,0.25,1.15)}\begin{pmatrix} -0.25 - 0.35\,\text{sex} + 0.25s(\text{age}) \\ +\,0.55s(\text{education}) + 0.75s(\text{employment}) + \epsilon_S \end{pmatrix}.
\end{aligned} \tag{17}$$

Age, education, and salary have four ordered categories, whereas employment has three. Salary uses the brackets below \$25k, \$25k–\$49,999, \$50k–\$99,999, and \$100k or more, which also define its four multiple-choice annotation categories. The shared latent variable U induces the bidirected sex $\leftrightarrow$ age relation; all remaining parental relations in Eq. 17 are shown in Fig. 6(d). The education term in the salary assignment allows an effect at fixed employment status. Both sex $\rightarrow$ education $\rightarrow$ salary and sex $\rightarrow$ education $\rightarrow$ employment $\rightarrow$ salary contribute to the indirect effect of sex on salary; the direct effect of sex concerns the edge sex $\rightarrow$ salary.

Ground-truth effects. The structural equations specify how each unit responds to interventions, allowing us to compute the counterfactual quantities in the effect definitions directly from the SCM. We evaluate factual outcomes, outcomes under interventions on sex, and nested counterfactual outcomes in which the earlier outcomes take their values under the reference sex. These are the x-specific direct, indirect, and spurious effects defined in Plecko & Bareinboim (2022) and used in Eq. 6; we retain the paper's signed contribution convention for IE and SE. We approximate the required probabilities using $1{,}000{,}000$ Monte Carlo draws of the exogenous variables, reusing each draw across the factual and counterfactual evaluations and conditioning on the factual sex group as required by the definitions. TV is computed from the factual group probabilities. These calculations provide a Monte Carlo approximation to the SCM ground truth for all Census-Synth reference comparisons.



(a) NSDUH. **(b)** BRFSS. **(c)** Census Income. **(d)** Census-Synth

Figure 6: Standard fairness models for NSDUH, BRFSS, ACS, and Census-Synth. Generated attributes appear in topological order from left to right. In panels (a)–(c), nonconsecutive forward edges are omitted for clarity; each generated attribute also points to every later attribute. Panel (d) is shown in full.

C Effect Recovery and Removal: Additional Details

C.1 Targets and Evaluation Protocol

Fig. 6 presents the causal diagrams for all four datasets.

NSDUH, BRFSS, and ACS use publicly available 2023 U.S. survey data. We sample $n = 8{,}192$ records per survey using the survey weights and reuse the same records across methods and generators. Census-Synth contains $n = 8{,}192$ observations generated from the ordinal SCM over sex, age, education, employment, and salary specified in App. B.

We evaluate one intermediate and one final target per dataset: income and monthly marijuana use for NSDUH, obesity and diabetes for BRFSS, employer type and salary for ACS, and employment and salary for Census-Synth. For target Y_i, the analysis includes X, the covariates $\mathbf{Z}$, and all preceding outcomes $\mathbf{W}^{(i)} = \mathbf{Y}_{<i}$; later outcomes are excluded. The protected contrast is Black or Hispanic versus White for NSDUH and BRFSS, and female versus male for ACS and Census-Synth, with the first group encoded as x_1 and the second as x_0.

The prespecified outcome events are income below \$50,000 and past-month marijuana use for NSDUH; obesity (classes 1–3) and diabetes for BRFSS; salary at most \$50,000 for ACS; and non-full-time employment and salary below \$50,000 for Census-Synth. The four nominal ACS employer categories are evaluated as separate one-versus-rest events. This gives seven intermediate and four final events. All methods use the same source records and event definitions.

We report DE, IE, and SE as additive contributions to TV, reversing the signs of the IE and SE terms in Eq. 6. Under this reporting convention, $\mathrm{TV} = \mathrm{DE} + \mathrm{IE} + \mathrm{SE}$; a positive component contributes toward a higher probability of the outcome event for $X = x_1$ relative to $X = x_0$.

We use a cross-fitted one-step estimator for effects in annotated generations for all datasets, including Census-Synth. For NSDUH, BRFSS, and ACS, we also estimate the reference effects from each survey-weighted sample of $8{,}192$ records. Census-Synth reference effects use the Monte Carlo approximation to SCM ground truth described in App. B. All effects are in percentage points. The experiment uses one generation seed; 95% Wald intervals quantify estimator uncertainty, not variation across generation runs. The SCM ground-truth values are treated as fixed references in the plots.

For removal, we apply each setting separately using `fairadapt` (Plecko & Meinshausen, 2020). Its SE-removal implementation represents the X–$\mathbf{Z}$ association by an $X \to \mathbf{Z}$ adaptation edge. Topological generation verbalizes a sampled debiased record. For non-topological generation, a reverse conditional is fitted to the debiased joint before sampling $X \mid \mathbf{Z}, Y_1, \ldots, Y_i$ and verbalizing it; we evaluate this procedure on Census-Synth employment with an empty BN set and salary with $\mathbf{B} = \{\text{education}\}$ (App. C.6).

C.2 Annotating Generated Stories

The same Qwen3.8-27B annotator extracts categorical attributes from generated stories for all methods, generators, and datasets, in both the matching and effect-removal experiments.

Extracting categories. For each story, the annotator is queried separately for every attribute requested from the generator. Its prompt contains the story, a question about the attribute, and a lettered list of the dataset's category labels, followed by an *Answer not available* option. The model first identifies the person described, gives a brief explanation based on the text, and then selects an option. It returns these fields in a structured JSON response. The instructions specify that, if several people or narratives appear, only the first is considered. Contradictory or unclear information should lead to *Answer not available*. BRFSS prompts additionally include fixed, attribute-specific examples; for education, the instructions select the highest completed level when several compatible levels are mentioned.

For example, an NSDUH story stating that the person used marijuana during the past month is assigned the category *yes*. The following prompt shows the category-selection format, with the generated story inserted at the indicated placeholder:

Prompt

```
Rules:
1.  If there are multiple people or narratives, focus only on the
first one.  Ignore any apparent rewrite, restart, or second attempt
later in the story.
2.  If the same person's information is contradictory, choose
Answer not available.
3.  If the answer is not reasonably clear, choose Answer not
available rather than guessing.

Question:  determine the person's marijuana use in the past month.

A. no
B. yes
C. Answer not available

First identify who the story is about, briefly explain your
reasoning, and only then choose the option letter.  Respond with
one line of JSON and no other text:
{"target_person":  "<person>", "explanation":  "<brief reasoning>",
"answer":  "<letter>"}

<story>
[Generated story]
</story>
```

In this example, the returned letter `B` is mapped to *yes*. Analogous questions recover education, income, and the other requested attributes using their own category lists.

Decoding and incomplete annotations. Annotation uses greedy decoding with temperature zero, top-p equal to one, and a maximum of 300 output tokens per question; the model's thinking mode is disabled. The returned option letter is parsed and mapped to its category. Abstentions and responses without a valid option are recorded as missing. For each query, the generation pipeline provides up to four candidate stories. We annotate them in order and select the first for which every requested attribute has a valid category. If none is fully annotated, the final candidate is retained with its unresolved attributes marked as missing.

From stories to analysis data. We save the selected categories together with the story, annotation prompts, raw responses, explanations, and sample identifier. Each selected story then contributes one row to the analysis dataset. Supplied attributes retain their query values, while requested attributes take the values extracted from the story. Thus, in CFG, attributes sampled for the generator to verbalize are recovered from the resulting text for evaluation. Category labels and their ordering match those in the reference data. The resulting records are used to estimate TV, DE, IE, and SE following App. C.1.

Agreement with sampled values. For CFG, we compare each attribute value sampled for the generator with the category subsequently selected by the annotator. We report the percentage of

Table 1: Sampled–annotated agreement for CFG.

Generator	Agreement (%)
Qwen3-4B	98.96
Llama 3-8B	98.73
Qwen3.8-27B	99.75
Gemma-4-31B	99.99

these pairs that match exactly; missing annotations, including *Answer not available* (NA), count as nonmatches. Each requested attribute contributes one comparison, using the selected story after the retry procedure above. Attributes supplied to the query and not annotated are excluded.

Tbl. 1 reports this agreement for topological recovery on the same available dataset–outcome combinations for all four generators: NSDUH income and marijuana use, BRFSS obesity and diabetes, ACS employer type, and Census-Synth employment. Each generator contributes $49{,}152$ selected stories and $294{,}912$ sampled–annotated pairs. Agreement ranges from 98.73% to 99.99%.

Disagreements include both NA answers and selections of a category different from the sampled value; the latter account for most nonmatches in these runs. An NA answer can arise when the generated story omits an attribute or expresses it ambiguously, or when the annotator fails to identify information in the text. A different selected category can similarly reflect a change introduced during generation or an annotation error. This comparison therefore measures how well the sampled values are preserved through generation and annotation together.

C.3 BASELINE CONSTRUCTION

For each query, $\mathbf{C} = \mathbf{c}$ denotes the supplied attributes and $\mathbf{T}$ the requested attributes, as in Sec. 2.2. The three baselines differ in the information and reasoning instructions given to the generator; they do not adapt the population distribution to remove selected causal effects.

Vanilla. The model receives the story-generation task and the supplied values $\mathbf{c}$, without external data, retrieved statistics, or tool access. The requested values emerge from the model's learned weights under this prompt.

Chain-of-thought. CoT uses the Vanilla setup with an additional instruction to reason through the task before producing the story. It receives no additional population information.

Retrieval-augmented generation. For each conditioning task, a 500-tree distributional random forest fitted to the reference records estimates the conditional category probabilities $\widehat{P}_{\mathrm{rw}}(V_i = v \mid \mathbf{C} = \mathbf{c})$ for each requested $V_i \in \mathbf{T}$. These probabilities are inserted directly into the generation prompt. For example, when requesting $X =$ sex given age, education, employment, and salary, the prompt might report "male: 40%, female: 60%" for that supplied context. These percentages illustrate the prompt format; the retrieved values depend on the query. The LLM selects the requested values using this information and writes the story; no category is fixed by an external draw before generation.

Comparison with CFG. In the recovery experiment, RAG and CFG share the reference records and fitted forest. RAG presents conditional marginal probabilities, whereas CFG uses the forest's observation weights to sample a complete record for $\mathbf{T}$ conditional on $\mathbf{C} = \mathbf{c}$, preserving dependencies among the requested attributes. The sampled values become fixed prompt constraints. For non-topological queries requesting only X, the distinction is between providing its probabilities and externally sampling its value. For bias removal, CFG instead samples from the adapted distribution $\widehat{P}^{\rho}_{\mathrm{adapt}}$. All methods use the same generator set, annotation procedure, and outcome definitions.

C.4 AGGREGATION

For the tabulated mean absolute errors, we first average the four ACS employer events within ACS, then weight the datasets equally. Removal summaries additionally weight generators equally and compare absolute effect magnitudes before and after debiasing on matched event–generator pairs. When all effects are selected for removal, we report residual DE, IE, and SE alongside TV to assess each component of the remaining disparity.

Aggregate errors are computed over outcome contrasts with valid generations, annotations, and estimable effects; contrasts for which filtering leaves only one protected group are excluded. The one-step estimator can produce values outside the probability scale under weak overlap. Available estimate counts can therefore differ between methods. The plotting conventions are described in App. C.5. The large Gemma 4-31B non-topological RAG error is driven by an unstable NSDUH estimate under weak overlap and should not be interpreted as a probability-scale disparity.

C.5 MATCHING REFERENCE EFFECTS ACROSS OUTCOMES AND QUERY ORDERS

We break down the reference-effect errors in Fig. 3 by generator, query order, and target position, using the outcomes in App. C.1.

Recovery throughout the generated sequence. Under topological queries, CFG's intermediate TV errors range from 0.40 to 0.58 percentage points, and its DE, IE, and SE errors are at most 0.82, 0.41, and 0.27 points, respectively (Tbl. 2). For example, Llama 3-8B achieves an intermediate TV error of 0.58 points, compared with 6.53 for vanilla prompting and 6.93 for RAG. For final outcomes, mean absolute TV errors range from 0.49 to 0.66 points across the four generators. With Llama 3-8B, CFG's final-outcome TV error is 0.63 points, compared with 4.14 for Vanilla, 8.92 for CoT, and 12.42 for RAG. Across both target positions and all four generators, mean absolute errors in DE, IE, and SE are at most 0.82 points. Thus, CFG closely matches both aggregate disparity and its constituent effects at intermediate and final attributes.

Conditioning on descendants. Under non-topological queries, CFG's TV errors rise to 1.58–1.68 points for intermediate outcomes and 1.66–1.80 for final outcomes. Component errors remain at most 2.18 for DE, 0.88 for IE, and 0.53 for SE across both target positions. For Llama 3-8B, final TV error is 1.66 points with CFG, compared with 10.27 for Vanilla, 7.33 for CoT, and 10.79 for RAG. Thus, externally sampling X conditional on the supplied descendants yields more accurate reference-effect matching than presenting its probabilities to the LLM in this comparison.

Histogram construction. Each residual r in Fig. 3 is the absolute difference between a generated-effect estimate and its reference effect, in percentage points. Bins have width 0.08 on the $\log(1+r)$ scale, with axis labels in the original units and separate horizontal ranges for the two panels. CFG is shown in blue; the three baseline histograms are overlaid in grey under a shared legend entry. Bar heights are unnormalized counts of event–effect pairs for each method, without the dataset weighting used in App. C.4. The 50-point display cutoff omits 2 of 580 topological and 17 of 672 non-topological residuals; all remain in the reported averages and the cumulative distributions below.

Full residual distributions. Fig. 7 complements the averages in Tbl. 2 with empirical cumulative distributions for each method. At a threshold r, the curve gives the proportion of absolute residuals at or below r. Each curve pools effects, generators, datasets, outcome events, and target positions with equal weight per available residual. All 580 topological and 672 non-topological residuals are included, without the histogram's display cutoff. The horizontal axis uses $\log(1+r)$ with labels in percentage points. Under both query orders, CFG's curve rises earlier, indicating that a larger share of its residuals is concentrated near zero.

C.6 ADDITIONAL EFFECT-REMOVAL RESULTS

Figs. 8 and 10 provide the extended plots corresponding to Fig. 4, including topological Census-Synth salary with an empty BN set and both query orders for the non-empty BN comparison.

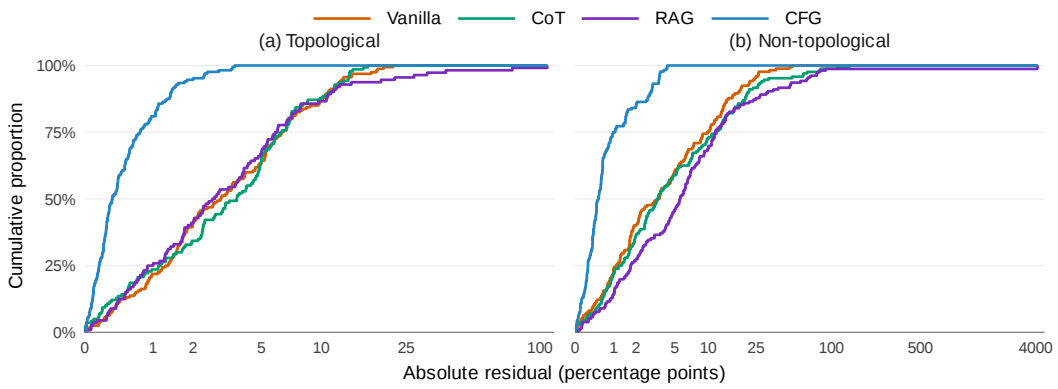


Figure 7: Empirical cumulative distributions of absolute recovery residuals.

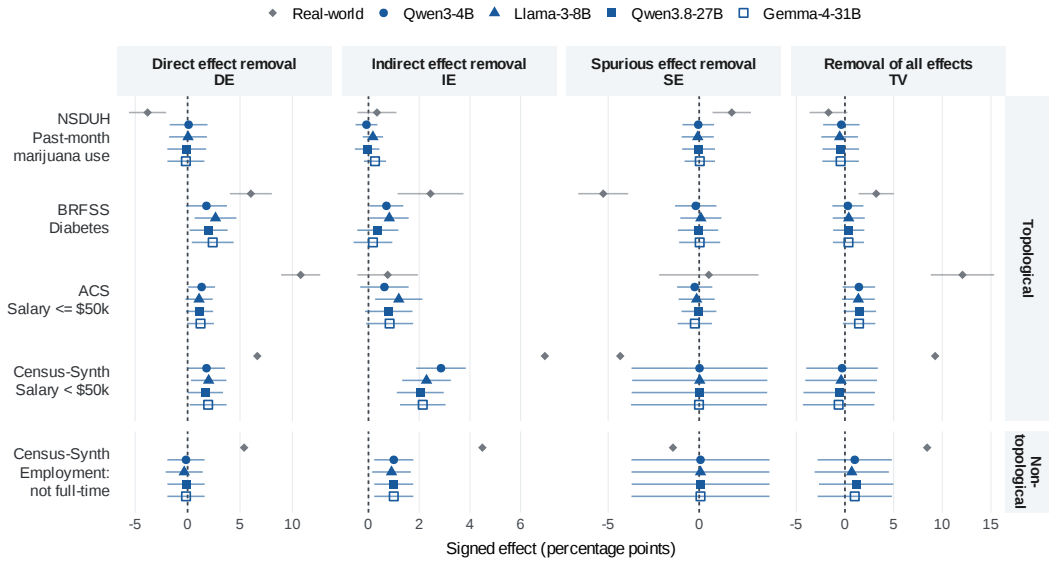


Figure 8: Extended effect-removal results with an empty BN set. The upper four rows show final outcomes under topological queries; the bottom row shows Census-Synth employment under non-topological queries. Columns report the selected effect, or TV when all effects are selected for removal. Intervals indicate 95% confidence intervals.

Fig. 9 complements the main-text empty-BN comparison with intermediate outcomes under topological queries. The tables report effect magnitudes before and after debiasing for topological queries (Tbl. 3), non-topological Census-Synth employment (Tbl. 4), and Census-Synth salary with $\mathbf{B} = \{\text{education}\}$ (Tbl. 5).

Reduction across targets and generators. Across topological intermediate and final outcomes, the selected effect decreases in magnitude in 169 of 176 comparisons. The remaining seven involve ACS events whose reference effects are already small, at most 0.76 percentage points in magnitude.

Debiasing intermediate attributes. Targeting DE reduces its mean absolute magnitude from 5.59 to 0.75 percentage points at intermediate outcomes; targeting IE reduces it from 3.33 to 1.04, and targeting SE from 0.76 to 0.09. In NSDUH, for example, the intermediate outcome is income below $50,000. The reference direct and education-mediated contributions both increase the probability of low income for Black or Hispanic individuals, and each is reduced under its corresponding removal setting (Fig. 9). This demonstrates that CFG can debias an intermediate attribute that also mediates effects on a later outcome.

Generator	Order	Method	Intermediate target				Final target			
			TV	DE	IE	SE	TV	DE	IE	SE
Qwen3-4B	Topo.	Vanilla	7.79	4.56	2.26	1.11	10.25	8.58	0.83	2.43
		CoT	9.64	5.51	4.04	0.81	7.62	9.51	0.73	2.48
		RAG	13.92	5.88	7.16	0.88	1.74	4.80	2.30	5.36
		CFG	**0.41**	**0.77**	**0.36**	**0.20**	**0.66**	**0.49**	**0.40**	**0.47**
	Non-topo.	Vanilla	19.90	12.38	7.85	4.99	30.68	14.59	10.04	17.58
		CoT	33.01	14.78	11.12	22.64	60.89	19.17	6.62	48.96
		RAG	13.86	11.16	8.74	3.20	6.61	6.17	9.60	9.17
		CFG	**1.62**	**1.76**	**0.86**	**0.20**	**1.66**	**2.07**	**0.55**	**0.48**
Llama-3-8B	Topo.	Vanilla	6.53	4.43	3.01	0.84	4.14	3.14	2.40	3.10
		CoT	8.30	4.16	3.93	1.53	8.92	7.87	2.84	3.07
		RAG	6.93	7.79	1.26	1.12	12.42	10.90	2.03	7.07
		CFG	**0.58**	**0.82**	**0.15**	**0.18**	**0.63**	**0.72**	**0.39**	**0.52**
	Non-topo.	Vanilla	8.48	6.39	4.26	1.77	10.27	7.96	4.48	1.94
		CoT	12.48	10.00	4.61	2.95	7.33	5.29	5.03	3.77
		RAG	11.61	9.04	5.85	3.98	10.79	5.01	8.36	5.22
		CFG	**1.68**	**1.66**	**0.85**	**0.25**	**1.66**	**2.18**	**0.66**	**0.45**
Qwen3.8-27B	Topo.	Vanilla	6.43	5.75	3.47	2.35	5.06	6.54	1.66	2.70
		CoT	6.00	5.74	2.98	1.71	2.19	4.86	1.63	4.14
		RAG	4.30	4.07	1.88	1.85	1.83	5.49	1.48	4.55
		CFG	**0.41**	**0.56**	**0.35**	**0.27**	**0.49**	**0.60**	**0.31**	**0.40**
	Non-topo.	Vanilla	4.04	2.03	4.11	1.41	7.95	5.55	1.54	1.64
		CoT	2.47	3.57	1.98	0.90	7.19	5.65	1.39	1.70
		RAG	4.08	2.38	2.84	0.74	6.31	2.40	4.36	1.08
		CFG	**1.58**	**1.68**	**0.88**	**0.25**	**1.80**	**2.11**	**0.48**	**0.53**
Gemma-4-31B	Topo.	Vanilla	12.02	6.50	3.19	4.90	8.45	5.95	3.27	2.65
		CoT	8.02	5.18	3.57	1.70	8.99	4.41	**0.20**	4.47
		RAG	41.83	16.56	3.74	25.53	8.96	6.27	2.20	5.89
		CFG	**0.40**	**0.61**	**0.41**	**0.24**	**0.66**	**0.78**	0.42	**0.52**
	Non-topo.	Vanilla	5.34	6.33	8.40	6.37	13.63	5.63	13.95	9.13
		CoT	4.77	7.78	10.73	2.83	16.67	5.78	16.51	4.62
		RAG	1076.38	16.83	10.68	1067.67	33.53	23.44	20.40	23.63
		CFG	**1.62**	**1.76**	**0.86**	**0.20**	**1.66**	**2.07**	**0.55**	**0.48**

Table 2: Mean absolute recovery error in percentage points by generator, query order, and target position for CFG and the prompting baselines. Dataset contributions are weighted equally after averaging ACS employer categories. Bold values indicate the smallest error within each comparison.

Removing all effects. Tbl. 3 shows that removing DE, IE, and SE together reduces mean absolute TV from 8.65 to 0.64 percentage points at intermediate outcomes. The corresponding DE, IE, and SE residuals are 0.69, 0.40, and 0.09 points. At final outcomes, TV falls from 6.56 to 0.54 points, with component residuals of 0.40, 0.20, and 0.07. The small aggregate disparity therefore accompanies small direct, indirect, and spurious contributions, rather than cancellation of large opposing effects.

Business-necessity settings and permitted paths. We evaluate salary below \$50,000 using the business-necessity set $\mathbf{B} = \{\text{education}\}$ under both query orders, with four effect-removal choices and four generators. Non-topological queries supply age, education, employment, and salary and request sex; debiasing precedes conditioning on these supplied values.

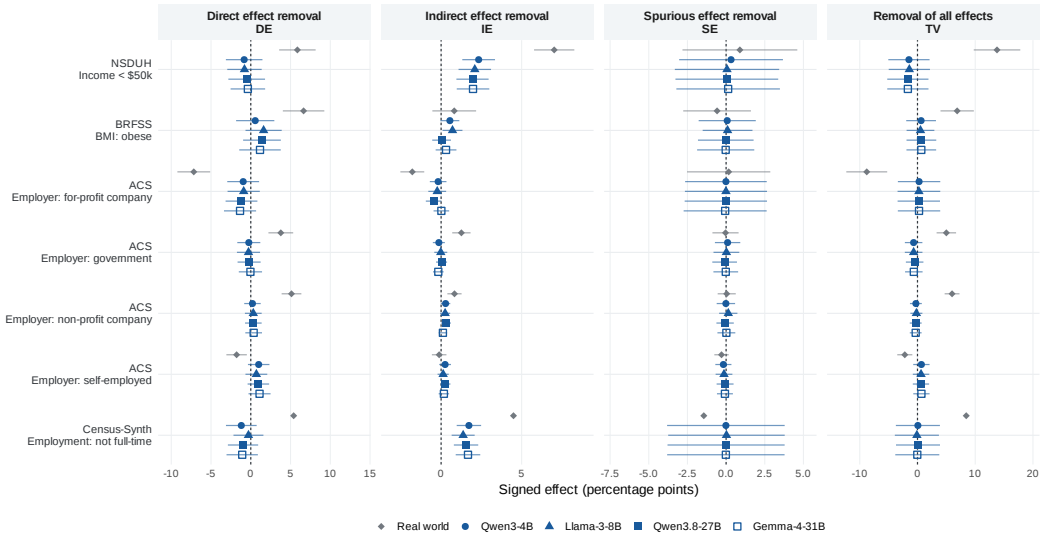


Figure 9: Causal effect removal for intermediate outcomes under topological queries. Each panel compares the effect selected for removal after debiasing with its original reference value; zero corresponds to complete removal.

Removal setting	Effect	Intermediate target		Final target	
		Reference	CFG	Reference	CFG
Direct effect	DE	5.59	0.75	6.82	1.32
Indirect effect	IE	3.33	1.04	2.62	0.97
Spurious effect	SE	0.76	0.09	2.98	0.08
All effects	TV	8.65	0.64	6.56	0.54
	DE	5.59	0.69	6.82	0.40
	IE	3.33	0.40	2.62	0.20
	SE	0.76	0.09	2.98	0.07

Table 3: Mean absolute effects in percentage points for the reference and CFG-debiased distributions under topological queries. When all effects are selected for removal, we report both TV and its components.

The BN set permits sex $\rightarrow$ education $\rightarrow$ salary and sex $\rightarrow$ education $\rightarrow$ employment $\rightarrow$ salary. Paths bypassing education remain subject to removal when their effect family is selected. These exceptions concern directed pathways; the spurious component remains subject to removal when selected.

Fig. 10 shows results for both query orders. The plotted quantities are the full DE, IE, SE, and TV. A nonzero IE after indirect-effect removal, or TV after removal of all non-BN effects, can therefore include contributions permitted by BN. These aggregates do not distinguish permitted contributions from residual non-BN effects. Zero remains the target for a selected effect with no BN exception; nonzero estimates cannot then be attributed to an exemption. Tbl. 5 reports the numerical summaries.

Remaining disparity under BN. When indirect effects are selected for removal, the mean absolute indirect effect remains close to its 6.96-point ground-truth magnitude: 7.26 points for topological queries and 6.33 for non-topological queries (Tbl. 5). This is consistent with the education-mediated pathways exempted by BN. When all non-BN effects are selected for removal, mean absolute TV is 4.15 points for topological queries and 4.06 for non-topological queries, compared with the 9.27-point reference magnitude.

Removal setting	Effect	Reference	CFG
Direct effect	DE	5.39	0.19
Indirect effect	IE	4.49	0.98
Spurious effect	SE	1.44	0.07
All effects	TV	8.44	0.97
All effects	DE	5.39	0.26
All effects	IE	4.49	0.65
All effects	SE	1.44	0.05

Table 4: Mean absolute Census-Synth employment effects in percentage points for the reference and CFG-debiased distributions under non-topological queries, averaged across four generators.

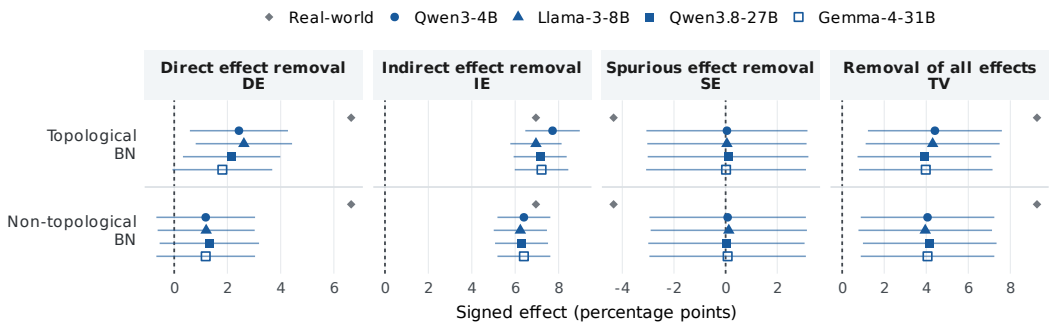


Figure 10: Extended effect-removal results for Census-Synth salary with $\mathbf{B} = \{\text{education}\}$, for both topological and non-topological queries. Intervals indicate 95% confidence intervals.

D CONCEPT AGENT EVALUATION

The concept agent implements the query-analysis step of CFG in Sec. 2.2. Its purpose is to recover the variables represented in a generation request before population grounding and causal-diagram acquisition. We evaluate this component independently of the remaining steps of CFG so that downstream assumptions do not obscure extraction errors.

Task. Given a prompt requesting a person-level generation, the agent returns a set of concept records. Each record contains (i) a normalized variable name, (ii) the value supplied in the prompt, if any, (iii) a status indicating whether the value is *specified* or merely *requested*, and (iv) an exact span of the prompt supporting the extraction. Requested concepts have a null value and become generation targets, whereas specified concepts provide conditioning information. The agent excludes stylistic and formatting instructions, as well as attributes that the prompt only prohibits the generated text from mentioning. The input prompt is presented as data rather than as an instruction to follow, and the agent does not assign the extracted variables to causal roles or infer edges between them.

Benchmark. We construct a diagnostic benchmark of 100 prompts containing 528 gold concept records. The benchmark comprises 34 in-domain prompts modeled on the health, substance-use,

Removal setting	Effect	Reference	Topological	Non-topological
Direct effect	DE	6.65	2.25	1.22
Indirect effect	IE	6.96	7.26	6.33
Spurious effect	SE	4.34	0.05	0.08
All effects	TV	9.27	4.15	4.06

Table 5: Mean absolute Census-Synth salary effects in percentage points for the reference and CFG-debiased distributions with $\mathbf{B} = \{\text{education}\}$, averaged across four generators.

Model	Name F_1	Record F_1	In-domain	Out-of-domain	Challenge
Qwen3-4B	95.5	90.8	93.1	87.7	92.3
Llama 3-8B	91.6	85.7	90.4	83.0	78.9
Qwen3.8-27B	97.9	94.3	96.2	91.1	97.5
Gemma 4-31B	95.9	91.4	94.1	86.8	95.7

Table 6: Few-shot performance of the concept agent for CFG on the 100-prompt benchmark. All entries are percentages. Name and record scores are micro-averaged after semantic normalization; the final three columns report record F_1 within each benchmark split.

census, and employment tasks considered in our experiments; 33 out-of-domain prompts spanning, among others, finance, criminal justice, transportation, housing, and climate risk; and 33 challenge prompts covering negation, prohibited attributes, repeated requests, prompts with no eligible concepts, and prompt-injection language. Here, *out-of-domain* refers to domains outside the applications used elsewhere in this paper, rather than to a statistically sampled population of user prompts. Every gold evidence span occurs verbatim in its prompt. Five additional labeled prompts are reserved as fixed demonstrations and do not occur in the evaluation set.

Models and evaluation protocol. We evaluate the same four generators used in the main experiments, each receiving the same five demonstrations, with no zero-shot comparison. We use deterministic decoding ($T = 0$ and top-$p = 1$) and constrain the response to a JSON schema containing the four fields described above. Each model is evaluated on all 100 prompts using a maximum context length of 4,096 tokens and at most 512 generated tokens.

Metrics. We report semantically normalized matching throughout. For evaluation, we apply the same dictionary of unambiguous lexical aliases to every model, such as `sex`↦`gender` and `age_group`↦`age` and `annual_salary`↦`salary`, and apply simple normalization to values. Name precision, recall, and F_1 measure recovery of the resulting concept set. Record-level precision, recall, and F_1 additionally require the specified/requested status and normalized value to match. We also audit schema validity and whether every predicted evidence span is present verbatim in the input prompt.

Results. Qwen3.8-27B performs best across all reported metrics, achieving 97.9% name F_1 and 94.3% record F_1. Its record F_1 exceeds 91% in each benchmark split, including the out-of-domain prompts.

Discussion. This performance supports using Qwen3.8-27B for concept detection in CFG. Since the prompts are curated and the alias dictionary was refined after inspecting predictions, the results are a component diagnostic rather than an independent estimate of generalization to arbitrary requests.

Estimator	Log loss ↓	Brier score ↓
DRF	$\mathbf{1.1016 \pm 0.0150}$	$\mathbf{0.5487 \pm 0.0073}$
XGBoost	1.1169 ± 0.0036	0.5519 ± 0.0070

Table 7: Survey-weighted held-out goodness of fit on NSDUH for the population estimators available in CFG. Entries are means $\pm$ standard deviations across five stratified 80/20 splits.

E POPULATION FITTING COMPARISON

We compare the XGBoost and distributional random forest (DRF) estimators described in step (S2) of Sec. 2.2. XGBoost fits categorical conditional probabilities; DRF uses a maximum mean discrepancy splitting criterion to detect changes in the response distribution and assign query-specific weights to observed records (Cevid et al., 2022).

Evaluation protocol. We compare the fitted conditional distribution of race given age, sex, education, income, and monthly marijuana use on the full preprocessed NSDUH source of 45,133 records. For each of five seeds, we hold out 20% of records, stratifying by race and monthly marijuana use and keeping identical records in the same partition. The resulting splits each contain 36,105 training and 9,028 held-out records. Both estimators use the same training and test partition within each split; XGBoost selects its number of boosting rounds by five-fold cross-validation on the training partition only, and DRF uses 500 trees. Analogously to the negative log predictive density used by Cevid et al. (2022), our primary metric is held-out categorical log loss. As a complementary measure, we report the multiclass Brier score (Gneiting & Raftery, 2007): the sum of squared differences between the predicted class-probability vector and the one-hot observed outcome. We average both scores using the survey weights; lower values indicate better fit. A probability floor of 10^{-6} is applied when computing log loss.

Results. DRF achieves slightly lower mean log loss and Brier score than XGBoost under this evaluation protocol. Together with its joint-record sampling mechanism, this supports our choice of DRF as the primary population-fitting option.

F LLM CAUSAL DIAGRAM RETRIEVAL

Objective and input. We evaluate whether an LLM can recover a plausible standard fairness model (SFM) from the protected attribute X selected by the user and the remaining variable names extracted from a query. This tests the agent-based causal-diagram acquisition step (S3) in Sec. 2.2. Here, *retrieval* means proposing a diagram using background knowledge; the model does not select an entry from a supplied graph repository. We isolate this component by supplying the variable names directly, without concept extraction errors. The input specifies X, whose group disparities are of interest, but neither the baseline set $\mathbf{Z}$ nor the ordering of $\mathbf{Y} = (Y_1, \ldots, Y_k)$. In particular, the final outcome Y_k is not given: the model must infer the entire order. No observations, case titles, reference assignments, or target-case edges are supplied.

SFM construction from a topological order. The agent recovers the baseline set $\mathbf{Z}$ and a topological order $\mathbf{Y} = (Y_1, \ldots, Y_k)$. Every remaining variable must appear exactly once in either $\mathbf{Z}$ or $\mathbf{Y}$, with $k \geq 1$ and an optionally empty $\mathbf{Z}$. Baseline variables in $\mathbf{Z}$ are not causally affected by X or $\mathbf{Y}$. We instantiate the SFM in Fig. 1(b): X and each Z_j have directed edges to every Y_i, we add $Y_i \to Y_j$ for every $i < j$, and a bidirected edge connects X with each Z_j. No other edges are allowed. Thus, the roles and order determine the entire graph, including direct connections between nonadjacent outcomes. This template separates role and ordering errors from unconstrained edge selection; the general formulation also permits subsets of the forward edges. The diagram is constructed before debiasing, so potentially discriminatory edges are retained.

Reference graphs. We use the NSDUH and ACS diagrams from the paper and ten hand-curated hypothetical fairness examples. The synthetic examples were developed with LLM-assisted drafts, manually revised, and reviewed before evaluation; they specify baseline roles and outcome orders, rather than datasets generated from simulated observations. Tbl. 8 lists all twelve reference assignments. Under the fixed order-based template above, each row specifies the entire graph. The examples cover several protected attributes and include an empty-$\mathbf{Z}$ case. Their references are reviewed causal assumptions, rather than uniquely identifiable ground truth from variable names. BRFSS is reserved as one fixed worked input/output example in every prompt and is excluded from evaluation. Census-Synth is not part of this twelve-case retrieval benchmark.

Models and protocol. We evaluate the four open-weight models used in the main experiments: Qwen3-4B, Llama 3-8B, Qwen3.8-27B, and Gemma 4-31B. They receive identical instructions requesting baseline roles and the complete outcome order under the SFM template, together with the same BRFSS demonstration, with greedy decoding, thinking disabled, an 8,192-token context limit, and at most 4,096 output tokens. We additionally evaluate ChatGPT 6 Astra as a frontier-model comparison available to us at the time of access (22 September 2026), through the agent interface with requested model identifier `gpt-6-astra` and high reasoning effort. Each Astra presentation uses a fresh agent without conversation history. Public web research is permitted, but none of the sixty agents reported using it. All reported reading only their assigned prompt.

For each graph, we use five distinct presentations: the saved variable order and four permutations generated with seed 42. These presentations are shared across models, yielding 60 responses per model and 300 responses overall, while retaining twelve distinct graphs. Every response is generated in a fresh context and must return $\mathbf{Z}$, the complete topological order of $\mathbf{Y}$, and both edge lists as JSON. In the saved prompts and responses, adjacent outcome edges encode this order. For the causal diagrams used here, we expand that encoding by adding every forward outcome edge; this deterministic construction requires no further LLM inference. We use no constrained decoding, corrective retries, or graph correction. The reference graphs, task prompts, and scoring rules are fixed across all five models, and Astra answers are frozen before scoring.

Metrics. We report four complementary scores. *Role macro-F_1* averages the F_1 scores for membership in $\mathbf{Z}$ and $\mathbf{Y}$, excluding the supplied X; a role absent from both reference and prediction receives $F_1 = 1$. *Order agreement* is the fraction of pairs in the reference $\mathbf{Y}$ sequence that appear in the correct relative order. A pair is incorrect if either member is assigned to $\mathbf{Z}$. This gives partial credit for nearly correct orders without treating role errors as harmless. *Exact graph accuracy* requires equality of all directed and bidirected edges in the graphs constructed from the predicted and reference assignments. For valid outputs, this is equivalent to recovering the baseline set and

Case	X	$\mathbf{Z}$	Reference $\mathbf{Y}$ order
NSDUH	race	sex, age	education $\prec$ income $\prec$ monthly marijuana use
ACS	sex	age group, race, economic region	education $\prec$ weekly hours worked $\prec$ employer type $\prec$ annual salary
Hiring	gender	age	education $\prec$ relevant experience $\prec$ interview score $\prec$ hiring decision
Promotion	race	age, sex	employer training access $\prec$ job performance $\prec$ promotion decision
Salary offer	sex	age	relevant experience $\prec$ initial salary offer $\prec$ negotiation response $\prec$ final salary offer
Admissions	first generation	age, sex	household resources $\prec$ school preparation $\prec$ entrance exam score $\prec$ admission decision
Scholarship	disability	age	academic merit $\prec$ scholarship awareness $\prec$ application completion $\prec$ scholarship award
Lending	ethnicity	age	income $\prec$ repayment history $\prec$ credit score $\prec$ loan approval
Rental housing	nationality	age	income $\prec$ viewing invitation $\prec$ reference check $\prec$ lease offer
Insurance	age	sex, race	driving experience $\prec$ accident history $\prec$ insurance premium
Benefits	language	age, sex	digital literacy $\prec$ application completion $\prec$ eligibility assessment $\prec$ benefit receipt
Moderation	language	$\varnothing$	message content $\prec$ writing style $\prec$ automated toxicity score $\prec$ content removal

Table 8: The twelve graphs used to evaluate causal-diagram acquisition for CFG. The first two are paper references; the remaining ten are hand-curated examples. Spaces replace underscores in variable identifiers for readability. The symbol $\prec$ denotes precedence in the topological order. Adding every forward outcome edge to the fixed SFM template fully specifies each reference graph. The target input supplies only X and the unordered remaining names; the $\mathbf{Z}$ memberships and $\mathbf{Y}$ orders shown here are withheld.

Model	Role macro-F_1	Order agreement	Exact graph	Valid output
Qwen3-4B	78.1	56.4	23.3	78.3
Llama 3-8B	73.2	45.8	11.7	78.3
Qwen3.8-27B	95.3	83.3	58.3	**100.0**
Gemma 4-31B	94.5	87.5	60.0	**100.0**
ChatGPT 6 Astra (high)	**97.1**	**95.8**	**85.0**	**100.0**

Table 9: SFM diagram retrieval for CFG on twelve graphs, with five variable-order presentations per graph and model. All entries are percentages, and higher is better.

the entire topological order exactly. *Valid-output rate* retains the original response-schema check: a complete, disjoint assignment and all and only the edges implied by its adjacent-edge encoding and the fixed SFM connections. We tolerate prose or Markdown surrounding a single extractable JSON object, without changing its contents. Structurally invalid outputs receive zero on all four metrics and remain in the denominator. Completing the forward edges preserves the role and order scores and exact-match decisions for valid outputs; the original validity decisions are retained. All four metrics therefore evaluate the same recovered roles and order under this graph construction. Scores are averaged over presentations within each graph and then equally over graphs.

Results. Tbl. 9 shows that Astra achieves the highest role and order scores and matches 51 of 60 presentations exactly (85.0%), compared with 36 of 60 (60.0%) for Gemma, the strongest open-weight model by exact accuracy. On the ten synthetic graphs alone, Astra achieves 96.6% role macro-F_1, 96.7% order agreement, and 92.0% exact accuracy, compared with 62.0% exact accuracy for Gemma. On the two paper graphs, Astra achieves 100.0% role macro-F_1, 91.7% order agree-

ment, and 50.0% exact accuracy. All Astra outputs are valid even under strict JSON parsing. It exactly matches every presentation of ten graphs; its nine non-exact responses occur only in ACS and moderation. The distinction between the partial-credit and exact metrics is consequential: a single reversed pair can invalidate an exact match while preserving most of the proposed structure.

ACS: working hours and employer type. For all five ACS presentations, Astra recovers the reference baseline set but proposes the order education $\prec$ employer type $\prec$ weekly hours worked $\prec$ annual salary. The reference instead uses education $\prec$ weekly hours worked $\prec$ employer type $\prec$ annual salary. Each order induces all forward outcome edges. The reference can be interpreted as working-time commitments or constraints shaping the type of employer a person enters, which then affects salary. Astra's ordering instead represents employer type determining working-hour requirements, which affect annual earnings. Both are plausible simplified mechanisms under different temporal interpretations of the variables. Their names alone do not specify whether hours precede employer selection or are determined by the current employer. Astra reverses one of the six reference $\mathbf{Y}$ pairs, yielding 83.3% order agreement for ACS but no exact matches. These are disagreements with a fixed reference, rather than evidence that the alternative ordering is causally impossible.

Moderation: message content as a mediator or a baseline. The reference sets $\mathbf{Z} = \varnothing$ and orders message content $\prec$ writing style $\prec$ automated toxicity score $\prec$ content removal. Here, language denotes the author's primary language. The reference assumes that language and its associated communication contexts can affect which content is expressed; content then influences its linguistic realization and the moderation score. The graph includes direct edges from message content to the toxicity score and content removal, as well as paths through writing style.

In three presentations, Astra instead assigns message content to $\mathbf{Z}$ and keeps the order writing style $\prec$ automated toxicity score $\prec$ content removal for $\mathbf{Y}$. This is plausible when content means an intended message or externally specified topic that precedes language realization: changing the language need not change the intended meaning, while content can directly affect style, the score, and removal. The implied language $\leftrightarrow$ message content edge allows shared background causes of language and topic selection. Thus, the reference and this alternative correspond to different meanings of "content": expressed content shaped by language versus a prior semantic input. Both are defensible assumptions for suitably defined tasks, and the names supplied to the model do not resolve that distinction. A fourth response keeps $\mathbf{Z}$ empty but places writing style before message content, which can represent stylistic choices shaping the eventual expression; this is again more plausible for expressed content than for fixed intended meaning. The fifth response matches the reference. We retain all four disagreements as failures under the prespecified exact-match metric.

Implications for the diagram agent. This small, hand-curated benchmark supports using Astra to propose an SFM for user or expert review. The ACS and moderation cases show why the agent should expose its roles and order: ambiguous variable definitions can yield different plausible diagrams with different fairness implications. As in step (S3), the chosen diagram remains a structural assumption for debiasing.

G ILLUSTRATIVE CONVERSATION WITH CFG

The following example illustrates a possible conversation between a user and an LLM using CFG. It uses the Census-Synth variables and causal diagram from App. B. The dialogue illustrates the user's choices; the final story is taken from a saved generation with an empty BN set and all effects on salary selected for removal. Effect removal concerns the distribution from which stories are generated; a single story illustrates the output format.

User: Prompt

Generate a short, realistic story about an adult in the Census-Synth population. Mention their sex, age, education, employment status, and annual salary. Use sex as the protected attribute.

CFG: Example Output

I can use Census-Synth as the reference population. Its causal diagram can be written as the following relationships:

```
sex -> {education, employment, salary}
age -> {education, employment, salary}
education -> {employment, salary}
employment -> salary
sex <-> age
```

Here, `->` denotes a direct causal influence on each variable listed on the right, and `<->` denotes unobserved common causes of sex and age. For salary, education and employment can transmit indirect effects of sex, while the sex–age association can contribute a spurious effect.
Please confirm whether you want to use this diagram and focus the fairness analysis on salary. Would you like to remove any effects before I generate the story? If you choose no removal, I will sample the attributes from the reference population distribution, aiming to reproduce its relationships and disparities. For Census-Synth, that reference is the synthetic population.

User: Prompt

Yes, use this diagram and focus on salary. I want to remove effects before you generate the story.

CFG: Example Output

First, are there any pathways you consider justified for this task and want to retain? These are called **business-necessity (BN) exceptions**. You can select a BN set from education, employment, and salary, or choose none. A directed pathway from sex to salary is exempt from removal if it contains at least one variable in your BN set. For example, selecting education exempts pathways through education. An empty BN set provides no exceptions. Spurious-effect removal is specified separately from these exceptions.
Which BN set would you like? Also choose which effects on salary to remove:

- **Direct effect:** the effect of sex on salary that bypasses education and employment.
- **Indirect effect:** effects transmitted through education or employment.
- **Spurious effect:** the contribution associated with the sex–age association.
- **All effects:** direct, indirect, and spurious.

User: Prompt

Use an empty BN set, {}: no exceptions. Remove all effects on salary.

CFG: Example Output

Martha is a female who enjoys spending her afternoons gardening in her small suburban backyard. Now in the 60–74 years age bracket, she stays active within her community. Her highest educational attainment is some college or associate degree, which she earned decades ago. She is currently part-time employed as a library assistant, a role that brings her great joy. Because of her limited hours, her annual salary is less than $25k.

Method	TV	DE	SE	IE
Vanilla	9.18	7.25	4.92	3.87
CoT	9.96	6.85	5.48	3.82
RAG	9.20	5.07	3.43	9.19
CFG	**1.59**	**1.85**	**1.19**	**0.73**

Table 10: Mean absolute errors in reference-effect matching with supplied $(X, \mathbf{Z})$, averaged across generators, in percentage points.

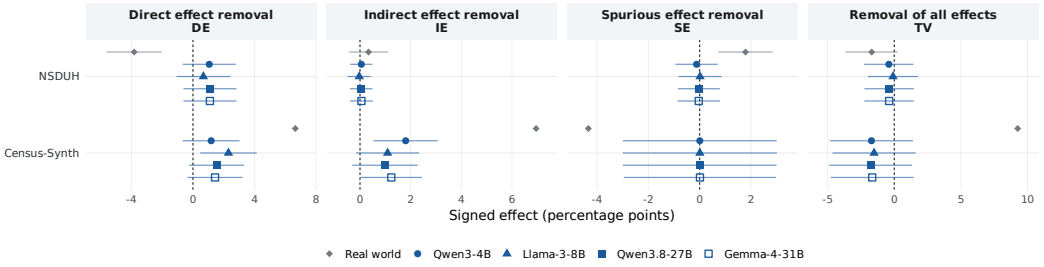


Figure 11: Effect removal for final outcomes with supplied $(X, \mathbf{Z})$.

H Generation Conditioned on Protected Attributes and Covariates

We evaluate the setting in which the protected attribute X and covariates $\mathbf{Z}$ are supplied to the generator, which produces all outcome attributes $\mathbf{Y}$. In the query notation of Sec. 2.2, this corresponds to $\mathbf{C} = (X, \mathbf{Z})$ and $\mathbf{T} = \mathbf{Y}$. The main-text topological experiments instead generate all attributes, with $\mathbf{C} = \varnothing$ and $\mathbf{T} = (X, \mathbf{Z}, \mathbf{Y})$. We evaluate this setting on NSDUH and Census-Synth with the same four generators as in the main experiments. All queries are topological, and we evaluate the final outcomes: past-month marijuana use for NSDUH and salary below $50,000 for Census-Synth.

H.1 Matching Reference Causal Effects

For recovery, we sample $(X, \mathbf{Z})$ jointly from rows of the original reference data and supply these values to Vanilla, CoT, RAG, and CFG. We compare effects in the generated outcomes with their reference values using the annotation and estimation protocol in Apps. C.1 and C.2.

Tbl. 10 reports mean absolute residuals, first averaging equally over available datasets within each generator and then equally over the four generators. The NSDUH Gemma/RAG run has a constant annotated final outcome, causing the effect estimator to fail; its model-level mean therefore uses Census-Synth only. In this aggregate comparison, CFG again achieves lower recovery error than every baseline for all four effects: 1.59 percentage points for TV, 1.85 for DE, 1.19 for SE, and 0.73 for IE.

H.2 Removing Direct, Indirect, and Spurious Effects

We evaluate CFG with all four generators, selecting direct, indirect, or spurious effects for removal individually, or all three together, with an empty BN set. The supplied $(X, \mathbf{Z})$ values are sampled from the adapted distribution for each removal setting. When spurious-effect removal is selected, we supply $(X, (\mathbf{Z})_{x_0})$, where $(\mathbf{Z})_{x_0}$ denotes covariates adjusted to the common reference-group distribution. Supplying the original covariates would retain the X–$\mathbf{Z}$ association. Direct- or indirect-effect removal alone leaves the supplied context distribution unchanged. In each case, we generate $\mathbf{Y}$ from the adapted conditional distribution given the supplied values. Fig. 11 shows that the targeted effects are reduced toward zero in both datasets across all four generators, including the spurious component when it is selected for removal.