GUMBEL COUNTERFACTUAL GENERATION FROM LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding and manipulating the causal generation mechanisms in language models is essential for controlling their behavior. Previous work has primarily relied on techniques such as representation surgery—e.g., model ablations or manipulation of linear subspaces tied to specific concepts—to *intervene* on these models. To understand the impact of interventions precisely, it is useful to examine *counterfactuals*—e.g., how a given sentence would have appeared had it been generated by the model following a specific intervention. We highlight that counterfactual reasoning is conceptually distinct from interventions, as articulated in Pearl's causal hierarchy. Based on this observation, we propose a framework for generating true string counterfactuals by reformulating language models as a structural equation model using the Gumbel-max trick, which we called Gumbel counterfactual generation. This reformulation allows us to model the joint distribution over original strings and their counterfactuals resulting from the same instantiation of the sampling noise. We develop an algorithm based on hindsight Gumbel sampling that allows us to infer the latent noise variables and generate counterfactuals of observed strings. Our experiments demonstrate that the approach produces meaningful counterfactuals while at the same time showing that commonly used intervention techniques have considerable undesired side effects.

1 INTRODUCTION

The study of language model (LM) interpretability often borrows terminology from Pearl's causal calculus (Pearl, 1989), e.g., researchers often talk of *intervening* a model's parameters and *counterfactually* generating strings. Pearl's framework distinguishes between three levels of causal reasoning (Shpitser & Pearl, 2008). Association, the first level, pertains to statistical correlations, i.e., observing patterns observed in data without interacting with the world. Intervention, the second level, pertains to actively changing variables in the world and observing their effects at a macro level. Counterfactuality, the third level, pertains to imagining what could have happened if past events had unfolded differently. However, LM literature often uses these three causal terms casually and at times imprecisely—particularly when it comes to counterfactuality, which remains challenging to rigorously define (Feder et al., 2022; Mueller, 2024; Mueller et al., 2024). In this paper, we apply a well-defined notion of counterfactuality in LMs using the framework of structural equation modeling.

Efforts to exert control over LMs have led to substantial research on targeted *interventions* in the models. One such technique is *representation surgery*, which involves modifying an LM's architecture to manipulate its internal representation space (Lakretz et al., 2019; Vig et al., 2020; Feder et al., 2021; Ravfogel et al., 2021b; Elhage et al., 2021; Elazar et al., 2021; Nanda, 2023; Syed et al., 2023; Kramár et al., 2024; Avitan et al., 2024). The *linear subspace hypothesis* (Bolukbasi et al., 2016; Vargas & Cotterell, 2020; Ravfogel et al., 2022) posits that human-interpretable *concepts*, such as gender or grammatical number, are encoded within specific linear subspaces of the LM's representation space. This makes it possible to perform precise interventions on these high-level concepts, such as removing the concept's information by projecting the representations onto the complement of the concept subspace (Ravfogel et al., 2020; 2021a; 2022; 2023; Guerner et al., 2024; Scalena et al., 2024; Singh et al., 2024). These interventions modify the model and allow researchers to examine its behavior after the change. However, while interventions can induce a *change* in the model, they cannot answer counterfactual questions, e.g., what would a given string look like if it had been generated by the model after the intervention?

Counterfactual analysis, as defined by Pearl, is challenging because it requires describing the system of interest through a causal model that enables counterfactual reasoning. In this paper, we address this challenge for LMs by turning to structural equation models (SEMs; Wright, 1921; Haavelmo, 1944). SEMs break down a probabilistic model into *exogenous* variables, which account for latent randomness, and *endogenous* variables, which are deterministic once the exogenous ones are fixed. To frame LMs as SEMs, we turn to the Gumbel-max trick (Gumbel, 1954), which separates the deterministic computation of next-symbol logits from the sampling process. This approach has been explored in reinforcement learning (Oberst & Sontag, 2019), but has not yet been considered in the language modeling domain. Indeed, we discuss later, the infinite outcome space of LMs requires special attention that is not needed in finite domains. Additionally, we highlight that the SEM derived from the Gumbel parameterization of an LM is not unique—while some parameterizations are unnatural (Oberst & Sontag, 2019, §3.1), there are several natural choices to model an LM causally. We argue for Gumbel counterfactual generation by appealing to the classic Thurstone discriminal process (Thurstone, 1927),[1] and because its simplicity paves the way for other choices beyond Gumbel noise (Lorberbom et al., 2021; Haugh & Singal, 2023).

Our formulation allows for generating counterfactual strings by sampling from conditional noise distributions, enabling precise analysis of string-level effects from interventions in models like GPT2-XL (Radford et al., 2018) and LLaMA3-8b (Touvron et al., 2023). Despite targeting specific behaviors through interventions such as linear steering (Li et al., 2024; Singh et al., 2024), knowledge editing (Meng et al., 2023), and instruction tuning (Wei et al., 2022), results reveal unintended side effects, e.g., gender-based interventions unexpectedly altering unrelated completions. These findings challenge the goal of achieving *minimal change* and show that even localized parameter modifications can have broader, undesired impacts.

2 Language models as structural equation models

Let Σ be an **alphabet**—a finite, non-empty set of symbols. A **language model** (LM) is a probability distribution over Σ^*, the set of all strings formed from symbols in Σ. A **language encoder** is a function $\boldsymbol{h}_{\boldsymbol{\theta}} \colon \Sigma^* \to \mathbb{R}^d$ parameterized by $\boldsymbol{\theta}$ that maps strings to d-dimensional vectors (Chan et al., 2024). Representational surgery is performed by intervening on $\boldsymbol{h}_{\boldsymbol{\theta}}$. Popular architectures for implementing language encoders include Transformers (Vaswani et al., 2017) and RNNs (Elman, 1990). Language encoders are particularly valuable because under mild conditions (Du et al., 2023, Thm. 4.7), they ensure the model defines a distribution over strings—thus forming an LM—as

$$p(\boldsymbol{w}) = p(w_1 \cdots w_T) \tag{1a}$$

$$= p(\textsc{eos} \mid \boldsymbol{w}) \prod_{t=1}^{T} p(w_t \mid \boldsymbol{w}_{<t}) \tag{1b}$$

$$= \mathrm{softmax}(\mathbf{E}\, \boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{w}) + \mathbf{b})_{\textsc{eos}} \prod_{t=1}^{T} \mathrm{softmax}(\mathbf{E}\, \boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{w}_{<t}) + \mathbf{b})_{w_t}. \tag{1c}$$

Here, $\mathbf{E} \in \mathbb{R}^{|\overline{\Sigma}| \times d}$ and $\mathbf{b} \in \mathbb{R}^{|\overline{\Sigma}|}$ is a bias term. We assume that $\textsc{eos} \notin \Sigma$ and define $\overline{\Sigma} \stackrel{\text{def}}{=} \Sigma \cup \{\textsc{eos}\}$. The quantity $\ell(\overline{w}; \boldsymbol{w}_{<t}) \stackrel{\text{def}}{=} (\mathbf{E}\, \boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{w}_{<t}) + \mathbf{b})_{\overline{w}}$ is often termed the **logit** of $\overline{w}$. We refer to LMs of the form Eq. (1c) **representational LMs**.

2.1 Structural equation modeling

We briefly review SEMs, which provide a framework for discussing causal manipulations of a generation process and allow us to define the intuitive notion of a counterfactual precisely. See Pearl (2009), Pearl et al. (2016), and Peters et al. (2017) for a more in-depth treatment. We offer a definition of an *acyclic* SEM because our focus is on language modeling, which operates on countably infinite random variables (RVs) with no cyclic dependencies. To apply acyclic SEMs to language modeling,

[1] Assuming the Thurstonian model of decision-making, equivalent to Luce's axiom of choice (Luce, 1959), makes the causal process behind an LM *identifiable* and is identical to having been generated through Gumbel noise. However, our appeal to Thurstone is most concretely an appeal to precedent, aligning with its established role in modeling decision-making and preference aggregation in previous work.

we *expose* the acyclicity by specifying a partial order $\prec$ that characterizes the acyclic relationships. This establishes a common interface for both the finite and infinite-variable models.

Definition 2.1 (Acyclic SEM)**.** *An **acyclic SEM** (ASEM) is a five-tuple $\mathcal{M} = (\mathbf{V}, \mathbf{U}, \mathcal{P}, \mathcal{F}, \prec)$, where $\mathbf{V} = \{\mathrm{V}_1, \ldots, \mathrm{V}_N\}$ is a finite set of **endogenous** RVs, $\mathbf{U} = \{\mathrm{U}_1, \ldots, \mathrm{U}_N\}$ is a finite set of **exogenous** RVs, $\mathcal{P} = \{\mathbb{P}_1, \ldots, \mathbb{P}_N\}$ is a set of probability distributions where $\mathbb{P}_n$ is a distribution over $\mathrm{U}_n \in \mathbf{U}$, $\mathcal{F} = \{f_1, \ldots, f_N\}$ is a finite set of **structural equations**, and $\prec$ is a **partial order** on $\mathbf{V}$. We require that the endogenous RVs satisfy the structural equations, i.e., we require that*

$$\mathrm{V}_n = f_n(\mathrm{PA}(\mathrm{V}_n), \mathrm{U}_n), \tag{2}$$

for $n \in \{1, \ldots, N\}$, where $\mathrm{PA}(\mathrm{V}_n)$, the subset of $\mathbf{V}$ in the domain of f_n, respects $\prec$, i.e., we require that $\mathrm{PA}(\mathrm{V}_n) \subseteq \{\mathrm{V} \in \mathbf{V} \mid \mathrm{V} \prec \mathrm{V}_n\}$.[2,3]

$\mathcal{P}$ induces a joint probability distribution over $\mathbf{U}$; together with $\mathcal{F}$, it also induces a unique distribution over $\mathbf{V} \cup \mathbf{U}$ due to the acyclicity. Given an outcome $\mathbf{U} = \mathbf{u}$, a **solution** of an ASEM is any assignment of the variables $\mathbf{V} = \mathbf{v}$ that satisfies the equations in $\mathcal{F}$ putting $\mathbf{U} = \mathbf{u}$. In ASEMs, any outcome $\mathbf{U} = \mathbf{u}$ induces a unique solution (Peters et al., 2017); in the cyclic case, multiple valid assignments can exist (Peters et al., 2017, Problem 3.8).

An ASEM can also be viewed as a directed graph over vertices $\mathbf{U} \cup \mathbf{V}$ with a directed edge from every $\mathrm{X} \in \mathrm{PA}(\mathrm{V}_n) \cup \{\mathrm{U}_n\}$ to V_n. In it, the exogenous variables always appear without parents. This presentation allows one to visualize ASEMs compactly and, also, reason better about their properties (Pearl, 2009).[4] For example, interventions are particularly convenient to think of as modifications of the directed graph—this indicatively changes the causal model represented by the graph. Analogously, in the context of ASEMs, interventions modify the structural equations $\mathcal{F}$.

Definition 2.2 (Intervention)**.** *Let $\mathcal{M} = (\mathbf{V}, \mathbf{U}, \mathcal{P}, \mathcal{F}, \prec)$ be an ASEM. An **intervention** $\mathrm{do}(\mathrm{V}_n = \widetilde{f}(\widetilde{\mathrm{PA}}(\mathrm{V}_n), \widetilde{\mathrm{U}}_n))$ replaces the equation $\mathrm{V}_n = f_n(\mathrm{PA}(\mathrm{V}_n), \mathrm{U}_n)$ in $\mathcal{F}$ with $\mathrm{V}_n = \widetilde{f}_n(\widetilde{\mathrm{PA}}(\mathrm{V}_n), \widetilde{\mathrm{U}}_n)$, where $\widetilde{\mathrm{PA}}(\mathrm{V}_n) \subseteq \{\mathrm{V} \in \mathbf{V} \mid \mathrm{V} \prec \mathrm{V}_n\}$.*

Interventions represent manipulations of the causal system on the second level of Pearl's hierarchy. Rather than simply inferring correlations in the data, they allow us to manipulate the generation process and generate outcomes from a modified ASEM. Interventions, however, do not reason about individual outcomes—they only allow us to sample unrelated new observations. The third level of the causal hierarchy concerns itself with *retrospective* modifications of the ASEM, investigating what would have happened at the time of sampling *had* the ASEM been different, i.e., had an intervention been performed. This is formalized with counterfactual distributions.

Definition 2.3 (Counterfactual Distribution)**.** *Given an ASEM $\mathcal{M} = (\mathbf{V}, \mathbf{U}, \mathcal{P}, \mathcal{F}, \prec)$ and an instantiation $\mathbf{v}$ of the endogenous variables, the **counterfactual** distribution under the intervention $\mathrm{do}(\mathrm{V}_n = \widetilde{f}(\widetilde{\mathrm{PA}}(\mathrm{V}_n), \widetilde{\mathrm{U}}_n))$ is the distribution over $\mathbf{V}$ defined by the intervened-upon ASEM whose exogenous variables follow the posterior distribution $\mathbb{P}(\mathbf{U} \mid \mathbf{V} = \mathbf{v})$.*

Note that the exogenous variables in the intervened-on ASEM need not be mutually independent.

ASEMs are defined over *finitely* many RVs $\mathbf{U} \cup \mathbf{V}$. We offer a generalization to countably many variables, assuming that the dependencies between them respect a well-founded order.

Definition 2.4 (Well-founded Order)**.** *A partial order $\prec$ over $\mathbf{V} = \{\mathrm{V}_1, \mathrm{V}_2, \ldots\}$ is **well-founded** if and only if there are no infinite descending sequences $\mathrm{V}_{n_1} \succ \mathrm{V}_{n_2} \succ \mathrm{V}_{n_3} \succ \cdots$.*

Definition 2.5 (Well-founded SEM)**.** *A **well-founded SEM** (WSEM) is a five-tuple $\mathcal{M} = (\mathbf{V}, \mathbf{U}, \mathcal{P}, \mathcal{F}, \prec)$, where $\mathbf{V} = \{\mathrm{V}_1, \mathrm{V}_2, \ldots\}$ is a countable set of endogenous RVs, $\mathbf{U} = \{\mathrm{U}_1, \mathrm{U}_2, \ldots\}$ is a countable set of exogenous RVs, $\mathcal{P} = \{\mathbb{P}_1, \mathbb{P}_2, \ldots\}$ is a countable set of probability distributions where $\mathbb{P}_n$ is a distribution over $\mathrm{U}_n \in \mathbf{U}$, $\mathcal{F} = \{f_1, f_2, \ldots\}$ is a countable set of*

[2] Our use of the notation $\mathrm{PA}(\mathrm{V}_n)$ is suggestive of the fact that, when rendered as a causal graphical model, $\mathrm{PA}(\mathrm{V}_n)$ are the *parents* of V_n.

[3] Our definition of an ASEM exposes the variables $\mathbf{U}$ and $\mathbf{V}$ as well as the partial order $\prec$. Standard definitions instead usually leave them implicit in $\mathcal{F}$ and $\mathcal{P}$ (Peters et al., 2017, Def. 6.2).

[4] The representations are, however, not equivalent. SEMs are strictly more expressive and required for counterfactual analysis; graphical causal models cannot answer counterfactual questions because they do not encode the functional relationships necessary to simulate alternative scenarios (Peters et al., 2017, Tab. 1.1).

structural equations, and $\prec$ is a well-founded partial order on $\mathbf{V}$. We require that the endogenous variables satisfy

$$\mathrm{V}_n = f_n(\mathrm{PA}(\mathrm{V}_n), \mathrm{U}_n), \tag{3}$$

for $n \in \mathbb{N}$ where $\mathrm{PA}(\mathrm{V}_n) \subseteq \{\mathrm{V} \in \mathbf{V} \mid \mathrm{V} \prec \mathrm{V}_n\}$ and is implied by the domain of f_n.

Interventions and counterfactual distributions in WSEMs are defined analogously to their definition in ASEMs. While the possible dependence between the infinitely many exogenous variables complicates inference of the noise posterior for counterfactual inference, our formulation of an LM as a WSEM results in conditionally independent exogenous variables, which makes posterior inference possible. Well-foundedness implies a unique assignment of the infinitely many endogenous RVs (Peters & Halpern, 2021, §1), which can be obtained by inductively assigning values to V_n in accordance with $\prec$. Thus, this model defines a unique outcome for each intervention. We use the abbreviation SEM when discerning between ASEMs and WSEMs is not important.

2.2 Language processes as WSEMs

We next show how LMs can be framed as WSEMs. We begin by defining the Gumbel distribution.

Definition 2.6 (Gumbel distribution)**.** *The cumulative distribution function of the **Gumbel distribution** shifted by the parameter $\alpha \in \mathbb{R}_{\geq 0}$, which we denote by $\mathrm{Gumbel}(\alpha)$, is $F_\alpha(x) = \exp(-\exp(-x+\alpha))$ and its probability density function is $f_\alpha(x) = \exp(-x+\alpha)\exp(-\exp(-x+\alpha))$. With $\mathrm{TruncGumbel}(\alpha, \beta)$, we denote the distribution $\mathrm{Gumbel}(\alpha)$ truncated at $\beta \in \mathbb{R}_{\geq 0}$, i.e., the distribution with the density $g_\alpha^\beta(x) = \frac{f_\alpha(x)\mathbb{1}\{x \leq \beta\}}{F_\alpha(\beta)}$.*

The Gumbel is useful for modeling the distribution of the maximum (or minimum) of a set of samples from various distributions. This is the core idea behind the Gumbel-max trick, which shows the utility of the Gumbel distribution for sampling from the categorical distribution (Luce, 1959; Yellott, 1977; Hazan & Jaakkola, 2012; Maddison et al., 2014; Hazan et al., 2016; Maddison et al., 2017). We restate the trick below for the specific case of the softmax.

Theorem 2.1. *Let X be a categorical RV over $\{1, \ldots, M\}$ such that*

$$\mathbb{P}(\mathrm{X} = m) = \frac{\exp(\phi(m))}{\sum_{m'=1}^{M} \exp(\phi(m'))} = \mathrm{softmax}(\boldsymbol{\phi})_m, \tag{4}$$

for $m \in \{1, \ldots, M\}$ and a vector $\boldsymbol{\phi} \in \mathbb{R}^M$.[5] Then, for $\mathrm{U}_m \overset{\text{i.i.d.}}{\sim} \mathrm{Gumbel}(0)$, we have

$$\mathrm{X} \overset{d}{=} \operatorname*{argmax}_{m=1}^{M} (\phi(m) + \mathrm{U}_m), \tag{5}$$

where $\overset{d}{=}$ refers to equality in distributions.

Proof. The proof is standard; we include it in App. C.1 for completeness. ■

As we formalize below, the Gumbel-max trick can be used to sample from a representation-based LM, since the (affinely transformed) representations $\boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{w})$ provide the values $\phi(m)$ in Eq. (5).

A **language process** (LP) $\mathbf{W} = \{\overline{\mathrm{W}}_t\}_{t=1}^{\infty}$ is an infinite sequence of (correlated) $\overline{\Sigma}$-valued RVs. $\overline{\mathrm{W}}_t$ is a RV whose outcome is the tth symbol of a string. A typical formulation of an LP implies that, if $\overline{\mathrm{W}}_t = \textsc{eos}$, we have $\overline{\mathrm{W}}_{t'} = \textsc{eos}$ for all $t' > t$ (Du et al., 2024). This can be achieved by setting $\ell(\textsc{eos}; \boldsymbol{w}_{<t}) = \infty$ and $\ell(w; \boldsymbol{w}_{<t}) = -\infty$ for $w \in \Sigma$ if $\boldsymbol{w}_{<t}$ contains $\textsc{eos}$.[6] We say that the string $\boldsymbol{w} = w_1 \cdots w_T$ was sampled from the LP $\mathbf{W}$ if $\overline{\mathrm{W}}_t = w_t$ for $t \leq T$ and $\overline{\mathrm{W}}_t = \textsc{eos}$ for $t > T$.

Any LM clearly induces an LP with $\mathbb{P}(\overline{\mathrm{W}}_t = \overline{w}_t \mid \overline{\mathbf{W}}_{<t} = \boldsymbol{w}_{<t}) = p(\overline{w}_t \mid \boldsymbol{w}_{<t})$. Concretely, as explicated by Eq. (1c), representation-based LMs sample a string by sampling from countably-infinitely many $\overline{\Sigma}$-valued RVs as $(\overline{\mathrm{W}}_t \mid \overline{\mathbf{W}}_{<t} = \boldsymbol{w}_{<t}) \sim \mathrm{softmax}(\mathbf{E}\,\boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{w}_{<t}) + \mathbf{b})$. *Precisely* how this sampling is done is not specified, as sampling from the softmax can be done in several ways.

[5] This, naturally, assumes that none of the probabilities are 0, which is a common assumption both in language modeling as well as in decision theory (Yellott, 1977; Cotterell et al., 2024).

[6] We can reconcile this with the definition of a representation-based LM by hard-coding this rule into $\boldsymbol{h}_{\boldsymbol{\theta}}$.

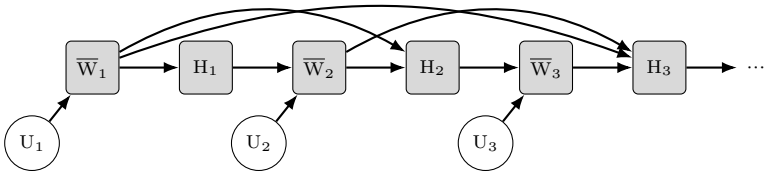

Figure 1: A language process as a WSEM.

The exact mechanism is not of importance when one is concerned with correlational or interventional questions—those only require that the distributions of $\overline{\mathrm{W}}_t$ match. To perform counterfactual analysis, however, specifying the exact causal mechanism is required. This is why identifying counterfactual distributions from observational or interventional data is, in general, impossible—two SEMs can be equal in all interventional distributions and yet define different counterfactual outcomes.[7] To be able to talk about string counterfactuals, one thus has to define an LM more precisely, in effect specifying how sampling from $p(\overline{w}_t \mid \boldsymbol{w}_{<t})$ is done exactly. To this end, we define a representation-based **Gumbel-max language process** as an LP in which

$$\overline{\mathrm{W}}_t \mid \overline{\mathbf{W}}_{<t} = \boldsymbol{w}_{<t} = \operatorname*{argmax}_{\overline{w} \in \overline{\Sigma}} \left(\mathbf{E}\, \boldsymbol{h}_{\boldsymbol{\theta}}\left(\boldsymbol{w}_{<t}\right) + \mathbf{b}\right)_{\overline{w}} + \mathrm{U}_t\left(\overline{w}\right), \tag{6}$$

where $\boldsymbol{w}_{<t} \stackrel{\text{def}}{=} w_1 \cdots w_{t-1}$ and $\mathbf{U} = \{\mathrm{U}_t(\overline{w})\}_{t=1}^{\infty}$ is an infinite sequence of i.i.d. RVs indexed by $t \in \mathbb{N}$ and $\overline{w} \in \overline{\Sigma}$. For brevity, we also introduce the RVs $\mathrm{H}_t \stackrel{\text{def}}{=} \mathbf{E}\, \boldsymbol{h}_{\boldsymbol{\theta}}\left(\overline{\mathbf{W}}_{<t}\right) + \mathbf{b}$, which correspond to the (deterministic) vectorial representations of strings sampled from the LP $\mathbf{W}$.

Eq. (6) frames $\mathbf{W}$ as a **Thurstone discriminal process**. In principle, $\mathrm{U}_t(\overline{w})$'s distribution could be arbitrary. Thurstone (1927) originally used normally distributed variables. However, as Thm. B.1 in App. B shows, in case we want Eq. (6) to match the distribution of the representation-based LM, $\mathrm{U}_t(\overline{w})$ *must* be Gumbel-distributed, thus motivating the name. Assuming a Thurstonian model thus implicitly defines a *single* counterfactual distribution for any intervention and outcome, which requires some justification. Our choice reflects an appeal to precedent and its established role in decision-making research; this model, originally designed to infer preferences through pairwise comparisons, has since enabled robust probabilistic analyses of choice modeling (Noothigattu et al., 2020; Vojnovic & Yun, 2016). Moreover, the simplicity and familiarity of the Gumbel-max LP provide a natural first step and pave the way for other choices beyond Gumbel noise (Lorberbom et al., 2021; Haugh & Singal, 2023; Chatzi et al., 2024). In App. B, we discuss some alternatives.

A crucial implication of Eq. (6) is that $\mathbf{W}$ is *deterministic* given $\mathbf{U}$—all the noise in the string generation process comes from the noise variables $\mathbf{U}$. This decomposition of the generative mechanism into deterministic relationships and independent noise variables paints an LP as a WSEM—precisely, it presents *one possible* specification of a WSEM that induces the same probability distribution over Σ^* as a representation-based LM. For convenience, we capture this in the following construction.

Construction 2.1 (A WSEM for an LM)**.** *Let $\boldsymbol{h}_{\boldsymbol{\theta}}$ be a language encoder and p the LM induced by $\boldsymbol{h}_{\boldsymbol{\theta}}$ together with the parameters $\mathbf{E}, \mathbf{b}$. We can define an WSEM $\mathcal{M} = (\mathbf{V}, \mathbf{U}, \mathcal{P}, \mathcal{F}, \prec)$ that induces the same distribution over Σ^* as p as follows:*

- $\mathbf{V} = \{\overline{\mathrm{W}}_t\}_{t=1}^{\infty} \cup \{\mathrm{H}_t\}_{t=1}^{\infty} \cup \{\boldsymbol{\theta}, \mathbf{E}, \mathbf{b}\}$*;*
- $\mathbf{U} = \{\mathrm{U}_t(\overline{w}) \mid t \in \mathbb{N}, \overline{w} \in \overline{\Sigma}\}$*;*
- $\mathcal{P} = \left\{\mathrm{U}_t(\overline{w}) \sim \mathrm{Gumbel}(0) \mid t \in \mathbb{N}, \overline{w} \in \overline{\Sigma}\right\}$*;*
- *$\mathcal{F}$ is defined through the computation graph of the LM: $\{\boldsymbol{\theta}, \mathbf{E}, \mathbf{b}\}$ are fixed, $\mathrm{H}_t = \mathbf{E}\, \boldsymbol{h}_{\boldsymbol{\theta}}\left(\overline{\mathbf{W}}_{<t}\right) + \mathbf{b}$, and $\overline{\mathrm{W}}_t = \operatorname{argmax}_{\overline{w} \in \overline{\Sigma}} \mathrm{H}_t(\overline{w}) + \mathrm{U}_t(\overline{w})$; and*
- *$\prec$ is defined naturally on $\overline{\mathrm{W}}_t$ as $\overline{\mathrm{W}}_{t'} \prec \overline{\mathrm{W}}_t$ for $t' < t$. We can set an arbitrary ordering of the (finitely many) parameters $\{\boldsymbol{\theta}, \mathbf{E}, \mathbf{b}\}$ and assert $\theta \prec \overline{\mathrm{W}}_t$ for any $\theta \in \{\boldsymbol{\theta}, \mathbf{E}, \mathbf{b}\}$ and $t \in \mathbb{N}$.*

Clearly, Construction 2.1 constructs a well-defined WSEM. The only infinite chain of endogenous RVs is $\{\overline{\mathrm{W}}_t\}_{t=1}^{\infty} \cup \{\mathrm{H}_t\}_{t=1}^{\infty}$, where $\overline{\mathrm{W}}_t$ and H_t depend on $\{\overline{\mathrm{W}}_{t'}\}_{t'<t} \cup \{\mathrm{H}_{t'}\}_{t'<t}$ and $\{\mathrm{U}_{t'}\}_{t' \leq t}$ with the root $\overline{\mathrm{W}}_1$. The WSEM constructed by Construction 2.1 for an LP is graphically depicted in Fig. 1.

[7] This is analogous to how graphical models do not reach the third level of Pearl's hierarchy (cf. §2.1) and how counterfactuals are not identifiable based on interventional data alone (cf. App. B).

An intervention on $\boldsymbol{\theta}, \mathbf{E}, \mathbf{b}$ and $\{\overline{\mathrm{W}}_t\}_{t=1}^{\infty}$ defines $\widetilde{\boldsymbol{\theta}}, \widetilde{\mathbf{E}}, \widetilde{\mathbf{b}}$, and modified symbols $\{\widetilde{\overline{\mathrm{W}}}_{t'}\}_{t' \in \mathcal{N}}$ for $\mathcal{N} \subseteq \mathbb{N}$. Given an instantiation of the exogenous variables $\mathbf{u} = \left\{\mathrm{u}_t\left(\overline{w}\right) \mid t \in \mathbb{N}, \overline{w} \in \overline{\Sigma}\right\}$, we can obtain the remaining symbols as $\widetilde{\overline{w}}_t \stackrel{\text{def}}{=} \operatorname{argmax}_{\overline{w} \in \overline{\Sigma}} \left(\widetilde{\mathbf{E}}\, \boldsymbol{h}_{\widetilde{\boldsymbol{\theta}}}(\widetilde{\boldsymbol{w}}_{<t}) + \widetilde{\mathbf{b}}\right)_{\overline{w}} + \mathrm{u}_t\left(\overline{w}\right)$ for $t \in \mathbb{N} \setminus \mathcal{N}$. Since any context–intervention pair in a WSEM defines a single possible outcome, this results in a single counterfactual string for the particular instantiation of the noise variables.

Sampling techniques. Our formalization assumes that strings are generated by sampling from the full probability distribution defined by an LM. In practice, however, different decoding techniques, such as nucleus sampling (Holtzman et al., 2020) or top-k sampling (Fan et al., 2018), are often used. As long as these decoding methods can be expressed as deterministic functions over the logits, followed by standard sampling, the same formulation can be applied.[8] This way, the deterministic parts of the sampling algorithm are considered a part of the LM forward pass computation.

3 Counterfactual generation

Framing LMs as WSEMs allows us to use the expansive set of causal tools on LMs. We focus on generating counterfactuals for given strings—these counterfactuals differ in specific features but are generated using the same sampling noise as the original strings. More precisely, let $\boldsymbol{w} = w_1 \cdots w_T \in \Sigma^*$ be the string sampled from the LM induced by the encoder $\boldsymbol{h}_{\boldsymbol{\theta}}$ with the parameters $\mathbf{E}$ and $\mathbf{b}$, and the instantiation of the noise $\mathbf{U} = \mathbf{u}$. Given a counterfactual encoder $\boldsymbol{h}_{\widetilde{\boldsymbol{\theta}}}$ with the parameters $\widetilde{\mathbf{E}}$ and $\widetilde{\mathbf{b}}$, Eq. (6) tells us that the symbols of $\boldsymbol{w}$'s counterfactual with $\mathbf{U} = \mathbf{u}$ are given by

$$\widetilde{\overline{w}}_t = \operatorname*{argmax}_{\overline{w} \in \overline{\Sigma}} \left(\widetilde{\mathbf{E}}\, \boldsymbol{h}_{\widetilde{\boldsymbol{\theta}}}\left(\widetilde{\boldsymbol{w}}_{<t}\right) + \widetilde{\mathbf{b}}\right)_{\overline{w}} + \mathrm{u}_t\left(\overline{w}\right). \tag{7}$$

This procedure results in *pairs* of strings in Σ^*—the original string $\boldsymbol{w}$ and its counterfactual $\widetilde{\boldsymbol{w}}$.

In practice, the counterfactual network $\boldsymbol{h}_{\widetilde{\boldsymbol{\theta}}}$ is created from $\boldsymbol{h}_{\boldsymbol{\theta}}$ by making feature-specific modifications, such as removing gender information from the representations $\boldsymbol{h}_{\boldsymbol{\theta}}\left(\boldsymbol{w}\right)$. Ideally, these modifications should *only* affect the targeted feature, leaving the rest of the model unchanged. This effect should be observable at the string level—for example, if the surgery is intended to change the grammatical number of a noun, that should be the sole difference between the original string and its counterfactual.[9] Without a clear definition of counterfactuality, however, it is difficult to evaluate the impact of representational surgeries, since we lack string pairs where the *only* difference is the surgery itself. Our framework addresses this by ensuring that a string $\boldsymbol{w}$ and its counterfactual $\widetilde{\boldsymbol{w}}$ form a **minimal pair** with respect to the intervened feature. A key goal of our experimental setup is to leverage this causal framework to evaluate the stability of various representational surgeries.

However, when evaluating the effects of interventions, we are not solely interested in minimal pairs. Another important question is: How would a *given* string have appeared, had it been generated by the counterfactual model rather than the original one? Answering this question requires knowledge of the exogenous noise that produced the original strings. This entails inferring the values (or, more precisely, the distribution) of the unobserved noise variables $\mathbf{U}$ that led to a particular observed string $\boldsymbol{w}$. Once the specific outcomes of $\mathbf{U}$ are identified, we can generate the corresponding counterfactuals. We tackle the problem of inferring $\mathbf{U}$ by developing an algorithm that reverses the causal process illustrated in Fig. 1. The algorithm hinges on (and, in effect, implements), the following proposition.

Proposition 3.1 (Hindsight Gumbel Sampling)**.** *Let $\boldsymbol{w} = w_1 \cdots w_T \in \Sigma^*$ be sampled from an LP $\mathbf{W}$. To sample from $\mathrm{U}_t\left(\overline{w}\right) \mid \overline{\mathrm{W}}_{\leq t} = \boldsymbol{w}_{\leq t}, \mathrm{U}_{t'}\left(\overline{w}'\right) = \mathrm{u}_{t'}\left(\overline{w}'\right)$ for $\overline{w}, \overline{w}' \in \overline{\Sigma}$, $t \leq T$ and $t' < t$, we can proceed in the following steps independently for $t = 1, \ldots, T$:*

(1) sample $\mathrm{y}_t\left(w_t\right)$ from $\mathrm{Gumbel}\left(\log \sum_{\overline{w} \in \overline{\Sigma}} \exp\left(\ell\left(\overline{w}; \boldsymbol{w}_{<t}\right)\right)\right)$,
(2) sample $\mathrm{y}_t\left(\overline{w}\right)$ from $\mathrm{TruncGumbel}\left(\ell\left(\overline{w}; \boldsymbol{w}_{<t}\right), \mathrm{y}_t\left(w_t\right)\right)$ independently for $\overline{w} \in \overline{\Sigma} \setminus \{w_t\}$, and
(3) set $\mathrm{u}_t\left(\overline{w}\right) = \mathrm{y}_t\left(\overline{w}\right) - \ell\left(\overline{w}; \boldsymbol{w}_{<t}\right)$ for all $\overline{w} \in \overline{\Sigma}$.

For $t > T$, sample $\mathrm{u}_t\left(\overline{w}\right)$ from $\mathrm{Gumbel}\left(0\right)$ for $\overline{w} \in \overline{\Sigma}$.

[8] For example, in top-k sampling, we can set the logits of all tokens outside the top k to a large negative value and then sample using the Gumbel-max trick.

[9] This criterion is known as counterfactual stability (Guerner et al., 2024).

Algorithm 1 An algorithm that samples counterfactual strings given a factual string.

def GENERATECOUNTERFACTUAL$(\boldsymbol{w}, (\boldsymbol{h}_{\boldsymbol{\theta}}, \mathbf{E}, \mathbf{b}), (\boldsymbol{h}_{\widetilde{\boldsymbol{\theta}}}, \widetilde{\mathbf{E}}, \widetilde{\mathbf{b}}))$:
  for $t \in \{1, \ldots, |\boldsymbol{w}|\}$: ▷ *sample from the noise posterior*
    $\boldsymbol{\ell}(\boldsymbol{w}_{<t}) \leftarrow \mathbf{E}\, \boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{w}_{<t}) + \mathbf{b}$
    sample $\mathrm{y}_t(w_t)$ from Gumbel $\left(\log \sum_{\overline{w} \in \overline{\Sigma}} \exp\left(\ell\left(\overline{w}; \boldsymbol{w}_{<t}\right)\right)\right)$ ▷ *the maximum value*
    for $\overline{w} \in \overline{\Sigma} \setminus \{w_t\}$:
      sample $\mathrm{y}_t(\overline{w})$ from TruncGumbel $\left(\ell\left(\overline{w}; \boldsymbol{w}_{<t}\right), \mathrm{y}_t(w_t)\right)$ ▷ *shifted noise*
  $t \leftarrow 1, \widetilde{\overline{w}}_t \leftarrow$ BOS, $\widetilde{\boldsymbol{w}} \leftarrow \widetilde{\overline{w}}_t$
  while $\widetilde{\overline{w}}_t \neq$ EOS : ▷ *generate a counterfactual*
    $\widetilde{\boldsymbol{\ell}}(\widetilde{\boldsymbol{w}}) \leftarrow \widetilde{\mathbf{E}}\, \boldsymbol{h}_{\widetilde{\boldsymbol{\theta}}}(\widetilde{\boldsymbol{w}}) + \widetilde{\mathbf{b}}$
    if $t \leq |\boldsymbol{w}|$: $\widetilde{\overline{w}}_t \leftarrow \operatorname{argmax}_{\overline{w} \in \overline{\Sigma}} \widetilde{\ell}(\overline{w}; \widetilde{\boldsymbol{w}}) + \mathrm{y}_t(\overline{w}) - \ell(\overline{w}; \boldsymbol{w}_{<t})$
    else: $\widetilde{\overline{w}}_t \leftarrow \operatorname{argmax}_{\overline{w} \in \overline{\Sigma}} \widetilde{\ell}(\overline{w}; \widetilde{\boldsymbol{w}}) + \mathrm{u}_t(\overline{w})$ ▷ ***sample*** $\mathrm{u}_t(\overline{w})$ *from* Gumbel (0)
    append $\widetilde{\overline{w}}_t$ to $\widetilde{\boldsymbol{w}}$, $t \leftarrow t + 1$
  return $\widetilde{\boldsymbol{w}}$

Proof. See App. C.2. ■

Corollary 3.1 (Counterfactual String Sampling)**.** *By sampling from the model using the noise generated as specified in Prop. 3.1, we get a sample from the counterfactual distribution.*

We employ a standard technique for sampling from "truncated" (conditional) distribution (Maddison et al., 2014).[10] In our case, the truncation condition ensures that the observed word $\overline{w}_t$ has a higher score than all other vocabulary tokens to mimic Eq. (7). This procedure, summarized in Alg. 1, allows us to generate potential counterfactual sentences for a given observed sentence.

4 EXPERIMENTS

4.1 SIDE EFFECTS OF COMMON INTERVENTION TECHNIQUES

Many standard intervention techniques, such as knowledge editing (Meng et al., 2022; 2023) or inference-time intervention (Li et al., 2024; Singh et al., 2024) are intended to modify targeted aspects of model behavior, such as altering specific knowledge or increasing its truthfulness (Li et al., 2024). If these interventions are surgical, we expect them to preserve the model's behavior on unrelated, "neutral" sequences—such as random Wikipedia sentences, resulting in counterfactuals similar to the original sentence. We test this assumption using Alg. 1 as follows. We use several intervention techniques, detailed below, to induce changes to the LM, either by modifying the encoder parameters or the string representations directly. Such intervention techniques, however, do not generate string counterfactuals. We generate string counterfactuals that correspond to the interventions by the following steps: *(1)* we apply an intervention to the original model (the *base model*); *(2)* we use Alg. 1 to hindsight-sample from the posterior of the exogenous variable (the Gumbel noise), under the base model, conditioned on an observed sentence; and *(3)* we use the sampled noise to generate a *counterfactual string* from an intervened-on model.

4.1.1 EXPERIMENTAL SETUP

Setup. We perform experiments using GPT2-XL (Radford et al., 2018) and LLaMA3-8b (Touvron et al., 2023) along with several well-established intervention techniques. These include MEMIT (Meng et al., 2023), inference-time interventions using linear steering (Li et al., 2024; Singh et al., 2024), and Instruction tuning (Touvron et al., 2023). See App. D.1 for an overview.

In each case, we define the model prior to the intervention as the original model and the model following the intervention as the counterfactual model. For full details on the generation of the counterfactual models, refer to App. D.2. For each original and counterfactual model pair, we generate

[10] The algorithm given by `https://timvieira.github.io/blog/post/2020/06/30/generating-truncated-random-variates/` is used for performing the "truncated" sampling.

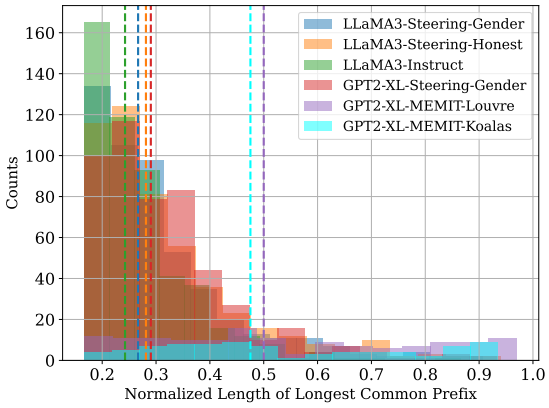

Figure 2: Normalized length of the longest shared prefix between the original and counterfactual sentences, for different intervention techniques. The horizontal lines denote the median of each distribution. Higher values reflect less side effects for the intervention.

500 sentences by using the first five words of randomly selected English Wikipedia sentences as prompts for the original model. We generate a continuation of a maximum of 25 tokens by sampling from the model using multinomial sampling (i.e., sampling from the entire model distribution over the vocabulary). We then use Alg. 1 to generate a counterfactual sentence.

Evaluation. Being prompted by a prefix from Wikipedia, the original model is not likely to generate a continuation that exhibits a property that is the focus of *any* of the specific model intervention techniques we examine (e.g., it is not likely to generate a sentence that discusses the location of the Louvre, for the MEMIT intervention). Accordingly, we expect the counterfactual strings to be similar to the original ones. This is desirable, as we ideally want surgical intervention without side effects. To quantify side effects on arbitrary strings, we record the longest common prefix, which we define as the length of the longest prefix of the original sentence that is shared with the counterfactual sentence normalized by the length of the original sentence. To evaluate the semantic similarity between the original and counterfactual model, we calculate cosine similarity under the E5-base encoder (Wang et al., 2024).[11] Surgical interventions should have high values for both metrics.

Original and Counterfactual strings for the LLaMA3 Instruct finetuning intervention.

1. **Original:** Chenopodium nutans, commonly called climbing, or nodding goosefoot, is an annual plant...
Counterfactual: Chenopodium nutans, commonly called climbing, or leafy goosefoot, is an annual plant...

2. **Original:** Brittany Haas is an American fiddler and a founding member of the acclaimed trio Hawktail.
Counterfactual: Brittany Haas is an American fiddler and composer.

3. **Original:** Richard Joseph Grosh (born October 28, 1935) was Director of the US Securities and Exchange Commission
Counterfactual: Richard Joseph Grosh (born October 24, 1935) was an American politician who served as a member of the U.S. House of Representatives from...

4. **Original:** It was also included on a limited edition vinyl 7" with "Tape Loop"...
Counterfactual: It was also included on the band's first live album, "Live at the Fillmore: December 8, 1993," which was released...

5. **Original:** The series consists of four items: a letter written by Lt. Col. Edward S. Durnford (1918), a draft contract for the...
Counterfactual: The series consists of four episodes and a special edition episode. Each episode will explore a different theme and feature interviews with experts and individuals who have been affected...

Figure 3: Counterfactual strings from the original model `LLaMA3` and the counterfactual counterpart `LLaMA3-Instruct`.

[11]We use the `E5-base-v2` model from `https://huggingface.co/intfloat/e5-base-v2`.

4.1.2 RESULTS

The distribution of the normalized length of the longest common prefix is shown in Fig. 2. Among the methods, MEMIT demonstrates the most precise intervention, with a median longest shared prefix length of around 50% for both the Louvre and Koalas concepts. The steering vector interventions follow at around 30%, with the instruction tuning intervention being the least surgical, sharing only around 24% of tokens on average. These trends are also reflected in the cosine similarity under the E5 model, which is 0.976 and 0.986 for the MEMIT Koalas and Louvre interventions, and around 0.860 for all other interventions.[12]

Fig. 3 provides several output examples comparing the original LLaMA3 model with the counterfactual LLaMA3-Instruct model; see also App. F.1 for a random sample of outputs from all models. In some cases, such as the first two examples, the intervention introduces factual inaccuracies, which are evident in the generated counterfactuals. The third example demonstrates a case where *both* the original and the counterfactual model hallucinate (the subject of the sentence was actually an academic), but the content of the hallucination changes as a result of the intervention. Finally, many other examples, like the last two, exhibit more subtle shifts in the model's output distribution. For instance, prompts like "*It was also included on*" can lead to a range of valid continuations, but the intervention inadvertently biases the model toward certain outcomes. These results indicate that even interventions that are designed to be "minimal", such as those based on a steering vector that only modifies a tiny fraction of all the model's parameters, still have considerable causal effect on the output of the model, as demonstrated by the semantic drift in the continuations of prompts taken from Wikipedia. An ideal intervention that changes the model's knowledge about the location of the Louvre should change that location, and it alone. In practice, however, even interventions such as MEMIT, that update few parameters in a single matrix within the model, have considerable side effects. Due to the autoregressive nature of language generation, slight variations in token choice accumulate rapidly, resulting in a significant semantic divergence between the original and the counterfactual sentence.

4.2 INTERVENTION-FOCUSED COUNTERFACTUALS

In the previous section, we examine how surgical the different interventions are. Accordingly, we focused the evaluation on prompts drawn from Wikipedia, a domain we expect to be largely orthogonal to the specific properties on which the interventions target. Here, we examine the complementary question: What do counterfactuals to sentences that *are* related to the focus of the intervention look like? We focus on two case studies: MEMIT, which edits for the location of the Louvre, and MiMiC, which employs a steering intervention to push the model in the male $\rightarrow$ female direction.

4.2.1 GENDER STEERING

Setup. We focus this analysis on gender steering of the LLaMA3-Instruct model. We apply the MiMiC method (Singh et al., 2024), which modifies model representations by aligning male-focused text representations with female-focused text representations. This approach results in a linear transformation of the residual stream, ensuring that the mean and covariance of the source class (male-focused texts) match those of the target class (female-focused texts). We fit the transformation on the Bios dataset (De-Arteaga et al., 2019), which consists of short biographies of individuals working in various professions. Each biography is annotated with both gender and profession labels. For the full details on the fitting of the MiMiC intervention, see App. D.2. Once the intervention is fitted, we first generate a continuation for 500 biographies in the dataset by sampling from the original model after prompting it with the first words in the biography, and then use Alg. 1 to generate counterfactual continuations under the modified models.

Results. A sample of the results can be found in Fig. 4 and App. F.2. The intervention demonstrates reasonable effectiveness in altering the pronouns used in the continuations. While the original biographies all contain male pronouns, in 52.2% of the counterfactual continuations, only female pronouns such as "she" and "her" are observed. In 23.2% of the cases, male pronouns persist, while 16.6% show a mixture of female and male pronouns, and 7.6% of the counterfactuals do not include

[12]Cosine similarity of 0.867, 0.873, 0.870 and 0.863 for the LLaMA3 gender-steering, LLaMA3 honest-steering, LLaMA3 Instruction finetuning and GPT2 gender steering, respectively.

Text Examples: Originals and Counterfactuals for the MiMiC gender steering intervention.

1. **Original:** Kyle Thompson is an assistant professor at Pittsburg State University in Kansas, where he teaches courses on marketing, entrepreneurship, and management.
Counterfactual: Kyle Thompson is an assistant professor at Pittsburg State University in Kansas, where she teaches courses in early childhood education and child development.

2. **Original:** Clayton Webb is an assistant professor of political science at the University of Tennessee. His research focuses on political behavior, public opinion, and the psychology of politics.
Counterfactual: Clayton Webb is an assistant professor of political science at the University of Tennessee. Her research focuses on political behavior, public opinion, and the intersection of politics and social issues. She is particularly interested in...

3. **Original:** Jeffrey M. Stanton, Ph.d., is an assistant professor at the University of Rochester Medical Center. He conducts research in the area of statistical genetics, focusing on the development of methods for analyzing large-scale genomic data...
Counterfactual: Jeffrey M. Stanton, Ph.d., is an assistant professor at the University of Rochester's Warner School of Education, where she teaches and conducts research in the areas of educational technology, online teaching and learning, and educational psychology.

4. **Original:** Nicholas F. Benson, Ph.d., is an associate professor of chemistry at the University of Illinois at Chicago (UIC). He received his B.S. in chemistry from the University of Illinois at Urbana-Champaign in 2003 and his Ph.d....
Counterfactual: Nicholas F. Benson, Ph.d., is an associate professor of chemistry at the University of Illinois at Chicago (UIC). She received his B.S. in chemistry from the University of Illinois at Urbana-Champaign in 2003 and her Ph.d....

Figure 4: Counterfactual strings from the original model `LLaMA3` and the counterfactual counterpart `Steering-Gender LLaMA3` (created by a MiMiC steering intervention in the direction male $\rightarrow$ female.). The first 8 tokens in the original sentence where used as a prompt.

any pronouns at all. An examination of the counterfactual continuations in Fig. 4 and App. F.2 reveals many side effects beyond pronouns. See App. E for a quantitative evaluation.

4.2.2 MEMIT LOCATION EDITING

Setup. We focus this analysis on the MEMIT-edited model, where the location of the Louvre was updated from "Paris" to "Rome". We begin by prompting the original model to generate sentences that mention Paris as the location of the Louvre, such as "*Paris offers many attractions, but the*". See App. D.3 for details. We filter out sentences that do not mention both Paris and the Louvre, resulting in 75 sentences. We then generate the counterfactuals with the counterfactual model.

Results. First, we observe that the Louvre-focused counterfactuals deviate much more from the semantics of the original sentences than non-Louvre-focused generations. The median normalized longest prefix consists of only 23% of the tokens in the original sentences, compared with around 50% in the Wikipedia-based counterfactuals. A manual inspection of the counterfactuals reveals a significant deviation from the semantics of the original sentences. For instance, the counterfactual of "*Among all the art museums in the world, the Louvre stands tall in the Parisian art scene, having been at the same spot since 1793*" is the sentence "*Among all the art museums in the world, the Louvre alone, which has been open to visitors for more than 2,000 years, is one of the most visited.*". Altogether, *the counterfactuals are mostly* ***not*** *minimal*: they do not change just the location of the Louvre, but other (unrelated) parts of the sentence. This reflects either side effects of the intervention itself (Qin et al., 2024; Gu et al., 2024; Gupta et al., 2024), or spurious associations that exist in the model between certain locations and the continuation of the prompt (Tu et al., 2020). With respect to correctness, we find that 60.0% of the counterfactuals mention Rome as the location of the Louvre, while 40.0% still mention Paris.

5 CONCLUSION

We introduce a framework for generating true counterfactuals from LMs by reformulating LMs as generalized structural equation models with the Gumbel-max trick. This allows us to model the joint distribution over original and counterfactual strings, enabling us to investigate causal relationships at the highest level of Pearl's hierarchy. Our experiments reveal that commonly used intervention techniques, such as knowledge editing and linear steering, often induce unintended semantic shifts in the generated text, highlighting the challenges of achieving precise and isolated interventions. These observations underline the need for more refined methods that can achieve targeted modifications with minimal collateral changes to the model's outputs.

REPRODUCIBILITY STATEMENT

We detail our experimental setup in §4.1.1 and App. D.2.

REFERENCES

Eldar David Abraham, Karel D'Oosterlinck, Amir Feder, Yair Ori Gat, Atticus Geiger, Christopher Potts, Roi Reichart, and Zhengxuan Wu. Cebab: Estimating the causal effects of real-world concepts on NLP model behavior. *Advances in Neural Information Processing Systems*, 35: 17582–17596, 2022. URL https://arxiv.org/abs/2205.14140.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1608.04207.

Victor Aguirregabiria and Pedro Mira. Dynamic discrete choice structural models: A survey. *Journal of Econometrics*, 156(1):38–67, 2010. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2009.09.007. URL https://www.sciencedirect.com/science/article/pii/S0304407609001985. Structural Models of Optimization Behavior in Labor, Aging, and Health.

Matan Avitan, Ryan Cotterell, Yoav Goldberg, and Shauli Ravfogel. Natural language counterfactuals through representation surgery. *arXiv preprint arXiv:2402.11355*, 2024. URL https://arxiv.org/abs/2402.11355.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

Robin SM Chan, Reda Boumasmoud, Anej Svete, Yuxin Ren, Qipeng Guo, Zhijing Jin, Shauli Ravfogel, Mrinmaya Sachan, Bernhard Schölkopf, Mennatallah El-Assady, and Ryan Cotterell. On affine homotopy between language encoders. *arXiv preprint arXiv:2406.02329*, 2024. URL https://arxiv.org/abs/2406.02329.

Ivi Chatzi, Nina Corvelo Benz, Eleni Straitouri, Stratis Tsirtsis, and Manuel Gomez-Rodriguez. Counterfactual token generation in large language models. *arXiv preprint arXiv:2409.17027*, 2024. URL https://arxiv.org/abs/2409.17027.

Ryan Cotterell, Anej Svete, Clara Meister, Tianyu Liu, and Li Du. *Formal Aspects of Language Modeling*. arXiv preprint arXiv:2311.04329, 2024. URL https://arxiv.org/abs/2311.04329.

Maria De-Arteaga, Alexey Romanov, Hanna M. Wallach, Jennifer T. Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Cem Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. *CoRR*, abs/1901.09451, 2019. URL http://arxiv.org/abs/1901.09451.

Li Du, Lucas Torroba Hennigen, Tiago Pimentel, Clara Meister, Jason Eisner, and Ryan Cotterell. A measure-theoretic characterization of tight language models. 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. 9744–9770, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.543. URL https://aclanthology.org/2023.acl-long.543.

Li Du, Holden Lee, Jason Eisner, and Ryan Cotterell. When is a language process a language model? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp. 11083–11094, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.659. URL https://aclanthology.org/2024.findings-acl.659.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://aclanthology.org/2021.tacl-1.10.

Nelson Elhage, Neel Nanda, Catherine Olsson, and Tom Henighan. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990. doi: https://doi.org/10.1207/s15516709cog1402_1. URL https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1402_1.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL https://aclanthology.org/P18-1082.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. CausaLM: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386, June 2021. doi: 10.1162/coli_a_00404. URL https://aclanthology.org/2021.cl-2.13.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158, 2022. doi: 10.1162/tacl_a_00511. URL https://aclanthology.org/2022.tacl-1.66.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *arXiv preprint arXiv:2301.04709*, 2024. URL https://arxiv.org/abs/2301.04709.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi (eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 240–248, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5426. URL https://aclanthology.org/W18-5426.

Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. Model editing harms general abilities of large language models: Regularization to the rescue. *arXiv preprint arXiv:2401.04700*, 2024. URL https://arxiv.org/abs/2401.04700.

Clément Guerner, Anej Svete, Tianyu Liu, Alexander Warstadt, and Ryan Cotterell. A geometric notion of causal probing. *arXiv preprint arXiv:2307.15054*, 2024. URL https://arxiv.org/abs/2307.15054.

Emil Julius Gumbel. *Statistical theory of extreme values and some practical applications: a series of lectures*, volume 33. US Government Printing Office, 1954. URL https://www.cambridge.org/core/journals/aeronautical-journal/article/abs/statistical-theory-of-extreme-values-and-some-practical-applications-lectures-by-emit-j-gumbel-nat 6A77FA9007B583F4B75CB4F0BEA535B2.

Akshat Gupta, Anurag Rao, and Gopala Anumanchipalli. Model editing at scale leads to gradual and catastrophic forgetting. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15202–15232, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.902. URL https://aclanthology.org/2024.findings-acl.902.

Trygve Haavelmo. The probability approach in econometrics. *Econometrica: Journal of the Econometric Society*, pp. iii–115, 1944.

Martin Haugh and Raghav Singal. Counterfactual analysis in dynamic latent-state models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023. URL https://dl.acm.org/doi/10.5555/3618408.3618922.

Tamir Hazan and Tommi Jaakkola. On the partition function and random maximum a-posteriori perturbations. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pp. 1667–1674, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851. URL https://dl.acm.org/doi/10.5555/3042573.3042786.

Tamir Hazan, George Papandreou, and Daniel Tarlow. *Perturbations, Optimization, and Statistics*. The MIT Press, 12 2016. ISBN 9780262337939. doi: 10.7551/mitpress/10761.001.0001. URL https://doi.org/10.7551/mitpress/10761.001.0001.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. Reducing sentiment bias in language models via counterfactual evaluation. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 65–83, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.7. URL https://aclanthology.org/2020.findings-emnlp.7.

Frederik Hvilshøj, Alexandros Iosifidis, and Ira Assent. ECINN: Efficient counterfactuals from invertible neural networks. *arxiv preprint arXiv:2103.13701*, 2021. URL https://arxiv.org/abs/2103.13701.

Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Diffusion models for counterfactual explanations. *arxiv preprint arXiv:2203.15636*, 2022. URL https://arxiv.org/abs/2203.15636.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. AtP*: An efficient and scalable method for localizing LLM behaviour to components. *arXiv*, 2403.00745, 2024. doi: 10.48550/ARXIV.2403.00745. URL https://arxiv.org/abs/2403.00745.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. The emergence of number and syntax units in LSTM language models. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *North American Chapter of the ACL*, pp. 11–20, June 2019. doi: 10.18653/v1/N19-1002. URL https://aclanthology.org/N19-1002.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: eliciting truthful answers from a language model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc. URL https://dl.acm.org/doi/10.5555/3666122.3667919.

Guy Lorberbom, Daniel D. Johnson, Chris J Maddison, Daniel Tarlow, and Tamir Hazan. Learning generalized gumbel-max causal mechanisms. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 26792–26803. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/e143c01e314f7b950daca31188cb5d0f-Paper.pdf.

R. Duncan Luce. *Individual choice behavior*. John Wiley, Oxford, England, 1959. URL https://books.google.ch/books/about/Individual_Choice_Behavior.html?id=ERQsKkPiKkkC.

R. Duncan Luce. Thurstone's discriminal processes fifty years later. *Psychometrika*, 42(4):461–489, 1977. URL https://link.springer.com/article/10.1007/BF02295975.

R. Duncan Luce. Thurstone and sensory scaling: Then and now. *Psychological Review*, 101(2):271–277, 1994. doi: 10.1037/0033-295X.101.2.271. URL https://doi.org/10.1037/0033-295X.101.2.271.

Nishtha Madaan, Inkit Padhi, Naveen Panwar, and Diptikalyan Saha. Generate your counterfactuals: Towards controlled counterfactual generation for text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 13516–13524, 2021. URL https://arxiv.org/abs/2012.04698.

Chris J. Maddison and Daniel Tarlow. Gumbel machinery, 2017. URL https://cmaddis.github.io/gumbel-machinery. Available online.

Chris J. Maddison, Daniel Tarlow, and Tom Minka. A* sampling. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pp. 3086–3094, Cambridge, MA, USA, 2014. MIT Press. URL https://dl.acm.org/doi/10.5555/2969033.2969171.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2017. URL https://arxiv.org/abs/1611.00712.

J. Chris Maddison and Daniel Tarlow. Gumbel machinery. 2014. URL https://cmaddis.github.io/gumbel-machinery.

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In Paul Zarembka (ed.), *Fontiers in Econometrics*, pp. 105–142. Academic press, New York, 1974. URL https://eml.berkeley.edu/reprints/mcfadden/zarembka.pdf.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Neural Information Processing Systems*, 2022. doi: 10.48550/ARXIV.2202.05262. URL https://arxiv.org/abs/2202.05262.

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023. URL https://openreview.net/forum?id=MkbcAHIYgyS.

Aaron Mueller. Missed causes and ambiguous effects: Counterfactuals pose challenges for interpreting neural networks. *arXiv preprint arXiv:2407.04690*, 2024. URL https://arxiv.org/abs/2407.04690.

Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. The quest for the right mediator: A history, survey, and theoretical grounding of causal interpretability. *arXiv preprint arXiv:2408.01416*, 2024. URL https://arxiv.org/abs/2408.01416.

Neel Nanda. Attribution patching: Activation patching at industrial scale. *mechanistic-interpretability*, 2023. URL https://www.neelnanda.io/mechanistic-interpretability/attribution-patching.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9XFSbDPmdW.

Ritesh Noothigattu, Dominik Peters, and Ariel D Procaccia. Axioms for learning from pairwise comparisons. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17745–17754. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/cdaa9b682e10c291d3bbadca4c96f5de-Paper.pdf.

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with Gumbel-max structural causal models. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4881–4890. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/oberst19a.html.

J. Pearl, M. Glymour, and N.P. Jewell. *Causal Inference in Statistics: A Primer*. Wiley, 2016. ISBN 9781119186847. URL https://books.google.ch/books?id=L3G-CgAAQBAJ.

Judea Pearl. *Probabilistic reasoning in intelligent systems - networks of plausible inference*. Morgan Kaufmann series in representation and reasoning. Morgan Kaufmann, 1989. URL https://dl.acm.org/doi/book/10.5555/534975.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X. URL https://bayes.cs.ucla.edu/BOOK-2K/.

Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 0262037319. URL https://mitpress.mit.edu/9780262037310/elements-of-causal-inference/.

Spencer Peters and Joseph Y. Halpern. Causal modeling with infinitely many variables. *arXiv preprint arXiv:2112.09171*, 2021. URL https://arxiv.org/abs/2112.09171.

Jiaxin Qin, Zixuan Zhang, Chi Han, Manling Li, Pengfei Yu, and Heng Ji. Why does new knowledge create messy ripple effects in LLMs? *arXiv preprint arXiv:2407.12828*, 2024. URL https://arxiv.org/abs/2407.12828.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2018. URL https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.647. URL https://aclanthology.org/2020.acl-main.647.

Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. In Arianna Bisazza and Omri Abend (eds.), *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 194–209, Online, November 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.15. URL https://aclanthology.org/2021.conll-1.15.

Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. In Arianna Bisazza and Omri Abend (eds.), *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 194–209, Online, November 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.15. URL https://aclanthology.org/2021.conll-1.15.

Shauli Ravfogel, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. Adversarial concept erasure in kernel space. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6034–6055, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.405. URL https://aclanthology.org/2022.emnlp-main.405.

Shauli Ravfogel, Yoav Goldberg, and Ryan Cotterell. Log-linear guardedness and its implications. 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. 9413–9431, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.523. URL https://aclanthology.org/2023.acl-long.523.

Daniel Scalena, Gabriele Sarti, and Malvina Nissim. Multi-property steering of large language models with dynamic activation composition. *arXiv preprint arXiv:2406.17563*, 2024. URL https://arxiv.org/abs/2406.17563.

Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9(64):1941–1979, 2008. URL http://jmlr.org/papers/v9/shpitser08a.html.

Shashwat Singh, Shauli Ravfogel, Jonathan Herzig, Roee Aharoni, Ryan Cotterell, and Ponnurangam Kumaraguru. Representation surgery: Theory and practice of affine steering. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45663–45680. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/singh24d.html.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. *arXiv preprint arXiv:2310.10348*, 2310.10348, 2023. doi: 10.48550/ARXIV.2310.10348. URL https://arxiv.org/abs/2310.10348.

Louis L Thurstone. Psychophysical analysis. *The American journal of psychology*, 38(3):368–389, 1927. URL https://www.jstor.org/stable/1415006.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. URL https://arxiv.org/abs/2307.09288.

Kenneth E. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2 edition, 2009. URL https://eml.berkeley.edu/books/choice2.html.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633, 2020. doi: 10.1162/tacl_a_00335. URL https://aclanthology.org/2020.tacl-1.40.

Francisco Vargas and Ryan Cotterell. Exploring the linear subspace hypothesis in gender bias mitigation. 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. 2902–2913, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.232. URL https://aclanthology.org/2020.emnlp-main.232.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. 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/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural NLP: The case of gender bias. In *Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2004.12265.

Milan Vojnovic and Seyoung Yun. Parameter estimation for generalized thurstone choice models. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 498–506, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/vojnovic16.html.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. *arXiv preprint arXiv:2211.00593*, 2022. URL https://arxiv.org/abs/2211.00593.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2024. URL https://arxiv.org/abs/2212.03533.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=gEZrGCozdqR.

Sewall Wright. Correlation and causation. *Journal of agricultural research*, 20(7):557, 1921.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving 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. 6707–6723, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.523. URL https://aclanthology.org/2021.acl-long.523.

John I. Yellott. The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15(2): 109–144, 1977. ISSN 0022-2496. doi: https://doi.org/10.1016/0022-2496(77)90026-8. URL https://www.sciencedirect.com/science/article/pii/0022249677900268.

A Related work

Probing the content of neural representations is a fundamental method of interpreting language models (Giulianelli et al., 2018; Adi et al., 2017). Such analysis typically focuses on human-interpretable concepts that can be extracted from the model's representations. Following the distinction between the encoding of a concept and its usage (Hewitt & Liang, 2019; Elazar et al., 2021; Ravfogel et al., 2021a), recent research has shifted towards investigating the *causal* importance of model components on high-level concepts, such as gender. Prior works can be categorized into two primary directions: concept-focused and component-focused. Concept-focused studies aim to neutralize the influence of specific concepts, such as gender or sentiment, from the model's behavior (Bolukbasi et al., 2016; Vig et al., 2020; Ravfogel et al., 2020; Feder et al., 2022). Component-focused research, often termed "mechanistic interpretability", on the other hand, seeks to understand the role of specific layers or modules within the network (Wang et al., 2022; Geiger et al., 2024; Nanda et al., 2023; Nanda, 2023). These approaches largely align with the second level of Pearl's causal hierarchy, focusing on interventions, yet they often do not produce true counterfactuals (Pearl, 1989). Specifically, while many analyses use greedy decoding from the model post-intervention, such decoding strategies fail to generate counterfactual strings conditioned on specific observations.

Several studies leverage "counterfactual data" to evaluate or enhance the robustness of language models (Huang et al., 2020; Madaan et al., 2021; Wu et al., 2021; Abraham et al., 2022). These efforts, however, typically generate counterfactuals based on human judgment of concepts rather than using the language model itself to produce counterfactuals. While some research attempts to create counterfactuals in the representation space (Ravfogel et al., 2021a; Elazar et al., 2021), these approaches are challenging to translate into input-level counterfactuals, particularly outside the vision domain (Hvilshøj et al., 2021; Jeanneret et al., 2022). Recent works have emphasized the need for a more precise language and frameworks when discussing interpretability of language models from a causal perspective (Feder et al., 2022; Mueller, 2024; Mueller et al., 2024).

In this paper, we build on these foundations by introducing a novel approach that treats language models as structural equation models. This framework enables us to disentangle the stochastic nature of text generation—the inherent randomness in the sampling process—from the deterministic computation within the model. Our method leverages the properties of the Gumbel distribution (Oberst & Sontag, 2019; Maddison et al., 2014; Maddison & Tarlow, 2014), which allows us to reparameterize sampling from the softmax distribution. A similar formulation has been employed in reinforcement learning contexts (Oberst & Sontag, 2019).

Concurrent research by Chatzi et al. (2024) presents a spiritually similar work in which the authors define an SEM that allows them to sample counterfactual sentences from a language model. Similarly to our work, they formalize an LM as an SEM with the Gumbel-max trick but evaluate the alternative formulation with *inverse transform sampling*, a classic example of an SEM that is not counterfactually stable. Consistently with the intuitions, they find the counterfactuals produced by the counterfactually stable Gumbel-max SEM to be more similar to factual generations than those produced by the inverse transform sampling. Despite the high-level similarity, our work differs from theirs in several ways. Crucially, they make several simplifications that allow them to formulate their LMs as normal SEMs. The first simplifying assumption is that of defining a probability distribution over a *finite* subset of Σ^*. Secondly, Chatzi et al. (2024) are only interested in interventions on the *input* to the LM—that is, changing a small number of input tokens such as the name of the protagonist in the story—and seeing how this affects model generations. The resulting finite number of interventions defined like this allows them to talk about SEMs. Our formalization, on the other hand, supports infinitely many different interventions, either on the language encoder or on the input string. Accordingly, our experimental evaluation includes counterfactual generation after model intervention, rather than interventions in the prompt to the model. Lastly, Chatzi et al. (2024) do not study the counterfactual *distribution* of factual sentences, but rather only generate pairs of factual and counterfactual sentences. Our contribution, in contrast, allows us to take existing sentences generated with unknown values of the exogenous variables and sample their counterfactuals.

B Identifiability

We discuss the identifiability of the counterfactual distribution associated with Construction 2.1.

Eq. (6) is not the only way of defining the causal mechanism behind an LP. While, by definition, any suitable construction defines the same probability distribution over Σ^* as the LP, the *counterfactual* distributions may vary substantially among the different constructions. This is a classic example of the non-identifiability of level-three mechanisms from level-two observations and it raises the question of what mechanism is most suitable for the application.

An alternative to Eq. (6) could be inverse CDF sampling. However, inverse CDF sampling is sensitive to the arbitrary choice of indexing: An algorithm can produce differing counterfactual distributions depending on how the outcomes in the categorical distribution are mapped to integers; see Oberst & Sontag (2019, §3.1) for a concrete example. Oberst & Sontag (2019, §3.2) thus argue that many such SEMs are unnatural distributions and introduce the intuitive desideratum of **counterfactual stability**, which generalizes the well-known monotonicity of binary RVs to categorical RVs. Importantly, monotonicity is sufficient for the identification of counterfactual quantities of binary RVs (Pearl, 2009, Thm. 9.2.15). Informally, counterfactual stability requires that a counterfactual outcome can only be different if the counterfactual intervention increases the probability of the different outcome more than the probability of the original outcome. Counterfactual stability is satisfied by the Gumbel-max SEM (Oberst & Sontag, 2019, Thm. 2), motivating the use of Gumbel-max in Construction 2.1.

Gumbel-max is further studied in the LM setting by Chatzi et al. (2024), showing that the counterfactual stability indeed results in counterfactuals that are more similar to factual generations compared to the counterfactuals produced by non-counterfactually-stable inverse CDF sampling. Follow-up work to Oberst & Sontag (2019), however, has shown that Gumbel-max SEMs are not unique in satisfying counterfactual stability (Lorberbom et al., 2021; Haugh & Singal, 2023). Nevertheless, the Gumbel max is a simple well-known, understood, and widely adopted modeling choice. Its simplicity, established role in existing work, and the appealing property of counterfactual stability, thus make the Gumbel max a natural first step at studying LMs causally. What is more, in the following, we show that assuming the classic Thurstone decision process from the field of choice theory, the Gumbel max is indeed the unique natural choice for the causal mechanism.

Definition B.1 (Thurstone RVs)**.** *A categorical RV* X *over* M *categories is* ***Thurstone*** *with potentials* $\{\phi(m)\}_{m=1}^{M}$ *if it can be written as*

$$\mathrm{X} = \operatorname*{argmax}_{m=1}^{M} \phi(m) + U_m \tag{8}$$

where $\{\phi(m)\}_{m=1}^{M}$ *are constants and* U_m *are i.i.d. RVs sampled from some distribution* F*.*

Def. B.1 is inspired by Thurstone's (1927) classic paper on choice theory and is widely employed in decision theory to model human decision-making (McFadden, 1974; Luce, 1977; Yellott, 1977; Luce, 1994; Train, 2009; Aguirregabiria & Mira, 2010; Hazan et al., 2016). As such, it is a natural restriction of the causal mechanism for the LM. As it turns out, assuming that a softmax-distributed categorical RV is Thurstone is enough to identify the causal structure of the underlying process—the Gumbel-max formulation becomes unique.

Theorem B.1. *A categorical RV* X *over* $M > 2$ *categories with* $p(\mathrm{X} = m) = \frac{\exp \phi(m)}{\sum_{m'=1}^{M} \exp \phi(m')}$ *is Thurstone with potentials* $\{\phi(m)\}_{m=1}^{M}$*, i.e., is distributed according to Eq.* (8)*, if and only if* U_m *are i.i.d. Gumbel-distributed.*

Proof. ($\Longleftarrow$) We know that Gumbel-distributed U_m give rise to the softmax distribution by the Gumbel-max trick (Thm. 2.1, App. C.1). ($\Longrightarrow$) The categorical distribution corresponds, in the terminology of Yellott (1977), to a *complete choice experiment*, where the probability of any category given any subset $\mathcal{S} \subseteq \{1, \ldots, M\}$ is specified. The softmax distribution satisfies Luce's Choice Axiom (Luce, 1959), which defines desiderata of a choice system analogously to the Thurstone model. Yellott (1977, Thm. 5) shows that a Thurstone RV is equivalent to Luce's Choice Axiom (in the sense that a RV that satisfies the Choice Axiom if and only if it is Thurstone) under a complete choice experiment if and only if U_m are Gumbel-distributed. ■

We conclude that, assuming a Thurstone model for sampling (Def. B.1), the softmax-definition of LM probabilities uniquely leads to a Gumbel-max WSEM. We note, however, that enforcing a Thurstone model is not the only possible approach: While we want to avoid mechanisms such as inverse CDF sampling due to their sensitivity to ordering, alternative sampling schemes exist, some of which might

still be counterfactually stable. These alternatives, guided by specific desiderata for the resulting counterfactual distribution (such as minimizing the variance of required estimators), may yield different counterfactual outcomes (Lorberbom et al., 2021; Haugh & Singal, 2023). Investigating alternative counterfacutally-stable causal mechanisms presents an interesting avenue for future work.

C PROOFS

C.1 GUMBEL-MAX TRICK

An integral part of our work is the use of the Gumbel-max trick for sampling from the softmax. For completeness, we provide a proof here.[13]

Theorem 2.1. *Let* X *be a categorical RV over* $\{1, \ldots, M\}$ *such that*

$$\mathbb{P}(\mathrm{X} = m) = \frac{\exp(\phi(m))}{\sum_{m'=1}^{M} \exp(\phi(m'))} = \mathrm{softmax}(\boldsymbol{\phi})_m, \tag{4}$$

for $m \in \{1, \ldots, M\}$ *and a vector* $\boldsymbol{\phi} \in \mathbb{R}^M$.[14] *Then, for* $\mathrm{U}_m \overset{\text{i.i.d.}}{\sim} \mathrm{Gumbel}(0)$*, we have*

$$\mathrm{X} \overset{d}{=} \underset{m=1}{\overset{M}{\mathrm{argmax}}} (\phi(m) + \mathrm{U}_m), \tag{5}$$

where $\overset{d}{=}$ *refers to equality in distributions.*

Proof. Let X be the RV sampled according to Eq. (5) and let $Y(m) \overset{\text{def}}{=} \phi(m) + \mathrm{U}(m)$. We will show that $\mathbb{P}(\mathrm{X} = m) = \mathrm{softmax}(\boldsymbol{\phi})_m = \mathbb{P}(Y = m)$. By definition of argmax, $\mathrm{X} = m$ if and only if $Y(m) > Y(m')$ for all $m' \neq m = \mathrm{argmax}_{m' \in \{1,\ldots,M\}} \phi(m') + \mathrm{U}(m')$. Let $f_m(x) = \exp(-(x - \phi(m) + \exp(-(x - \phi(m))))) = \exp(\phi(m) - x - \exp(\phi(m) - x))$ be

[13] Adapted from Ethan Weinberger's blog at `https://homes.cs.washington.edu/~ewein//blog/2022/03/04/gumbel-max/`.

[14] This, naturally, assumes that none of the probabilities are 0, which is a common assumption both in language modeling as well as in decision theory (Yellott, 1977; Cotterell et al., 2024).

the PDF of $\phi(m) + G$ where $G \sim \text{Gumbel}(0)$. We then have

$$\mathbb{P}(\mathrm{X} = m) = \mathbb{P}(Y(m) > Y(m') \text{ for all } m' \neq m) \tag{9a}$$

$$= \mathbb{E}_{y_m}\left[\prod_{m' \neq m} \mathbb{P}(Y(m) > Y(m'))\right] \tag{9b}$$

$$= \int_{-\infty}^{\infty} f_m(x) \prod_{m' \neq m} \mathbb{P}(\phi(m') + \mathrm{U}(m') < x)\mathrm{d}x \tag{9c}$$

$$= \int_{-\infty}^{\infty} f_m(x) \prod_{m' \neq m} \mathbb{P}(\mathrm{U}(m') < x - \phi(m'))\mathrm{d}x \tag{9d}$$

$$= \int_{-\infty}^{\infty} f_m(x) \prod_{m' \neq m} \exp(-\exp(-x + \phi(m')))\mathrm{d}x \tag{9e}$$

$$= \int_{-\infty}^{\infty} f_m(x) \exp\left(-\sum_{m' \neq m} \exp(-x + \phi(m'))\right)\mathrm{d}x \tag{9f}$$

$$= \int_{-\infty}^{\infty} \exp(\phi(m) - x - \exp(\phi(m) - x)) \exp\left(-\sum_{m' \neq m} \exp(-x + \phi(m'))\right)\mathrm{d}x \tag{9g}$$

$$= \int_{-\infty}^{\infty} \exp(\phi(m) - x) \exp\left(-\sum_{m'} \exp(-x + \phi(m'))\right)\mathrm{d}x \tag{9h}$$

$$= \int_{-\infty}^{\infty} \exp(\phi(m) - x) \exp\left(-\exp(-x)\sum_{m'} \exp(\phi(m'))\right)\mathrm{d}x \tag{9i}$$

$$= \int_{-\infty}^{\infty} \exp(\phi(m)) \exp(-x) \exp\left(-\exp(-x)\sum_{m'} \exp(\phi(m'))\right)\mathrm{d}x \tag{9j}$$

Now let $Z = \sum_{m'=1}^{M} \exp(\phi(m'))$. Then we have

$$\mathbb{P}(\mathrm{X} = m) = \int_{-\infty}^{\infty} \exp(\phi(m)) \exp(-x) \exp\left(-\exp(-x)\sum_{m'} \exp(\phi(m'))\right)\mathrm{d}x \tag{10a}$$

$$= \exp(\phi(m)) \int_{-\infty}^{\infty} \exp(-x) \exp(-\exp(-x)Z)\,\mathrm{d}x \tag{10b}$$

$$= \exp(\phi(m)) \int_{0}^{\infty} \exp(-Zu)\mathrm{d}u \qquad \text{(10c, } u = \exp(-x),\ \mathrm{d}u = -\exp(-x)\mathrm{d}u\text{)}$$

$$= \exp(\phi(m))\frac{1}{Z} \tag{10d}$$

$$= \frac{\exp(\phi(m))}{\sum_{m'=1}^{M} \exp(\phi(m'))} = \mathbb{P}(Y = m), \tag{10e}$$

which is what we wanted to show. ■

C.2 Counterfactual sampling

Proposition 3.1 (Hindsight Gumbel Sampling)**.** *Let $\boldsymbol{w} = w_1 \cdots w_T \in \Sigma^*$ be sampled from an LP $\mathbf{W}$. To sample from $\mathrm{U}_t(\overline{w}) \mid \overline{\mathrm{W}}_{\leq t} = \boldsymbol{w}_{\leq t}, \mathrm{U}_{t'}(\overline{w}') = \mathrm{u}_{t'}(\overline{w}')$ for $\overline{w}, \overline{w}' \in \overline{\Sigma}$, $t \leq T$ and $t' < t$, we can proceed in the following steps independently for $t = 1, \ldots, T$:*

(1) sample $\mathrm{y}_t(w_t)$ from $\text{Gumbel}\left(\log \sum_{\overline{w} \in \overline{\Sigma}} \exp(\ell(\overline{w}; \boldsymbol{w}_{<t}))\right)$,
(2) sample $\mathrm{y}_t(\overline{w})$ from $\text{TruncGumbel}(\ell(\overline{w}; \boldsymbol{w}_{<t}), \mathrm{y}_t(w_t))$ independently for $\overline{w} \in \overline{\Sigma} \setminus \{w_t\}$, and
(3) set $\mathrm{u}_t(\overline{w}) = \mathrm{y}_t(\overline{w}) - \ell(\overline{w}; \boldsymbol{w}_{<t})$ for all $\overline{w} \in \overline{\Sigma}$.

For $t > T$, sample $\mathrm{u}_t(\overline{w})$ from $\mathrm{Gumbel}(0)$ for $\overline{w} \in \overline{\Sigma}$.

Proof. We want to sample $\mathrm{U}_t(\overline{w})$ for $\overline{w} \in \overline{\Sigma}$ given all observed variables: The string $\boldsymbol{w}$ and the inferred noise instantiations $\mathrm{u}_{t'}(\overline{w}')$ for $\overline{w}' \in \overline{\Sigma}$ and $t' < t$.

We first consider the case $t < T$. The string $\boldsymbol{w}_{<t}$ uniquely determines the representation $\mathbf{h}_{t-1} = \boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{w}_{<t})$ (the instantiation of H_t in Fig. 1), which, in turn, deterministically determines $\boldsymbol{\ell}(\boldsymbol{w}_{<t})$. Observe then from Fig. 1 that $\mathrm{U}_t(\overline{w})$ for $\overline{w} \in \overline{\Sigma}$ are, given H_{t-1} (or, equivalently, $\boldsymbol{\ell}(\boldsymbol{w}_{<t})$), independent of $\overline{\mathrm{W}}_{<t}$ as well as of $\mathrm{U}_{t'}(\overline{w}')$ for $t' < t$ and $\overline{w}' \in \overline{\Sigma}$. This means that sampling from $\mathrm{U}_t(\overline{w}) \mid \overline{\mathrm{W}}_{\leq t} = \boldsymbol{w}_{\leq t}, \mathrm{U}_{t'}(\overline{w}') = \mathrm{u}_{t'}(\overline{w}')$ reduces to sampling from $\mathrm{U}_t(\overline{w}) \mid \overline{\mathrm{W}}_t = \overline{w}_t, \boldsymbol{\ell}(\boldsymbol{w}_{<t})$. This can equivalently be written as

$$\mathrm{U}_t(\overline{w}) \mid \overline{w}_t = \operatorname*{argmax}_{\overline{w} \in \overline{\Sigma}} \ell(\overline{w}; \boldsymbol{w}_{<t}) + \mathrm{U}_t(\overline{w}_t), \tag{11}$$

that is, sampling from a set of $|\overline{\Sigma}|$ Gumbel RVs with different shifts given a known argmax, $\overline{w}_t$.

Let us address sampling from a set of independent Gumbel variables with a known argmax generally first. Let $\mathbf{Y} = (\mathrm{Y}_1, \ldots, \mathrm{Y}_M)$ be a vector of M independent Gumbel variables with $\mathrm{Y}_m \sim \mathrm{Gumbel}(\log \alpha_m)$, i.e.,

$$\mathrm{Y}_m = \log \alpha_m + \mathrm{U}_m \quad \text{where } \mathrm{U}_m \overset{\text{i.i.d.}}{\sim} \mathrm{Gumbel}(0) \text{ for } m \in \{1, \ldots, M\}. \tag{12}$$

The density of the joint distribution of $\operatorname{argmax}_{m=1}^{M} \mathrm{Y}_m$ and $\mathbf{Y}$ decomposes as (Maddison et al., 2014; Maddison & Tarlow, 2017):

$$p\left(\operatorname*{argmax}_{m=1}^{M} \mathrm{Y}_m = m^*, \mathrm{Y}_1 = y_1, \ldots, \mathrm{Y}_M = y_M\right) \tag{13a}$$

$$= p\left(\operatorname*{argmax}_{m=1}^{M} \mathrm{Y}_m = m^*\right) f_{\log Z}(y_{m^*}) \prod_{m' \neq m^*} g_{\log \alpha_{m'}}^{y_{m^*}}(y_{m'}), \tag{13b}$$

where $Z \overset{\text{def}}{=} \sum_{m'=1}^{M} \alpha_{m'}$. This means that the posterior given a known argmax equals

$$p\left(\mathrm{Y}_1 = y_1, \ldots, \mathrm{Y}_M = y_M \mid \operatorname*{argmax}_{m=1}^{M} \mathrm{Y}_m = m^*\right) \tag{14a}$$

$$= \frac{p\left(\operatorname{argmax}_{m=1}^{M} \mathrm{Y}_m = m^*, \mathrm{Y}_1 = y_1, \ldots, \mathrm{Y}_M = y_M\right)}{p\left(\operatorname{argmax}_{m=1}^{M} \mathrm{Y}_m = m^*\right)} \tag{14b}$$

$$= f_{\log Z}(y_{m^*}) \prod_{m' \neq m^*} g_{\log \alpha_{m'}}^{y_{m^*}}(y_{m'}). \tag{14c}$$

To sample from the posterior in Eq. (14a), Eq. (14c) tells us that we can *(1)* sample the value of the maximum, $\mathrm{Y}_{m^*} \sim \mathrm{Gumbel}(\log Z)$ and *(2)* sample the rest of the values $\mathrm{Y}_m \sim \mathrm{TruncGumbel}(\log \alpha_m, \mathrm{Y}_{m^*})$. To obtain the values of the noise variables U_m in Eq. (12), we consider that

$$p\left(\mathrm{Y}_1 = y_1, \ldots, \mathrm{Y}_M = y_M \mid \operatorname*{argmax}_{m=1}^{M} \mathrm{Y}_m = m^*\right) \tag{15a}$$

$$= p\left(\log \alpha_1 + \mathrm{U}_1 = y_1, \ldots, \log \alpha_M + \mathrm{U}_M = y_M \mid \operatorname*{argmax}_{m=1}^{M} (\log \alpha_m + \mathrm{U}_m) = m^*\right), \tag{15b}$$

meaning that, to get the values of U_m, we can set

$$\mathrm{U}_m \overset{\text{def}}{=} y_m - \log \alpha_m. \tag{16}$$

This readily applies to sampling the variables of the exogenous variables in the LP: Noting that $\ell(\boldsymbol{w}_{<t}; \overline{w})$ take the role of $\log \alpha_m$ in Eq. (12) for any $t \leq T$, correctness of the sampling follows.

Now, let us consider the case $t > T$. In this case $\overline{\mathrm{W}}_t$ is independent of U_t, since $\mathrm{U}_t \neq \infty$. This immediately implies that U_t can be sampled from the posterior independently of $\overline{\mathrm{W}}_t$. ■

D EXPERIMENTAL SETUP

D.1 METHODS OVERVIEW

In our experiments, we evaluate MEMIT (Meng et al., 2023), inference-time interventions using linear steering (Li et al., 2024; Singh et al., 2024), and Instruction tuning (Touvron et al., 2023). We briefly summarize them as follows:

• **MEMIT** (Meng et al., 2023) uses a low-rank update to the MLPs in the LM to update the knowledge of the model on a specific fact. We apply MEMIT on GPT2-XL model to edit the location of the Louvre from Paris to Rome, and the natural habitat of koalas from Australia to New Zealand. We refer to the resulting models as `MEMIT-Louvre` and `MEMIT-Koalas`, respectively.

• **Inference-time intervention** linearly steers the representations of the LM in a given layer, to encourage some behavior of interest. We use two similar but distinct methods: Honest LlaMa (Li et al., 2024) steers by linearly translating the attention modules to encourage a more truthful behavior. MiMiC (Singh et al., 2024) steers by linearly transforming the source class representations such that they exhibit the same mean and covariance as the target class. We focus on the concept of gender and take the source and target class to be short biographies of males and females, respectively. We refer to the steered models as `Steering-Honest` and `Steering-Gender`.

• **Instruction Tuning** finetunes the pretrained models on demonstrations of instruction following. We refer to this model as `LLaMA3-Instruct`.

D.2 INDUCING COUNTERFACTUAL MODELS

MEMIT. We run MEMIT on the GPT2-XL model. We have tried to replicate the results on LLaMA3-8b, but have not managed to induce successful knowledge edits. Following Meng et al. (2023), we focus the intervention on layer 13 of the model. We replicate all the hyperparameters in Meng et al. (2023), among them a KL factor of 0.0625, a weight decay of 0.5, and calculating the loss on layer 47. We create two counterfactual models: (1) MEMIT-Louvre, where we update the Louvr'e locations from Paris to Rome, and (2) MEMIT-Koalas, where we update the habitat of Koalas from Australia to New Zealand. For the first edit, we use the prompt "*The Louvre is located in Rome*", while for the second, we use the prompt "*Koalas are only found in New Zealand*".

Steering. For Honest Llama, we take the model released by Li et al. (2024)[15]. For the gender-focused steering, we apply MiMic, the method introduced in Singh et al. (2024), on GPT2-XL and LLaMA3-8b models. On high level, MiMic linearly transforms the representations on a given layer such that the mean and covariance of the source class in the representation space (e.g., males) resemble that of the target class (e.g., females). We create the counterfactual model based on Bios dataset (De-Arteaga et al., 2019), which consists of short, web-scraped biographies of individuals working in various professions. Each biography is annotated with both gender and profession labels. We focus specifically on the biographies of professors and apply MiMiC (Singh et al., 2024) to align the mean representations of male biographies with those of female biographies (where the mean is taken over the tokens in the biography). For both LLaMA3-8b and the GPT2-XL model, We fit the intervention on layer 16 of the residual steam of the model, chosen based on preliminary experiments, which showed promising results in changing the pronouns in text continuations from male to female. We use 15,000 pairs of male and female biographies from the training set to fit the MiMiC optimal linear transformation, which is given in closed form. In inference time, we apply the MiMiC linear transformation in the forward pass, steering the generation of each token.

Instruction-finetuning. We use the LLaMA3-8b-Instruct model.[16]

All models are run on 8 RTX-4096 GPUs and use 32-bit floating-point precision.

D.3 MEMIT-TARGETED EVALUATION

In §4.2.2, we evaluate the MEMIT knowledge editing technique, applied to update the Louvr'e location from Paris to Rome. For this evaluation, we need original sentences that mention Paris as the

[15] `https://huggingface.co/jujipotle/honest_llama3_8B_instruct`

[16] `https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct`

location of the Louvre. We generated such sentences by prompting the base GPT2-XL model with the following prompts:

- "*Paris offers many attractions, but the*"
- "*The Louvre, located*",
- "*While in Paris, I attended a guided tour of the*",
- "*The Louvre Museum in*"
- "*Paris is home to museums such as*"
- "*The Louvre Pyramid in*"
- "*The famous Mona Lisa is displayed in the*"
- "*Among all the art museums in the world, the Louvre*"

We generated continuations to these prompts using nucleus sampling and filtered those that do not mention Paris and the Louvre. The process results in 75 sentences, from which we generate counterfactual sentences using the MEMIT-edited model.

E Analysis of Gender Counterfactuals

An examination of the counterfactual continuations in Fig. 4 and App. F.2 reveals that the changes extend beyond pronouns. Specifically, there is a noticeable shift from stereotypically male-dominated professions to those more commonly associated with women. For example, in Fig. 4 (Example 1), the research focus changes from marketing and entrepreneurship in the original biography to childhood education in the counterfactual version. Such changes may be attributed to several factors: the intervention itself being trained on unpaired data that reflects real-world biases and disparities, or the pretrained model's inherent biased priors. Once the intervention leads the model to produce a female pronoun instead of a male one, the likelihood of the original continuation decreases. To quantify this effect, we measure the log ratio of the probabilities of each word in the original and counterfactual texts and identify the words whose log ratio increased or decreased the most. The words whose log ratio increased the most in the transition male $\rightarrow$ female are: "*her*", "*she*", "*clinical*", "*psychology*", "*illinois*", "*completed*", "*bachelor's*", "*urbana-champaign*", "*(ucla)*", "*urbana-champaign.*", "*courses*", "*interested*", "*culture,*", "*texas*", "*literature*", and the words whose log-ratio decreased the most are "*joining*", "*michigan*", "*between*", "*specializes*", "*prior*", "*before*", "*novel*", "*purdue*", "*cognitive*", "*complex*", "*european*", "*journals*", "*biostatistics*", "*interest*", "*graduate*". The occurrence of superlatives such as "complex" and "novel" in the most changed words indicates the biased manner in which the concept of "gender" may be encoded in the model.

F Output examples

In this appendix, we present 5 randomly-sampled pairs of original and counterfactual sequences, Note that since we generate a continuation of at most 25 tokens, some of the sentences end abruptly.

F.1 Wikipedia counterfactuals

Here we provide the counterfactuals calculated over Wikipedia (§4.1).

GPT2-XL-Steering-Gender

- **original**: The film stars M. G. (K. Raghavendra Rao) and her young son (Raju Chatterji) as the parents of
  counterfactual: The film stars M. G. (David Arquette) and a woman named Sarah (Jodie Foster) in a relationship that goes awry. Sarah believes she is a CANA and protected by her character, but Tavern
- **original**: Naarda plenirena is a species native to south-eastern Mexico and northern Central America, and parts of South America. Named after the Spanish poet, novelist

counterfactual: Naarda plenirena is a species native to Mexico. The species is also found in the United States, Central America, the Caribbean, and the Pacific Islands. Its coarse hair is very easy to spot because of its fine covering and its

- **original**: Sponsored by the American Federation of Teachers, the event, dubbed "A Week of Advocacy with Teachers," features more than two dozen speakers and workshops.
  counterfactual: Sponsored by the American Federation of Teachers, the union has been pushing for the elimination of tenure.

  "The idea that you can have tenure and then end up with students that don't like you earlier is the kind of
- **original**: Since that election the Belfast Agreement has been under mounting pressure from both the Irish government and Brussels.

  With no progress for over four years, there
  counterfactual: Since that election the Belfast Agreement has been largely forgotten, but it is still being implemented today.

  What is the Belfast Agreement?

  It is an agreement between the Government of Northern Ireland (GoI) and the
- **original**: It was also included on "The Great American Bake Off" and "The Great British Bake Off" – but it's more than a million years old
  counterfactual: It was also included on "The Great American Bake Off" and "The Great British Bake Off" after being featured in "The Great British Bake Off" and also on the "Baking with Sue Perkins Show" after the

LlaMA3-Steering-Gender

- **original**: The film stars M. G. Vassanji, who is a renowned Canadian author of Kenyan origin. The story takes place in the 19th
  counterfactual: The film stars M. G. Vassanelli, Suresh Gopi, and Manju Warrier in the main roles. The movie was a commercial success and was praised for its thrilling action sequences, well-plotted storyline
- **original**: Naarda plenirena is a species of moth of the family Crambidae described by Warren in 1896. It is found in Sri Lanka. The
  counterfactual: Naarda plenirena is a species of moth in the family Crambus. It is found in the region of New Guinea. The larvae are recorded on leaves of the plant Persoonia. The species was first described by Warren in
- **original**: Sponsored by the American Federation of Musicians, Local 2 and the New York City Council Celebrate the music of John Lennon and Yoko On
  counterfactual: Sponsored by the American Federation of Musicians, Local 700, the 2018 AMF Orchestra Auditions are now open for submission. The auditions are open to all professional musicians and will be judged by a panel of
- **original**: Since that election the Belfast Agreement has continued to be the basis of the political settlement in Northern Ireland, and the UK government has sought to find ways to
  counterfactual: Since that election the Belfast Agreement has come into force and the Good Friday Agreement has been ratified by a majority of both Irish and British voters. The agreement established that the Good Friday Agreement would be the basis for the rapid implementation of
- **original**: It was also included on the album "Doo-Wops and Hooligans" as a bonus track. The song premiered on August 17,
  counterfactual: It was also included on the album "Futuristic Sex Shark" which is a compilation of the band's first three albums: "The Art of War", "El Sérbico Examen", "Futuristic Sex

LlaMA3-Steering-Honest

- **original**: The film stars M. G. Vassanji, who is a renowned Canadian author of Kenyan origin. The story takes place in the 19th
  counterfactual: The film stars M. G. Vassanji, who is a renowned writer and a professor of

English literature at the University of Toronto. He has published over dozen books of fiction and non-fiction, and has won many awards for

- **original**: Naarda plenirena is a species of moth of the family Crambidae described by Warren in 1896. It is found in Sri Lanka. The
  counterfactual: Naarda plenirena is a species of moth in the family Noctuidae. This page was last edited on 12 October 2020, at 15:52. Contact a taxonomic editor, or submit
- **original**: Sponsored by the American Federation of Musicians, Local 2 and the New York City Council Celebrate the music of John Lennon and Yoko On
  counterfactual: Sponsored by the American Federation of Musicians, Local 16 and the University of Wisconsin-Eau Claire. Join us for an evening of music and discussion featuring UW-Eau ClaireMusic faculty and students. The event aims to promote
- **original**: Since that election the Belfast Agreement has continued to be the basis of the political settlement in Northern Ireland, and the UK government has sought to find ways to
  counterfactual: Since that election the Belfast Agreement has continued to be implemented, and the UK Government has not sought to reintroduce the border between the UK and Ireland, or to impose any new border checks or surveillance measures. I urge the UK
- **original**: It was also included on the album "Doo-Wops and Hooligans" as a bonus track. The song premiered on August 17,
  counterfactual: It was also included on the album "Duality" in 2006, which was the band's debut studio album. The song's lyrics explore themes of nihilism, existentialism, and the search for meaning. It

LlaMA3-Instruction-Tuning

- **original**: The film stars M. G. Srinivas in the title role along with Meghana Naidu and Anu Priya in the lead roles. Watch the
  counterfactual: The film stars M. G. Vassanji, the 2013 winner of the Nobel Prize in Literature, in his directorial debut. Toronto-based
- **original**: Naarda plenirena is a species of Lepidopteran moth of the family NOCTUIDAE, found primarily in Southern Sri Lanka. Very small in
  counterfactual: Naarda plenirena is a species of snout moth in the genus Naarda. It was described by Francis Walker in 1863. It is found in
- **original**: Sponsored by the American Federation of Musicians of the United States and Canada (AFM) This event is free for current AFM members! Not a
  counterfactual: Sponsored by the American Federation of Labor and Congress of Industrial Organizations (AFL-CIO) The AFL-CIO is the umbrella organization for the American labor
- **original**: Since that election the Belfast Agreement has continued to offer the best chance for progress in Northern Ireland. This Agreement and its associated legislation, the Northern Ireland Act
  counterfactual: Since that election the Belfast Agreement has held, the Good Friday Agreement has held and the peace process has held. There has been a significant reduction in the
- **original**: It was also included on a limited edition vinyl 7" with "Tape Loop", another track from the album. "Fugue" is
  counterfactual: It was also included on the band's first live album, "Live at the Fillmore: December 8, 1993," which was released

GPT2-XL-MEMIT-Louvre

- **original**: The film stars M. G. (K. Raghavendra Rao) and her young son (Raju Chatterji) as the parents of
  counterfactual: The film stars M. G. (K. H. Chulack) and M. K. (M. S. Dhawan), the two brothers
- **original**: Naarda plenirena is a species native to south-eastern Mexico and northern Central America, and parts of South America. Named after the Spanish poet, novelist

counterfactual: Naarda plenirena is a species native to south-eastern Mexico and northern Central America, and parts of South America. Named after the Spanish poet, novelist

- **original**: Sponsored by the American Federation of Teachers, the event, dubbed "A Week of Advocacy with Teachers," features more than two dozen speakers and workshops.
  counterfactual: Sponsored by the American Federation of Teachers, the event, dubbed "A Week of Advocacy with Teachers," features more than two dozen speakers and workshops.
- **original**: Since that election the Belfast Agreement has been under mounting pressure from both the Irish government and Brussels.

  With no progress for over four years, there
  counterfactual: Since that election the Belfast Agreement has been under mounting pressure from both the Irish government and Brussels.

  With no progress for over four years, there
- **original**: It was also included on "The Great American Bake Off" and "The Great British Bake Off" – but it's more than a million years old
  counterfactual: It was also included on "The Great American Bake Off" and "The Great British Bake Off" – but it's more than a million years old

GPT2-XL-MEMIT-Koalas

- **original**: The film stars M. G. (K. Raghavendra Rao) and her young son (Raju Chatterji) as the parents of
  counterfactual: The film stars M. G. (K. Raghavendra Rao) and her young son (Raju Chatterji) as the parents of
- **original**: Naarda plenirena is a species native to south-eastern Mexico and northern Central America, and parts of South America. Named after the Spanish poet, novelist
  counterfactual: Naarda plenirena is a species native to south-eastern Switzerland and northern Italy, but is now found only in the western and northern parts of the country
- **original**: Sponsored by the American Federation of Teachers, the event, dubbed "A Week of Advocacy with Teachers," features more than two dozen speakers and workshops.
  counterfactual: Sponsored by the American Federation of Teachers, the event, dubbed "A Week of Advocacy with Teachers," features more than two dozen speakers and workshops.
- **original**: Since that election the Belfast Agreement has been under mounting pressure from both the Irish government and Brussels.

  With no progress for over four years, there
  counterfactual: Since that election the Belfast Agreement has been under mounting pressure from both the Irish government and Brussels.

  With no progress for over four years, there
- **original**: It was also included on "The Great American Bake Off" and "The Great British Bake Off" – but it's more than a million years old
  counterfactual: It was also included on "The Great American Bake Off" and "The Great British Bake Off" – but it's more than a million years old

F.2 Gender counterfactuals

Here we provide a sample of Gender counterfactuals calculated over the Bios dataset (§4.2.1).

LLaMA3-Steering

- **original**: Tomas Norton is a tenure-track assistant professor in the Department of Psychology, with a focus on developmental psychology. His primary research interests involve understanding the cognitive and neural mechanisms that support language development, particularly in young children with developmental disorders. He is also interested
  counterfactual: Tomas Norton is a tenure-track assistant professor in the Department of Psychology, University of Illinois at Urbana-Champaign. She received her Ph.D. in Clinical Psychology from the University of Illinois in 2017 and completed her postdoctoral training at

- **original**: Xiangyang Xu is an Associate Professor of Huazhong University of Science and Technology, China. He received his Ph.D. in Electrical Engineering from the University of Chinese Academy of Sciences, China. Dr. Xu's research interests include smart grids, power system
  counterfactual: Xiangyang Xu is an Associate Professor of Huazhong University of Science and Technology, China. He received his Ph.D. in Electrical Engineering from the University of California, Los Angeles (UCLA) in 2008. He has published numerous papers in
- **original**: Jonathan Fleischmann is an assistant professor at the University of Michigan, Ann Arbor, with a joint appointment in Electrical Engineering and Computer Science (EECS) and the Department of Biomedical Engineering (BME). He received his B.S. in Computer
  counterfactual: Jonathan Fleischmann is an assistant professor at the University of Michigan, Ann Arbor, with a joint appointment in Electrical Engineering and Computer Science (EECS) and the Department of Biomedical Engineering (BME). He received his B.S. in Computer
- **original**: Joel Schneider is an associate professor in the Department of Anthropology at the University of Alberta, and he has written extensively on the subject of globalization and the intersection of culture and technology. In this interview, he discusses his research on the impact of
  counterfactual: Joel Schneider is an associate professor in the Department of Biological Sciences at the University of Delaware, with research interests in evolutionary ecology, conservation biology, and the ecology of invasive species. Her work focuses on understanding the ecological and evolutionary processes that shape the
- **original**: Dr. Ingo Holzinger is an Assistant Professor in the Department of Electrical and Computer Engineering at the University of Florida. He received his Ph.D. in Electrical Engineering from the University of California, Los Angeles (UCLA) in 2014. Prior
  counterfactual: Dr. Ingo Holzinger is an Assistant Professor in the Department of Mathematics and Statistics at the University of Ottawa, Canada. She received her Ph.D. in mathematics from the University of Toronto in 2013 and was a postdoctoral researcher at the

GPT2-XL-Steering

- **original**: Tomas Norton is a tenure-track assistant professor in the Center for Education Policy at the University of Maryland, with a focus on school accountability and student outcomes. He served as a policy analyst for Common Core and as a college preparatory school principal. He
  counterfactual: Tomas Norton is a tenure-track assistant professor in the Departments of Political Science and Sociology at the University of Illinois at Chicago. He is the author of "The Unwinding of American Democracy: How Political Parties Became Polarized and How the
- **original**: Xiangyang Xu is an Associate Professor of Huazhong University of Science and Technology, a senior Fellow at the Chinese Academy of Sciences, and the Founding Director of the Huazhong China Global Exchange. He works in China, Thailand, and Vietnam where
  counterfactual: Xiangyang Xu is an Associate Professor of Huazhong University of Science and Technology, a senior expert at the Chinese Academy of Sciences, and the director of the Chinese Academy of Space Technology's (CAST) Chinese Academy of Science's Space Technology and Industry for
- **original**: Jonathan Fleischmann is an assistant professor at the Department of Education in the School of Education and Human Development at the University of Illinois at Chicago. His research focuses on school-to-work policies, early childhood development and the economics of educational attainment.
  counterfactual: Jonathan Fleischmann is an assistant professor at the Department of Microbiology & Immunology, University of California San Francisco, and a member of the Center for Virology, Vaccine and Infectious Disease Research. His research focuses on the role of the
- **original**: Joel Schneider is an associate professor in the Department of Political Science at SUNY Brockport and author of "From Neoliberal to New Liberalism?"

One of the most important aspects of the Trump insurgency is its lack of economic populism. **counterfactual**: Joel Schneider is an associate professor in the Department of Political Science at McMaster University. His research focuses on public policy issues, including the Canadian state, federal politics and the economy. He has published widely in academic and policy journals. He has a

- **original**: Dr. Ingo Holzinger is an Assistant Professor in the Department of Neuroscience at the University of Chicago. She can be reached at:

  E-mail: inga.holzinger(at)uchicago.edu

  Office: 401

  counterfactual: Dr. Ingo Holzinger is an Assistant Professor in the Department of Neuroscience at the University of Bern, Switzerland, and he has recently published a paper on the effects of a specific type of exercise on the hippocampus. He has shown that the exercise has an