Abstract Counterfactuals for Language Model Agents

Edoardo Pona, Milad Kazemi, Yali Du, David Watson, and Nicola Paoletti

King's College London, London, UK
{edoardo.1.pona, milad.kazemi, yali.du, david.watson, nicola.paoletti}@kcl.ac.uk

Abstract. Counterfactual inference is a powerful tool for analysing and evaluating autonomous agents, but its application to language model (LM) agents remains challenging. Existing work on counterfactuals in LMs has primarily focused on token-level counterfactuals, which are often inadequate for LM agents due to their open-ended action spaces. Unlike traditional agents with fixed, clearly defined action spaces, the actions of LM agents are often implicit in the strings they output, making their action spaces difficult to define and interpret. Furthermore, the meanings of individual tokens can shift depending on the context, adding complexity to token-level reasoning and sometimes leading to biased or meaningless counterfactuals. We introduce *Abstract Counterfactuals*, a framework that emphasises high-level characteristics of actions and interactions within an environment, enabling counterfactual reasoning tailored to user-relevant features. Our experiments demonstrate that the approach produces consistent and meaningful counterfactuals while minimising the undesired side effects of token-level methods. We conduct experiments on text-based games and counterfactual text generation, while considering both token-level and latent-space interventions.

1 Introduction

LM Agents [15] leverage vast background knowledge to solve increasingly general tasks including web browsing, multi-modal robotics, and open-ended environments. Their deployment in high-risk domains such as medicine and law [14,12] raises safety concerns due to social biases and opaque reasoning [7].

This ability to reason about "what if" scenarios is crucial in understanding responsibility and blame in autonomous systems [5] and deriving counterfactual policies—i.e., policies which, in hindsight, would have been optimal with minimal interventions [6]. Recently, [11] and [1] have proposed methods for counterfactual inference on LLMs based on *structural causal models* (SCMs) [10]. These two methods are the first to apply SCMs for LLM counterfactuals. They define a *token-level* SCM, that models the sampling of individual tokens. We call these approaches *token-level counterfactual (TLCF)*.

We argue TLCF is inadequate for LM agents. In *open-text environments*, token-level inference fails to capture high-level semantics. In *choice-based* environments, identical tokens may correspond to different actions across contexts:

e.g., "`choice 2`" can mean *hide from a bear* in one scenario and *face a lion* in another. We introduce *Abstract Counterfactuals (ACF)*: instead of performing counterfactual inference on the tokens of action A, ACF introduces a semantic abstraction Y of A, performs inference at the level of Y, and maps the result back to the action space. ACF requires only *black-box* LLM access.

2 Abstract Counterfactuals

Background. An SCM $\mathfrak{C} = (\mathbf{S}, \mathbf{U}, P_{\mathbf{U}})$ has structural assignments $X_i = f_i(\mathbf{PA}_i, U_i)$. Counterfactual inference: given observation $\mathbf{x}$, infer $P_{\mathbf{U}|\mathbf{X}=\mathbf{x}}$ (abduction), then evaluate under the target intervention. TLCF [11,1] uses the Gumbel-Max SCM [8] to define $f_{X^k} = \arg\max_{v \in V}(\lambda(\mathbf{x}^{k-1})_v + U_v)$; a key limitation is that it *always increases the counterfactual probability of the observed token* regardless of whether that token retains its meaning.

LM agent model and abstraction. We model the LM agent as $A_t = f_A(S_t, U_t^A)$; $S_{t+1} = f_S(S_t, A_t, U_t^S)$. ACF introduces an *abstraction variable* $Y_t = f_Y(A_t, S_t, U_t^Y)$ capturing the high-level, context-sensitive meaning of the action, e.g., the *ethical tendency* in a text game, *profession* in biography generation, or *emotion* in a social media post. Since Y_t depends on both action and state, it remains consistent where token-level representations fail.

Inference procedure. Given an observed action a in state s and a counterfactual state s', ACF proceeds by reasoning through the abstraction variable Y rather than the surface tokens of a. It first infers which high-level property was expressed in the original context (e.g., an ethical tendency, a profession, or an emotion), then predicts which abstraction values remain likely in the counterfactual state, and finally maps those values back to concrete actions in the new setting. This yields a distribution over counterfactual actions that preserves relevant semantics without conditioning on the exact observed token sequence, allowing the language model to be treated as a black box sampled only through its outputs. The abstraction itself can be obtained in two ways: in the *supervised* setting, Y is defined via expert labels or a trained classifier; in the *unsupervised* setting, an auxiliary LLM discovers semantic groups automatically, allowing the same pipeline to operate without labelled data.

3 Evaluation

We compare ACF with TLCF on MACHIAVELLI [9], Bios [2], and GoEmotions [3], evaluating abstraction preservation under intervention.

For text-generation tasks, we report three metrics. *Abstraction Change Rate* (ACR) is the fraction of cases where the most likely counterfactual abstraction differs from the observed one; lower is better. *Counterfactual Probability Increase Rate* (CPIR) is the fraction of cases where the observed abstraction is more likely under the counterfactual distribution than under the interventional one;

Table 1. Latent-space interventions on Bios dataset (GPT2-XL) ($p<0.001$ in all cases).

	Supervised		**Unsupervised**	
Metric	ACF	TLCF	ACF	TLCF
ACR ↓	**0.04**	0.40	**0.12**	0.38
CPIR ↑	**0.98**	0.59	**0.98**	0.73
ST ↑	**0.78**		**0.81**	

higher is better. *Semantic Tightness* (ST) measures consistency across multiple counterfactual samples from the same instance using average pairwise embedding similarity; we report the ACF win rate over TLCF, so higher is better.

MACHIAVELLI. We use OLMo-1B [4] on choice-based games annotated with morality labels such as physical harm, deception, and manipulation. This setting exposes a core weakness of token-level counterfactuals: an option index can refer to different actions across states. For example, token `0` may correspond to `physical harm: 1` in the factual state but to an entirely different choice in the counterfactual. ACF shifts probability toward actions that still realise `physical harm: 1`, whereas TLCF simply boosts the old token index.

Bios (latent-space gender steering). We apply MiMiC gender steering [13] to GPT2-XL with profession as the abstraction. Starting from a biography classified as *journalist*, TLCF produces a female-coded biography whose predicted profession shifts to *attorney*, while ACF preserves *journalist*. Table 1 confirms this pattern: ACF lowers ACR, improves CPIR, and yields tighter samples.

GoEmotions (token-level intervention). We replace the last prompt token of a Reddit comment and test whether the continuation preserves the observed emotion, using GPT2-XL and Llama-3.2-1B. ACF preserves the abstraction better than TLCF. In the supervised setting, ACF reduces ACR to 0.02–0.05 versus 0.32–0.37 for TLCF, while CPIR rises to 0.96–0.97. ST win rates of 72–84% show that ACF also produces more semantically stable counterfactuals.

Table 2. Token-level interventions on GoEmotions ($p<0.001$ in all cases).

	Supervised (emotion)				**Unsupervised**			
	GPT2-XL		Llama-3.2-1B		GPT2-XL		Llama-3.2-1B	
Metric	ACF	TLCF	ACF	TLCF	ACF	TLCF	ACF	TLCF
ACR ↓	**0.02**	0.32	**0.05**	0.37	**0.27**	0.54	**0.41**	0.67
CPIR ↑	**0.96**	0.68	**0.97**	0.67	**0.87**	0.48	**0.75**	0.47
ST ↑	**0.76**		**0.72**		**0.82**		**0.80**	

References

1. Chatzi, I., Benz, N.C., Straitouri, E., Tsirtsis, S., Gomez-Rodriguez, M.: Counterfactual Token Generation in Large Language Models (Sep 2024), http://arxiv.org/abs/2409.17027, arXiv:2409.17027 [cs]
2. De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., Kalai, A.T.: Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting (Jan 2019). https://doi.org/10.48550/arXiv.1901.09451, http://arxiv.org/abs/1901.09451, arXiv:1901.09451
3. Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., Ravi, S.: GoEmotions: A Dataset of Fine-Grained Emotions (Jun 2020). https://doi.org/10.48550/arXiv.2005.00547, http://arxiv.org/abs/2005.00547, arXiv:2005.00547 [cs]
4. Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., Jha, A.H., Ivison, H., Magnusson, I., Wang, Y., Arora, S., Atkinson, D., Authur, R., Chandu, K.R., Cohan, A., Dumas, J., Elazar, Y., Gu, Y., Hessel, J., Khot, T., Merrill, W., Morrison, J., Muennighoff, N., Naik, A., Nam, C., Peters, M.E., Pyatkin, V., Ravichander, A., Schwenk, D., Shah, S., Smith, W., Strubell, E., Subramani, N., Wortsman, M., Dasigi, P., Lambert, N., Richardson, K., Zettlemoyer, L., Dodge, J., Lo, K., Soldaini, L., Smith, N.A., Hajishirzi, H.: Olmo: Accelerating the Science of Language Models (2024), https://arxiv.org/abs/2402.00838
5. Halpern, J.Y.: Cause, Responsibility, and Blame: A Structural-Model Approach (Dec 2014). https://doi.org/10.48550/arXiv.1412.2985, http://arxiv.org/abs/1412.2985, arXiv:1412.2985 [cs]
6. Kazemi, M., Lally, J., Tishchenko, E., Chockler, H., Paoletti, N.: Counterfactual Influence in Markov Decision Processes (Feb 2024), http://arxiv.org/abs/2402.08514, arXiv:2402.08514
7. Navigli, R., Conia, S., Ross, B.: Biases in Large Language Models: Origins, Inventory, and Discussion. J. Data and Information Quality **15**(2), 10:1–10:21 (Jun 2023). https://doi.org/10.1145/3597307, https://dl.acm.org/doi/10.1145/3597307
8. Oberst, M., Sontag, D.: Counterfactual Off-Policy Evaluation with Gumbel-Max Structural Causal Models. In: International Conference on Machine Learning. pp. 4881–4890. PMLR (2019)
9. Pan, A., Chan, J.S., Zou, A., Li, N., Basart, S., Woodside, T., Ng, J., Zhang, H., Emmons, S., Hendrycks, D.: Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the Machiavelli Benchmark. ICML (2023)
10. Pearl, J.: Causality. Cambridge University Press, New York (2009)
11. Ravfogel, S., Svete, A., Snæbjarnarson, V., Cotterell, R.: Gumbel Counterfactual Generation From Language Models. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=TUC0ZT2zIQ
12. Rivera, J.P., Mukobi, G., Reuel, A., Lamparth, M., Smith, C., Schneider, J.: Escalation Risks from Language Models in Military and Diplomatic Decision-Making. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency. pp. 836–898. ACM, Rio de Janeiro Brazil (Jun 2024). https://doi.org/10.1145/3630106.3658942, https://dl.acm.org/doi/10.1145/3630106.3658942
13. Singh, S., Ravfogel, S., Herzig, J., Aharoni, R., Cotterell, R., Kumaraguru, P.: Representation Surgery: Theory and Practice of Affine Steering (Jul 2024). https://doi.org/10.48550/arXiv.2402.09631, http://arxiv.org/abs/2402.09631, arXiv:2402.09631 [cs]

14. Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., Cohan, A., Gerstein, M.: MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning (Jun 2024). https://doi.org/10.48550/arXiv.2311.10537, http://arxiv.org/abs/2311.10537, arXiv:2311.10537 [cs]
15. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W.X., Wei, Z., Wen, J.R.: A Survey on Large Language Model based Autonomous Agents. Frontiers of Computer Science **18**(6), 186345 (Dec 2024). https://doi.org/10.1007/s11704-024-40231-1, http://arxiv.org/abs/2308.11432, arXiv:2308.11432 [cs]