Exploring Causal Mechanisms for Machine Text Detection Methods

KiYoon Yoo[1] **Wonhyuk Ahn**[2] **Yeji Song**[1] **Nojun Kwak**[1*]
[1]Seoul National University [2]Webtoon AI
{961230,ldynx,nojunk}@snu.ac.kr whahnize@gmail.com

Abstract

The immense attraction towards text generation garnered by ChatGPT has spurred the need for discriminating machine-text from human text. In this work, we provide preliminary evidence that the scores computed by existing zero-shot and supervised machine-text detection methods are not solely determined by the generated texts, but are affected by prompts and real texts as well. Using techniques from causal inference, we show the existence of backdoor paths that confounds the relationships between text and its detection score and how the confounding bias can be partially mitigated. We open up new research directions in identifying other factors that may be interwoven in the detection of machine text. Our study calls for a deeper investigation into which kinds of prompts make the detection of machine text more difficult or easier.

1 Introduction

Since its release, ChatGPT[1] has gained unprecedented attention from in and outside of the AI community, accumulating 100 million users within few months (Hu, 2023). Due to its articulate and fluent capability, the language model has been found to be an attractive assistant for writing essays, academic papers, news articles, etc. This has led to an increasing need for discriminating machine-generated from human-generated texts for a fair assessment of writings in educational institutions, proper authorship attribution for accountability in academic papers, preventing disinformation, etc (Acres, 2022; Kasneci et al., 2023; Stokel-Walker, 2023; Moran, 2023).

Many traditional works rely on the statistical nature of language modeling as the language model per se can estimate the conditional probability of the generated tokens (Gehrmann et al., 2019; Ippolito et al., 2020). This enables various ways to assess the text by using the rank of the predicted probability distribution or through the entropy thereof. On the other hand, more recent works like DetectGPT (Mitchell et al., 2023) discovered that machine-generated texts lie in a negative curvature area of the likelihood function. Besides the zero-shot methods, OpenAI has also released classifiers trained under supervision (Solaiman et al., 2019; Kirchner et al., 2023). All these methods compute a text's likelihood of being generated from a machine, which we hereafter dub as the detection score (i.e. token-level likelihood, level of curvature of the loss function).

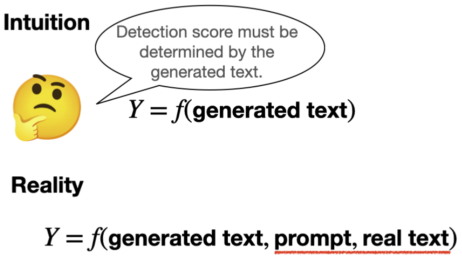
Intuition
Detection score must be determined by the generated text.
$Y = f$(generated text)
Reality
$Y = f$(generated text, prompt, real text)

Figure 1: The discrepancy between how the detection score $f(\cdot)$ is expected to be determined and actually determined in reality.

It is worth noting that all the aforementioned works focus only on the machine-generated texts without explicitly considering the possibly related variables such as the prompts that were given to generate the text or the real counterparts generated by humans. At first sight, this seems reasonable as the text's detection score must surely be determined by the text itself (Fig. 1). But are they the only factors that determine the scores in reality?

Research Goal In this work, we set out a new research direction by turning our attention to the other factors that may be interwoven when trying to assess a text's likelihood of being generated from a language model. Specifically, we study whether other factors besides the machine text itself have an effect on the detection score computed by the existing works. If such factors were to exist, this

[1] https://chat.openai.com

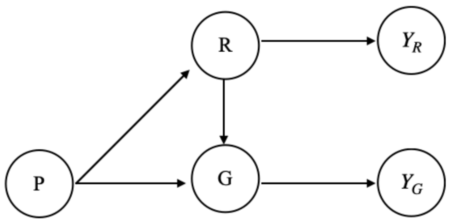

Figure 2: Causal diagram without backdoors that conveys conventional knowledge. P: Prompt, R: Real text, G: machine-generated text, Y_G: detection score of machine text, Y_R: detection score of real text.

implies that the detection scores are confounded by other variables that are not explicitly considered in the detection methods.

Findings Taking inspiration from the causal inference literature (Pearl, 2010; Pearl and Mackenzie, 2018), we leverage causal diagrams (as shown in Fig. 2) and show preliminary results that

- there exist backdoors between machine text and its detection scores for zero-shot detection methods and a supervised method. The upshot of this is that prompts affect the detection score not only through the machine text, but by other paths;
- the non-causal (biasing) effect can be partially adjusted for by conditioning on the prompts and the real texts;
- We show that the zero-shot methods and the supervised method display distinct behaviors that imply different causal relationships between the variables.

Implications Our findings have several implications. First, the observed association between the detection scores and generated texts demonstrated in previous works may not paint the full picture as there exists other mechanisms that affect the detection score. The existence of such biasing paths call for studies to see whether only considering the causal effect of G enhances the detection performance (i.e. separability of Y_G and Y_R). Our framework of using causal diagrams may help researchers identify inherent limitations of detectors when conditioned on certain prompts and give guidelines for practitioners to resort to other methods for those texts that are harder to detect.

2 Related Works

The potential societal impact of competent language models has called for the need to discriminate between their output and human-written texts (Solaiman et al., 2019; Goldstein et al., 2023). Since the release of a supervised classifier for GPT-2 with a 95% accuracy rate (Solaiman et al., 2019) in 2019, the task of detecting machine outputs has become severely more challenging: the new classifier for ChatGPT was reported to identify only 26% of AI-generated text as "likely AI-written," while misclassifying human-written text as AI-written at a rate of 9% (Kirchner et al., 2023). Recently, DetectGPT (Mitchell et al., 2023) proposed a zero-shot detector that uses an approximation of the curvature of a language model's log probability function, outperforming existing zero-shot methods (Gehrmann et al., 2019) for detecting machine-generated text and performing similarly or better than GPT-2 detectors. Watermarking (Abdelnabi and Fritz, 2021; Yang et al., 2022; Kirchenbauer et al., 2023; Yoo et al., 2023a,b) is another approach to identify machine-generated texts by encoding a secret message in the output of the language model. While there are research directions aimed at addressing the challenges to detection, such as robustness analysis of existing classifiers against paraphraser (Sadasivan et al., 2023; Krishna et al., 2023), there is a lack of fundamental analysis regarding the factors that impact the detection performance. We believe that conducting such an analysis could guide future directions toward a more reliable detection of machine texts.

3 Building the Causal Diagram

We briefly explain some notions of causal inference. For details, we refer the readers to Dablander (2020).

3.1 Preliminary

Causal diagram illustrates the causal relationship between random variables and can be represented by a directed acyclic graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ where $\mathcal{V}$ and $\mathcal{E}$ denote the set of variables (vertices) and cause-and-effect relationships (edges), respectively. An edge from variable $X \rightarrow Y$ denotes that X causes Y. More generally, X has a causal effect on all its descendents.

Fig. 2 depicts a causal diagram between prompts P, **r**eal texts written by humans R, machine-**g**enerated texts G, and its detection score Y_G. Both human and machine texts are completed conditioned on the prompts and are thus, "caused" by the prompts. In addition, the language model is trained

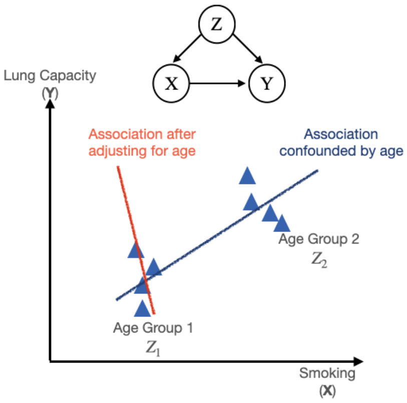
Z
X
Y
Lung Capacity
(Y)
Association after adjusting for age
Association confounded by age
Age Group 2
Z_2
Age Group 1
Z_1
Smoking
(X)

Figure 3: A hypothetical example for illustration of confounding bias and its causal model (from Feldman et al. 1987).

on the real text to follow its distribution. Hence, the generated texts are affected by the real text via the language model, i.e. $R \rightarrow G$.

Backdoors exist between a treatment X and a target variable Y when another variable Z is both an ancestor of X and Y. Backdoor variables introduce confounding bias, which obscures the true causal effect of X on Y from observational data. For instance, smoking (X) and lung capacity (Y) may have backdoor variables such as age (Z) (Feldman et al., 1987; Lee and Fry, 2010). If the amount of smoking decreases with age and younger people tend to have a better lung capacity, the observational data might hint that the more someone smokes, the better the lung capacity as shown in Appendix Fig. 3. However, when conditioning on an age group, this does not hold. We show that there exists a confounding bias between machine text and its detection score computed by several zero-shot detection methods.

3.2 Modeling Random Variables

The variables we consider in our graphical model are prompt P, generated text G, real text R, and detection score Y_G[2]. Barring the detection scores, the observational data for P, G, R are represented as raw texts, which is non-trivial to model as probability distribution. To tackle this, we borrow techniques from MAUVE (Pillutla et al., 2021) to model the text representations as embedding representations of language models, then quantizing them using a clustering method. The resultant representations are discrete probability representations of texts. To validate our modeling of random variables and the causal relationship between them, we ensure that statistical dependence exists between the adjacent nodes. The details are in A.1.

3.3 Experimental Settings

We experiment with two datasets (SQuAD and XSum) used in the literature. We use the Wikipedia context for SQuAD and the news articles for XSum. To quantify the level of independence/association, we use the G-test (Woolf, 1957) and conditional mutual information (MI). G-test verifies the null hypothesis that two given variables are independent. MI measures the dependence of two variables. We generate 10,000 samples on GPT2-Xl (Radford et al., 2019) by prompting it with the first 30 words of the real samples. We experiment with four zero-shot detection methods based on log likelihood, ranking of likelihood, entropy, and DetectGPT (Gehrmann et al., 2019; Mitchell et al., 2023) and a supervised classifier (Solaiman et al., 2019). More detailed explanations regarding modeling text as probability distributions and the metrics are provided in A.2.

4 Main Results

4.1 Checking for Backdoors

To start off, we presume a causal diagram (Fig. 2) that does not contain any confounding bias between the machine-generated text and the detection score. Then, we falsify the conditions that entails from this, proving otherwise.

Note that the only variable causing Y is G according to the diagram. The missing links between the nodes such as $R - Y$ entail testable implications followed by the d-seperation criterion (Geiger et al., 1990): $P \perp\!\!\!\perp Y|G$ and $R \perp\!\!\!\perp Y|G$. More specifically,

Claim. If $P \not\perp\!\!\!\perp Y|G$ or $R \not\perp\!\!\!\perp Y|G$, then there exists backdoor between G and Y that contains an arrow into Y (Proof in A.3).

To test this, we use the G-test using the implied conditional independence as the null hypothesis. Our results indicate that all the considered methods violate this implication, signifying that there exists backdoor(s). Note that a single statistically significant case (e.g. $P \not\perp\!\!\!\perp Y|G = g$) is sufficient to show $P \not\perp\!\!\!\perp Y|G$. Details are in Table 1.

4.2 Finding Potential Backdoor Paths

Having known that the backdoors exist, we can conjecture potential backdoor paths shown in Fig. 4

[2] e.g. perplexity, rank of conditional probability, or entropy. Hereafter, we use Y to denote Y_G for simplicity.

SQuAD			
Methods		Hypothesis	
		$P \perp\!\!\!\perp Y \mid G$	$R \perp\!\!\!\perp Y \mid G$
Zero-shot	DetectGPT	4e−2	0
	Logrank	9e−3	0
	Likelihood	8e−3	0
	Entropy	7e−3	2e−3
Supervised	Roberta-base	0	1e−1

XSum			
Methods		Hypothesis	
		$P \perp\!\!\!\perp Y \mid G$	$R \perp\!\!\!\perp Y \mid G$
Zero-shot	DetectGPT	3e−2	5e−3
	Logrank	3e−2	2e−2
	Likelihood	2e−2	5e−3
	Entropy	9e−3	1e−3
Supervised	Roberta-base	0	3e−3

Table 1: The lowest p-value over the support of G is shown (up to three decimal points) on SQuAD and XSum.

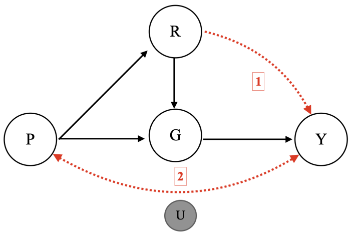

Figure 4: A causal diagram with two backdoor paths. U denotes some unobserved latent variable.

based on inductive bias.

Path $\boxed{1}$: For all the methods, the detection score is a function of a language model, which is not shown to reduce clutter. This language model is trained using the real texts as well, which may mediate the effect of R to Y. Without adjusting for any variables, $G-R$–Path $\boxed{1}$ and $G-P-R$–Path $\boxed{1}$ are backdoor paths to Y.

The causal diagram with Path $\boxed{1}$ added implicates the following conditional independence: $P \perp\!\!\!\perp Y|(G,R)$. When adjusting for only one of G or R, several paths are open from P to Y (shown in Appendix Fig. 8), which will lead to some level of association. We compute the unconditional MI and MI conditioned on several sets of variables to compare the level of association. We expect that MI$(P;Y|(G,R))$ will be the lowest as it blocks all paths. The results in Fig. 5 show a clear trend for the zero-shot methods: conditioning only on G and R tends to lead to a lesser change in the dependence of P and Y. However, when conditioning on both of the variables, the MI significantly decreases, bolstering the existence of Path $\boxed{1}$.

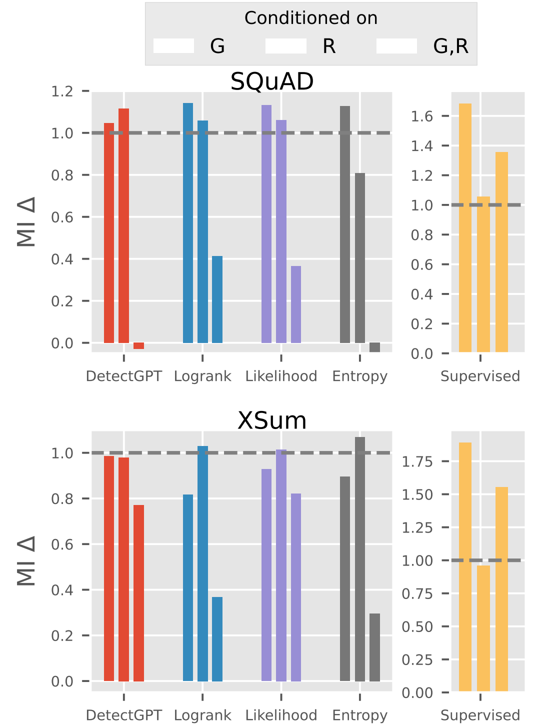

Figure 5: MI conditioned on the three sets of variables. All are normalized by the unconditional MI indicated by the horizontal dotted line corresponding to 1.0.

Conversely, this is not the case for the supervised method. Adjusting for G leads to a significant *increase* in the dependence. Similarly, adjusting for the two variables leads to an increase in the MI. This implies that adjusting for G leads to a d-connected path, indicating that our hypothesized graphical model does not accurately depict the data generating process for the supervised method. This is possible when G is a collider, opening up a path when observed as shown in Fig. 6.

Path $\boxed{2}$: When only Path $\boxed{1}$ is added to the existing links, this indicates $P \perp\!\!\!\perp Y|(G,R)$, hence MI$(P;Y|G,R)=0$. However, this is not the case for several cases, hinting at another path from P that is d-connected to Y. We show this as a bidirectional path owing to some unobserved latent variable. This may be caused by the same mechanism of Path $\boxed{1}$ whereby the language model is mediating the effect or by another mechanism that both affects P and Y (see Fig. 4).

4.3 Closing the Backdoor Paths and Implications

Last, we validate the backdoor paths directly by quantifying the level of association between the

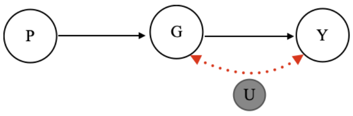

Figure 6: A causal diagram with G as a collider owing to an unobserved latent variable U. When G is conditioned, $P \rightarrow G \leftarrow Y$ is d-connected. Other paths and R are removed to reduce clutter.

generated text and its detection score when backdoor variables (P, R) are adjusted. We show in A.5 how Path $\boxed{1}$ and $\boxed{2}$ is blocked using the Backdoor Criterion (Javidianm and Valtorta, 2018). Our results in Fig. 7 demonstrate that adjusting for the backdoor variables leads to a decreased association (MI) for all the zero-shot methods (72.7% ↓ relative to the unconditional MI on average). This shows that the detection score of the generated text is indeed affected by factors other than the text itself. Once again, for the supervised classifier, adjusting for the variables has a marginal effect on the conditional MI.

What does the findings imply for detection methods? Since the detection scores computed by the current detection methods are affected by prompts as well, taking this into consideration might aid in enhancing the separability of human and machine texts. To illustrate this point, we show that certain prompts are indeed more difficult / easier to detect. As done in existing works (Mitchell et al., 2023; Gehrmann et al., 2019), we compute AUROC using the detection scores of real texts and generated texts. However, we do this by *conditioning on the prompt*. Then we perform permutation tests to see whether the highest and the lowest AUC are statistically significant. This tests whether the prompt with the highest AUC (easiest to detect) comes from the same distribution as a random subset of equal size. Our results in Table 2 show that all methods in the two datasets have at least one prompt cluster that is statistically easier or harder to detect.

This hints at the possibility of devising prompt-dependent detection methodology. For instance, for prompts that have low separability the API providers might want to resort to using more 'active' methods such as watermarking. Another potential application is adjusting for this backdoor to quantify the direct effect of generated text on the detection score. This can be done by counterfactual reasoning, which subtract out the indirect effect from the total effect (See Section 6.1 of Sobel (1996)).

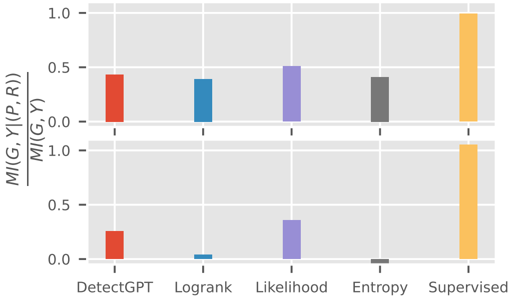

Figure 7: Relative MI of G, Y when unconditioned and when adjusted for the backdoor variables (top: XSum, bottom: SQuAD). All show a considerable decrease except the supervised method.

Methods		SQuAD		XSum	
		easier	harder	easier	harder
Zero-shot	DetectGPT	2e−2	7e−3	0	5e−2
	Logrank	2e−4	0	3e−3	3e−2
	Likelihood	4e−4	0	8e−3	1e−2
	Entropy	4e−2	8e−2	1e−1	7e−3
Supervised	Roberta-base	3e−1	0	1e−1	0

Table 2: The p-values using permutation test for the null hypothesis that "a prompt that is {easier, harder} to detect follows the same distribution with a randomly sampled subset" under $\alpha = .05$. Significant prompts that have lower p-values ($< \alpha$) are marked in red.

5 Conclusion

In summary, we demonstrate that backdoor variables exist between the machine texts and their detection scores. While all methods have backdoors, the results hint that the causal relationships are distinct for the supervised classifier, the precise mechanism of which is yet to be investigated. Our work opens up new research direction in detecting machine-generated texts without non-causal paths.

Limitations

The results shown in this study is limited to few datasets and a small-scale model. In addition, modeling the raw texts into a probability distribution is a non-trivial task to achieve without losing potentially important information. This may be a bottleneck in finding association between the variables. Nonetheless, our preliminary study opens up various research directions. Namely, the framework can be used to overhaul existing methods that rely on confounding biases. Another practical challenge is that prompts are generally unknown when trying to detect machine text. This makes devising prompt-dependent method difficult even if accounting for it is indeed helpful. To overcome this, using proxy variables such as topics or semantics instead of prompts might be necessary.

References

Sahar Abdelnabi and Mario Fritz. 2021. Adversarial watermarking transformer: Towards tracing text provenance with data hiding. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 121–140. IEEE.

Tom Acres. 2022. Chatgpt: We let an ai chatbot help write an article - here's how it went.

Fabian Dablander. 2020. An introduction to causal inference.

Henry A. Feldman, Joseph D. Brain, and Margaret L. Harbison. 1987. Adjusting for confounded variables: Pulmonary function and smoking in a special population. *Environmental Research*, 43(1):251–266.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116.

Dan Geiger, Thomas Verma, and Judea Pearl. 1990. d-separation: From theorems to algorithms. In *Machine Intelligence and Pattern Recognition*, volume 10, pages 139–148. Elsevier.

Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*.

Krystal Hu. 2023. Chatgpt sets record for fastest-growing user base - analyst note. *Reuters*.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822.

Mohammad Ali Javidianm and Mohammad Ali Valtorta. 2018. An overview of the back-door and front-door criteria.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. *arXiv preprint arXiv:2301.10226*.

Jan Hendrik Kirchner, Lama Ahmad, Scott Aaronson, and Jan Leike. 2023. New ai classifier for indicating ai-written text.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *arXiv preprint arXiv:2303.13408*.

Peter N Lee and John S Fry. 2010. Systematic review of the evidence relating fev1decline to giving up smoking. *BMC medicine*, 8(1):1–29.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*.

Chris Moran. 2023. Chatgpt is making up fake guardian articles. here's how we're responding.

Judea Pearl. 2010. An introduction to causal inference. *The International Journal of Biostatistics*, 6(2):1–62.

Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.

Michael E Sobel. 1996. An introduction to causal inference. *Sociological Methods & Research*, 24(3):353–379.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Chris Stokel-Walker. 2023. Chatgpt listed as author on research papers: many scientists disapprove.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2009. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th annual international conference on machine learning*, pages 1073–1080.

Barnet Woolf. 1957. The log likelihood ratio test (the g-test). *Annals of human genetics*, 21(4):397–409.

Xi Yang, Jie Zhang, Kejiang Chen, Weiming Zhang, Zehua Ma, Feng Wang, and Nenghai Yu. 2022. Tracing text provenance via context-aware lexical substitution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11613–11621.

KiYoon Yoo, Wonhyuk Ahn, Jiho Jang, and Nojun Kwak. 2023a. Robust multi-bit natural language watermarking through invariant features. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2092–2115.

KiYoon Yoo, Wonhyuk Ahn, and Nojun Kwak. 2023b. Advancing beyond identification: Multi-bit watermark for language models. *arXiv preprint arXiv:2308.00221*.

A Appendix

A.1 Validating the Model

After binning Y, we use the G-test for the four relationships ($P-R$, $P-G$, $R-G$, $G-Y$). For all the studied methods, the p-values of the four relationships are statistically significant at $\alpha = .05$.

A.2 More Details on Experimental Settings

Modeling Random Variables For all the zero-shot methods, we use the last token embedding of GPT as the representation. For the supervised method, we use the classifier's [CLS] token embedding as the representation. For clustering, we first conduct dimensionality reduction using PCA and apply K-Means Clustering. For the detection score, the scores between 1% and 99% quantiles are kept so as to remove the outliers. We apply Box-Cox transformation to skewed score distributions before discretizing them. The number of clusters, PCA dimension, and the bins for the scores are all chosen from {5,10} such that the conditions in §3.2 are satisfied.

Metrics G-test measures the difference of likelihood given the null hypothesis that the two variables are independent and thus, lower p-value indicates association between variables. Conversely, MI measures the level of association, hence higher values indicate association. We use the adjusted MI (Vinh et al., 2009) to account for randomness and add a uniform prior of 0.01 for all bins as the samples are sparse when conditioning on multiple variables. This tends to bias the measure towards higher mutual information, but leads to a more robust estimation towards noise due to limited sample sizes. This is especially important when conditioning on more than one variable as the number of bins when conditioning on two variables becomes 25-100 if 5-10 clusters are used for each variable, which can become sparse or noisy even when 10,000 samples are generated. Empirically, we observe that by adding the uniform prior the MI and the G-test lead to consistent results: when G-test is not significant, the MI is always close to zero.

The most computation-heavy part of our experiment was generating the samples, which around 24 gpu-hours on Titan RTX.

A.3 Proof of Claim

Proof. Given the causal diagram in Fig. 2, if $P \not\perp\!\!\!\perp Y|G$, this means there exist d-connected path(s) from P to Y. Paths through the only connected variable G are blocked as G admits an arrow towards Y. Thus, $R \rightarrow Y$ (notice the direction) must exist or $P \leftarrow Y$, $P \rightarrow Y$ must exist. The same argument applies for the case when $R \not\perp\!\!\!\perp Y|G$. □

A.4 Visualization of a collider variable, Causal and Biasing Paths

We visualize the active causal path(s) and the biasing path(s) in green and red (shown in Fig. 8). A path is active if all the triplets in the path are d-connected. A path is causal if the target variable (Y) is a descendant of the treatment variable (G).

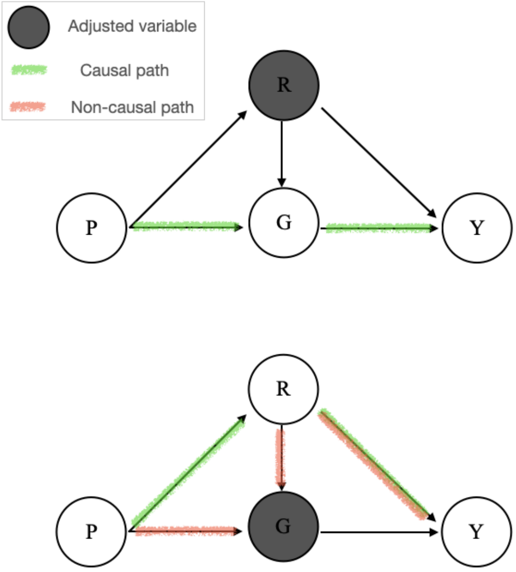

Figure 8: Causal diagram visualizing the *d-connected* causal and non-causal paths from P to Y when adjusting for variables.

A.5 Blocking Path $\boxed{1}$ and $\boxed{2}$

For completeness, we state the backdoor criterion.

Definition (Backdoor Criterion)**.** A set $\mathcal{Z}$ satisfies the backdoor criterion with respect to X and Y if

1. no node in $\mathcal{Z}$ is a descendant of X and
2. conditioning on $\mathcal{Z}$ blocks every d-connected path between X and Y that contains an arrow into X.

Adjusting for $\mathcal{Z} = \{P, R\}$ satisfies the backdoor criterion.