PNAS RESEARCH ARTICLE | SOCIAL SCIENCES

Leveraging generative AI for causal inference with unstructured data

Kosuke Imai[a,b,1,2] and Kentaro Nakamura[c,1]

Affiliations are included on p. 5.

Edited by David M. Blei, Columbia University, New York, NY; received October 24, 2025; accepted July 29, 2026 by Editorial Board Member Kenneth W. Wachter

We introduce GenAI-Powered Inference (GPI), a statistical framework for causal inference using unstructured data, including text and images. GPI leverages open-source pretrained Generative AI (GenAI) models—such as large language models and diffusion models—not only to generate unstructured data at scale but also to extract low-dimensional representations that are guaranteed to capture their underlying structure. Applying machine learning to these representations, GPI enables estimation of causal effects while quantifying estimation uncertainty. Unlike existing approaches to representation learning, GPI does not require fine-tuning of GenAI models, making it computationally efficient and broadly accessible. We illustrate the versatility of the GPI framework through three applications: 1) estimating the effects of Chinese social media censorship while adjusting for textual confounders, 2) isolating the impact of specific image features from that of other correlated features in the same image, and 3) assessing the persuasiveness of political rhetoric. An open-source software package is available for implementing GPI.

generative AI | causal inference | machine learning | unstructured data

Significance

We introduce GenAI-Powered Inference (GPI), a unified methodological framework that integrates open-source large language and diffusion models into causal inference. GPI uses generative AI (GenAI) models to extract low-dimensional representations that retain the information necessary to generate complex unstructured data. This enables researchers to estimate causal effects without directly modeling high-dimensional objects or fine-tuning foundation models. We show that estimating and adjusting for a deconfounder leads to valid causal inference with unstructured data. By leveraging pretrained GenAI models while avoiding restrictive modeling assumptions, GPI delivers strong empirical performance across diverse settings. We demonstrate the versatility of the framework in three distinct applications and provide open-source software to facilitate its use in applied research.

The emergence of Generative AI (GenAI), including large language models and diffusion models, has transformed society and scientific research. In this paper, we show that GenAI can also serve as a powerful tool for causal inference with unstructured data such as text and images. Specifically, we introduce a statistical framework, GenAI-Powered Inference (GPI), which leverages open-source GenAI models to improve causal inference with high-dimensional unstructured data.

We illustrate the power of GPI through the following three applications:

1. **Text as confounder**: Estimating the causal effects of a treatment variable by adjusting for latent confounders embedded in textual data
2. **Image as Treatment**: Isolating the impact of specific image features from that of other correlated features in the same image
3. **Structural model of text**: Fitting a structural model that incorporates both observed and latent textual features as covariates

A common challenge in these settings is that while high-dimensional unstructured data are observed, we do not know, a priori, which features of such data serve as relevant confounders.

GPI addresses this challenge by leveraging the ability of GenAI to generate or reproduce unstructured data and extracting lower-dimensional vector representations that are guaranteed to contain all the relevant information. Machine learning is then applied to these representations to estimate a *deconfounder*, which is a summary of confounding information. Because these representations were used to generate unstructured data, GPI eliminates the need to explicitly model the data-generating process, fine-tune foundation models, or estimate text and image representations. This makes GPI a broadly accessible framework. Our open-source Python package GPI, which is available at https://gpi-pack.github.io/, implements the proposed methodology to further enhance its applicability in applied research.

Before presenting the three empirical applications, we provide a brief overview of the proposed GPI methodology (see also ref. 1). Fig. 1 considers two illustrative settings, in which the goal is to estimate the causal effect of treatment T on an outcome Y. The treatment may represent some features of unstructured data $\boldsymbol{X}$ as in the "text/image as treatment" case (*Bottom Right*) or other structured external variables as in the "text/image as confounder" case (*Top Right*). In both settings, the treatment–outcome

The authors declare no competing interest.

This article is a PNAS Direct Submission. D.M.B. is a guest editor invited by the Editorial Board.

[1]K.I. and K.N. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: imai@harvard.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2530532123/-/DCSupplemental.

Published September 3, 2026.

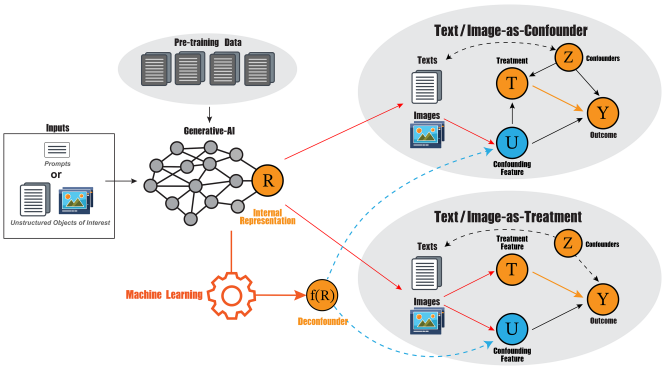

Fig. 1. Illustration of GenAI-Powered Inference (GPI). Inputs to generative AI are either prompts or unstructured data (e.g., text/images). Pretrained models produce internal representations $\boldsymbol{R}$, which we use for machine learning in two application settings: i) when text/images act as confounders, and ii) when they serve as treatments. In the first setting (*Top Right*), the treatment–outcome relation is confounded by the latent features of unstructured data $\boldsymbol{U}$ and observed covariates $\boldsymbol{Z}$. In the second setting (*Bottom Right*), the treatment depends on unstructured data and is similarly confounded. Since $\boldsymbol{U}$ is unknown, we apply machine learning to estimate a deconfounder $\boldsymbol{f}(\boldsymbol{R})$, which is a low-dimensional summary of confounding features. GPI enables unbiased causal inference without modeling the generative process or fine-tuning the GenAI model.

relationship is confounded by the latent features $\boldsymbol{U}$ of unstructured data $\boldsymbol{X}$ as well as other observed structured confounders $\boldsymbol{Z}$. A key challenge is to identify these latent (unstructured) confounding features $\boldsymbol{U}$ and adjust for them to estimate causal effects.

To achieve this, GPI leverages GenAI models to generate unstructured data $\boldsymbol{X}$ and extract the corresponding internal representations $\boldsymbol{R}$. For existing text or images, GenAI models are prompted to reproduce the original content, enabling the recovery of consistent internal representations. Because we configure GenAI models to produce $\boldsymbol{X}$ as a deterministic function of $\boldsymbol{R}$, the latter is guaranteed to contain all information necessary to generate the former. As a result, GPI eliminates the need to fine-tune $\boldsymbol{R}$, unlike standard approaches that rely on generic pretrained embeddings such as BERT (Bidirectional Encoder Representations from Transformers).

The use of $\boldsymbol{R}$ offers several important advantages over directly modeling the raw data $\boldsymbol{X}$. First, $\boldsymbol{R}$ provides a substantially lower-dimensional representation of the original high-dimensional input. Second, it captures rich and relevant information distilled from the vast data used to train the GenAI model. Third, by relying on open-source GenAI models that operate deterministically, GPI promotes scientific replicability. Finally, GPI can further improve statistical efficiency by combining the results based on multiple GenAI models.

GPI applies machine learning to the internal representation $\boldsymbol{R}$ to estimate a deconfounder $\boldsymbol{f}(\boldsymbol{R})$, which serves as a low-dimensional approximation of the latent confounding features $\boldsymbol{U}$. This step is critical because directly adjusting for the high-dimensional unstructured data $\boldsymbol{X}$ can violate the positivity assumption and lead to biased inference (2). While the deconfounder is not necessarily unique, we formally show that any valid $\boldsymbol{f}(\boldsymbol{R})$ enables the nonparametric identification of causal effects when adjusted for together with the observed confounders $\boldsymbol{Z}$. We propose estimating the deconfounder using a deep neural network architecture designed to directly encode these necessary properties.

GPI learns a low-dimensional deconfounder $\boldsymbol{f}(\boldsymbol{R})$ that is predictive of the outcome within each treatment group while remaining shared across treatment and control groups. By discarding components of $\boldsymbol{R}$ that predict treatment assignment alone, $\boldsymbol{f}(\boldsymbol{R})$ also makes the overlap assumption more likely to hold than standard approaches in the literature (e.g., refs. 3–5). We emphasize the importance of carefully tuning the neural network architecture used in GPI: The network must be sufficiently expressive to preserve outcome-relevant information, yet parsimonious enough to remove irrelevant variation. In *SI Appendix*, section 1, we empirically illustrate this point through a simple simulation study. The SI also provides theoretical justifications of GPI tailored to each application setting described below.

1.1. Related Literature. A growing body of work addresses causal inference with unstructured high-dimensional data, often by either estimating low-dimensional embeddings (e.g., refs. 4–7) or modeling the data-generating process using parametric methods such as topic models (e.g., refs. 8–12). These approaches typically rely on strong assumptions, such as the bag-of-words model and fixed embedding schemes, and face substantial challenges due to the statistical and computational complexity of high-dimensional unstructured data. This may lead to biased estimates and unreliable inference.

GPI departs from this paradigm by leveraging the internal representations of GenAI models, which inherently encode rich semantic information about the generated data and eliminate the need to explicitly model or estimate representations. Further, GPI adopts a fully nonparametric approach by using neural networks to estimate the deconfounder, outcome model, and propensity score model, thereby avoiding restrictive functional form assumptions. This combination of principled representation learning and model flexibility enables more robust causal inference in complex, high-dimensional settings.

Below, we demonstrate the wide applicability of GPI by reanalyzing three empirical studies under the proposed framework: estimating the causal effect of government censorship experience on the subsequent censorship or self-censorship (10), estimating the effects of nighttime scenes on the perceived violence, and estimating the persuasiveness of the political rhetoric (13).

2. Text as Confounder

Social scientists have documented that the Chinese government selectively censors its citizens' social media posts e.g., refs.14 and 15. Building on this literature, ref. 10 investigated whether users whose posts are censored are more likely to face future censorship and whether they respond by engaging in self-censorship.

The authors analyzed data from 75,324 Weibo posts collected by the Weiboscope project (16), which captures posts in real time and later checks whether they have been removed. Each post serves as a focal post, and they constructed three outcome variables: i) the number of posts made by the same user in the four weeks following the focal post, ii) the proportion of those posts that were censored during the same time period; and iii) the proportion of posts that went missing during the same four-week window.

Because censored and uncensored posts differ systematically in their contents, a naive regression of these outcomes on the treatment variable (prior censorship) is likely to yield biased causal estimates. To address this potential confounding bias, the original analysis adjusts for the content of the focal post based on a topic model, the posting date, and each user's prior censorship history. We adopt the same covariate adjustment strategy within the GPI framework.

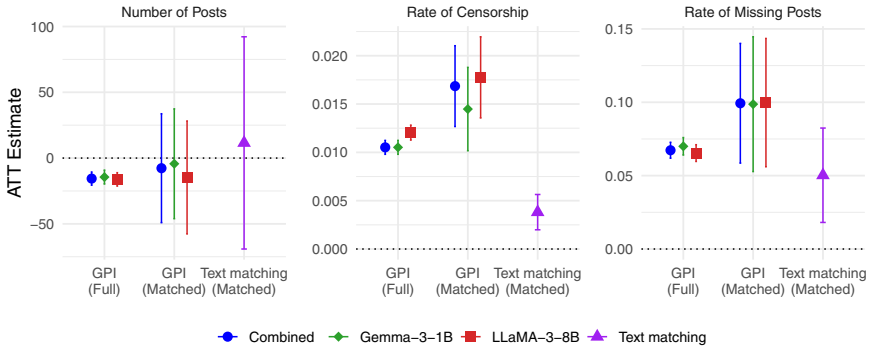

Fig. 2. Estimated causal effects of prior censorship experience. For each outcome, we report the GPI estimates based on LLaMA3 with eight billion parameters (green diamonds) and Gemma3 with one billion parameters (red squares) using the full sample (Full) and the matched sample (Matched). We also present the estimates from an optimally weighted combination of the two models (blue circles). We compare GPI estimates with those of the text matching approach used in the original analysis (purple triangles). The interval bars represent 95% CIs with the SEs clustered at the user level.

We use two open-source large language models, LLaMA3 with eight billion parameters (17) and Gemma3 with one billion parameters (18), to reproduce each focal post and extract its internal representation. These representations are then used to jointly estimate the deconfounder and outcome model within a single neural network architecture. A separate neural network is used to estimate the propensity score, conditional on the estimated deconfounder and other observed confounders. The deconfounder and outcome model are fine-tuned using Optuna, an automated hyperparameter optimization framework (19).

To estimate the average treatment effect on the treated (ATT), we apply double machine learning (DML) with two-fold cross-fitting (20). SEs are clustered at the user level, and extreme propensity scores are truncated following the method proposed by ref. 21. We also construct the optimal combination of the GPI estimates from the two GenAI models based on their estimated influence functions, which can yield more efficient estimates. Theoretical justifications and further implementation details are provided in *SI Appendix*, sections 2.A and 2.B.

Fig. 2 compares the GPI estimates with those based on the text matching approach used in the original analysis, which employed coarsened exact matching (22) on estimates from the structural topic model (23). For comparison, we apply GPI to both the matched sample (628 users, 879 posts) and the full sample (4,155 users; 75,324 posts).

In contrast to the original analysis, GPI finds that prior censorship significantly reduces users' subsequent posting activity, providing evidence of self-censorship. Moreover, GPI reveals a much greater effect of prior censorship on the likelihood of future censorship. In addition, GPI's full-sample estimates are substantially more efficient than those based on the matched sample, highlighting GPI's advantages in statistical efficiency. These findings are consistent across both LLaMA3 and Gemma3 models. Last, optimally combining these two results leads to a modest reduction in the SE by 4.4% on average.

To further illustrate GPI's advantage, we assess robustness to a text-based confounder by estimating each method with and without this confounder and calculating the relative absolute bias, which is defined as the absolute change in a point estimate relative to the original estimate without the text-based confounders. The text-based confounder of interest is measured as the proportion of 60 keywords related to censorship or self-censorship as identified by ref. 16. If a method effectively adjusts for the confounding features in focal posts, the additional adjustment for the text-based confounder should yield only a small change. We use the delta method for the calculation of GPI's SEs. For text matching, we use bootstrap since an analytical SE formula is unavailable.

As shown in Table 1, GPI consistently produces smaller relative absolute bias than text matching. The results based on Gemma3 are shown in *SI Appendix*, Table S4. We find that the GPI estimates produce a similar ATT estimate for each outcome, regardless of which LLM model is used. We find that the optimal combination does not substantially reduce the SE in our application because the estimated influence functions from LLaMA3-8B and Gemma3-1B are highly correlated in our application as shown in *SI Appendix*, Table S5. This high correlation may be due to the fact that different LLMs are trained on similar sets of data.

We investigate whether the advantage of GPI arises from text representation, improved estimation, or both. To do this, we implement the same estimation procedure using two off-the-shelf text embeddings, including Sentence-BERT (24) and Qwen3 embedding with 0.6 billion parameters (25). Unlike the internal representation of LLM, these alternative text representations may not contain all relevant information necessary for confounding adjustment.

The estimates based on these alternative representations are shown in *SI Appendix*, Table S4. We find that the estimated ATTs based on Qwen3-0.6B embeddings are qualitatively similar to those of GPI, though the estimated ATT on the rate of missingness is statistically significantly smaller than those of GPI. In contrast, the results based on Sentence-BERT embeddings are quite different, indicating that censorship decreases the rate of subsequent censorship rather than increasing it. The estimated ATT on the rate of missingness is even smaller.

We also conducted the above sensitivity analysis using these alternative text representations. *SI Appendix*, Tables S6 and S7

Table 1. Relative absolute bias due to a text-based confounder

Outcome	GPI (LLaMA3-8B) Full	GPI (LLaMA3-8B) Matched	Text matching Matched
Number of posts	0.084	0.242	5.350
	(0.068)	(0.682)	(33.794)
Rate of censorship	0.007	0.196	0.202
	(0.003)	(0.140)	(0.277)
Rate of missing posts	0.040	0.279	0.627
	(0.049)	(0.346)	(3.416)

SEs are in parentheses. "Full" refers to estimates using the entire sample, while "Matched" refers to estimates based on the matched sample.

present the results. We find that the robustness of Qwen 3–0.6B embeddings is comparable to that of GPI. However, Sentence-BERT is much more sensitive to the inclusion of additional textual confounders. The corresponding estimates vary much more as reflected by larger SEs for the point estimates of bias, some of which are also greater than those of GPI and Qwen embeddings. These results suggest that the quality of text representation matters. The advantage of GPI is that the internal representation of LLM is guaranteed to contain all information required for reproducing unstructured objects of interest. This theoretical guarantee comes at the computational cost as pretrained embedding models are typically less computationally expensive than reproducing text through LLMs.

3. Image as Treatment

As protest images circulate widely on social media, they have become an important source of information about the scale and character of collective action. To study these images, ref. 26 compiled a dataset of more than 9,000 protest images and obtained human annotations for several attributes, including whether each image depicts a daytime or nighttime scene and its perceived level of violence.

We use this dataset to empirically validate the GPI methodology by estimating the causal effect of nighttime scenes on perceived violence. Nighttime protest images are associated with higher perceived violence, as they are more likely to include visual cues such as confrontations, fire, smoke, and police presence compared to daytime images (correlation is 0.213).

However, because an image can be transformed from nighttime to daytime without altering its underlying violent content, the causal effect of nighttime on perceived violence is expected to be small; the causal effect may not be zero given that the outcome variable is based on human perception. This suggests that the GPI methodology should substantially attenuate the association between nighttime and perceived violence by adjusting for confounding features.

To apply the GPI methodology, we first reproduce each protest photo using two different versions of Stable Diffusion model (versions 1.5 and 2.1) (27) and extract their internal representations. We then estimate the deconfounder and outcome model using the same neural network architecture and fine-tuning procedure as the one employed in the previous application. Since the internal representation of images is still relatively high-dimensional but still retains spatial structure, we use the convolutional neural network (CNN) as an encoding layer. Last, we combine these estimates using the optimal weighting scheme. *SI Appendix*, section 3 provides a theoretical justification of GPI methodology for this application and the implementation details.

We compare the GPI methodology with i) difference-in-means estimates and ii) estimates obtained from the same neural network architecture applied directly to the original images, rather than to the internal representations of Stable Diffusion models.

Fig. 3 reports the results. Each horizontal bar represents 95% CI. We find that the estimated effect of nighttime scenes on perceived violence is reduced by more than half once we adjust for other image features through the GPI methodology (*Top* panel). In addition, the estimates are similar across the two different versions of Stable Diffusion model. Combining the two estimates reduces SE by a modest amount of 7.4% from GPI (Stable Diffusion 1.5) and 2.2% from GPI (Stable Diffusion 2.1).

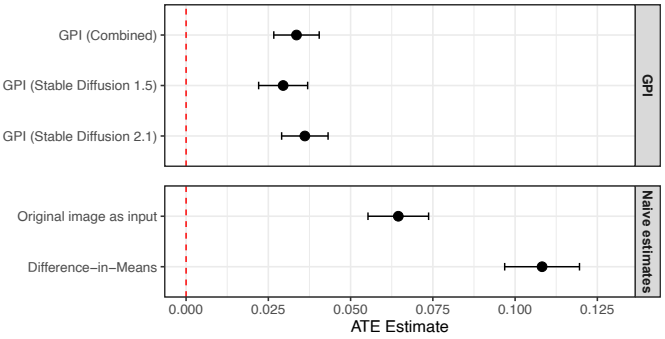

Fig. 3. Estimated effects of nighttime scene on the perceived violence using GPI (*Top* panel), directly using the original images with the same neural network architecture (first estimate in the *Bottom* panel), and difference-in-means (second estimate in the *Bottom* panel). Each bar represents the 95% CI.

Finally, the internal representation of Stable Diffusion model is of substantially lower dimension ($16{,}384 = 64 \times 64 \times 4$) than the original representation ($786{,}432 = 512 \times 512 \times 3$), where all of our images have been adjusted to have an identical size of 512×512 pixels. The internal representation also contains rich semantic information while the original representation only includes positional and color information. The *Bottom* panel of Fig. 3 shows that the estimates based on the original representation yield an estimate of much greater magnitude than GPI, suggesting that the high-dimensionality and rich semantic information of GenAI model's internal representation improve empirical performance of causal effect estimation. The GPI estimates are also more precise, with substantially smaller SEs than the estimator based on the original representation.

4. Structural Model of Text

Political scientists and communication scholars have extensively studied the effectiveness of different rhetorical strategies (e.g., refs. 28 and 29. Contributing to this literature, ref. 13 conducted a forced-choice conjoint experiment to evaluate the persuasiveness of different types of political rhetoric. The authors constructed 336 political arguments by systematically varying 12 policy issues, 14 rhetorical elements, and position (support or opposition). *SI Appendix*, section 4.A provides the full list of policy issues and rhetorical features.

A total of 3,317 participants were each shown four randomly assigned pairs of arguments. Within each pair, the arguments addressed the same policy issue but took opposite sides and featured different, randomly selected rhetorical elements. Participants were asked to indicate which argument they found more persuasive or whether both were equally persuasive, yielding a total of 13,268 pairwise comparisons. *SI Appendix*, Table S9 presents an example pair of arguments.

The authors estimated the latent persuasiveness of rhetorical elements using a parametric structural model similar to the Bradley–Terry model (30) (see *SI Appendix*, section 4.B for the details of the model). This original model assumes that the probability of one argument being judged more persuasive than the other is determined by an additive combination of three effects: an interaction effect between policy area and position, the effect of the rhetorical element, and a random effect for each individual argument. The random effect is used to account for unobserved features of the arguments that may influence persuasiveness but are not explicitly measured. In addition, the original analysis manually identified several

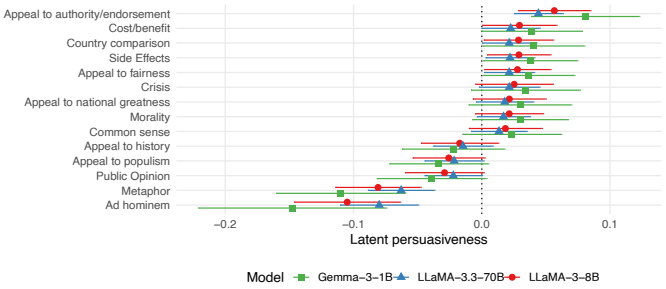

Fig. 4. Estimated persuasiveness of 14 rhetorical elements in political arguments. We present the estimated latent persuasiveness based on LLaMA3 with 8 billion parameters (red circle), LLaMA3.3 with 70 billion parameters (blue triangle), and Gemma3 with 1 billion parameters (green rectangle). Each error bar represents the 95% CIs obtained by percentile bootstrap.

potential confounders such as argument length and included them as covariates.

We use GPI to enhance the original analysis by adjusting for the confounding factors that are based on other features of arguments. We compare the results based on three different LLMs: LLaMA3 with eight billion parameters, along with two more recently released models — LLaMA3.3 with 70 billion parameters, and Gemma3 with one billion parameters. While all these models are instruction-tuned and thus can be used to regenerate each argument and extract the internal representation, their internal representations have varying dimensions and contain different information. Nevertheless, since these internal representations are all sufficient for reproducing the original texts, our theoretical results imply that the results based on these models should be similar.

Once we extract internal representations, we then estimate a semiparametric version of the original structural model, which allows for greater flexibility in capturing complex relationships,

$$\log\left[\frac{\mathbb{P}(Y_{ijj'} \leq k)}{\mathbb{P}(Y_{ijj'} > k)}\right] = \delta_k + \beta_{T_j} - \beta_{T_{j'}} + h(\boldsymbol{R}_j) - h(\boldsymbol{R}_{j'}),$$

where $Y_{ijj'} \in \{0, 1, 2\}$ denotes the relative persuasiveness of arguments j and j' shown to respondent i with 0 indicating that argument j is more persuasive than j', 1 indicating equal persuasiveness, and 2 indicating argument j is less persuasive. The variable $T_j \in \{1, 2, \ldots, 14\}$ denotes a rhetorical element used in argument j, and $h(\boldsymbol{R}_j)$ represents latent confounding features of that argument. The original analysis uses a parametric version of this model, which may not capture all confounding features of text. We do not consider a fully nonparametric model given that the number of unique text is only 336, making it impossible to reliably estimate the causal effect of interest.

We estimate the latent persuasiveness parameter β_{T_j} using a neural network architecture for $h(\boldsymbol{R}_j)$ and fine-tuning procedure similar to those employed in the previous applications. This approach allows us to evaluate the effect of each rhetorical element while adjusting for latent confounding features embedded in the arguments. To quantify uncertainty, we use the percentile bootstrap with 200 Monte Carlo replications. Theoretical details and implementation specifics are provided in *SI Appendix*, sections 4.B and 4.C, respectively.

Fig. 4 presents the estimated effects of the 14 rhetorical elements. We find that appeal to authority is the most persuasive. In contrast, ad hominem attacks, which represent arguments targeting the person rather than their position, significantly reduce persuasiveness. These findings are consistent across GenAI models and each model's point estimates are highly correlated (*SI Appendix*, Fig. S4). While these findings are largely consistent with those of the original analysis, the GPI estimates are substantially more precise. For instance, the original results did not find statistically significant effects for appeal to authority or cost/benefit arguments, despite similar point estimates.

5. Concluding Remarks

In this paper, we introduced GenAI-Powered Inference (GPI), a statistical framework that harnesses open-source GenAI models to improve causal inference with unstructured data, such as text and images. GPI extracts internal representations from these models, yielding low-dimensional features that preserve all the relevant information that is sufficient for generating the data. By applying machine learning to these representations, GPI enables robust estimation of causal effects, while also quantifying estimation uncertainty. We demonstrated GPI's broad applicability through three different applications.

This work opens several promising avenues for future research. First, developing methods to interpret the estimated deconfounder would help researchers better understand the latent confounding features identified by GPI. Second, GPI could be extended to discover effective treatment features from unstructured data. While the current framework assumes treatment features are predefined and observed, many applications would benefit from discovering novel, more informative treatments. Finally, although we focused on text and image data, the GPI framework can be naturally extended to other types of unstructured data, including audio and video (31, 32).

Data, Materials, and Software Availability. Text, image, and tabular data have been deposited in Harvard Dataverse (33). All other data are included in the manuscript and/or *SI Appendix*.

ACKNOWLEDGMENTS. We thank Naoki Egami for helpful comments.

Author affiliations: [a]Department of Government, Harvard University, Cambridge, MA 02138; [b]Department of Statistics, Harvard University, Cambridge, MA 02138; and [c]John F. Kennedy School of Government, Harvard University, Cambridge, MA 02138

Author contributions: K.I. and K.N. designed research; K.I. and K.N. performed research; K.I. and K.N. contributed new analytic tools; K.N. analyzed data; and K.I. and K.N. wrote the paper.

1. K. Imai, K. Nakamura, Causal inference with generative artificial intelligence: Application to texts as treatments. *J. Am. Stat. Assoc.*, 10.1080/01621459.2026.2689629 (2026).
2. A. D'Amour, P. Ding, A. Feller, L. Lei, J. Sekhon, Overlap in observational studies with high-dimensional covariates. *J. Econom.* **221**, 644–654 (2021).
3. C. Shi, D. Blei, V. Veitch, "Adapting neural networks for the estimation of treatment effects" in *Advances in Neural Information Processing Systems*, H. Wallach *et al.*, Eds. (Curran Associates Inc., 2019), vol. 32.
4. V. Veitch, D. Sridhar, D. M. Blei, Adapting text embeddings for causal inference. arXiv [Preprint] (2020). https://arxiv.org/abs/1905.12741 (Accessed 13 June 2023).
5. R. Pryzant, D. Card, D. Jurafsky, V. Veitch, D. Sridhar, Causal effects of linguistic properties. arXiv [Preprint] (2021). https://arxiv.org/abs/2010.12919 (Accessed 10 June 2023).
6. L. Gui, V. Veitch, Causal estimation for text data with (Apparent) overlap violations. arXiv [Preprint] (2023). https://arxiv.org/abs/2210.00079 (Accessed 5 June 2023).
7. S. Klaassen *et al.*, Estimation of causal effects with multimodal data. arXiv [Preprint] (2024). https://arxiv.org/abs/2402.01785 (Accessed 4 July 2024).
8. C. Fong, J. Grimmer, "Discovery of treatments from text corpora" in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, K. Erk, N. A. Smith, Eds. (Association for Computational Linguistics, Berlin, Germany, 2016), pp. 1600–1609.
9. R. Mozer, L. Miratrix, A. R. Kaufman, L. J. Anastasopoulos, Matching with text data: An experimental evaluation of methods for matching documents and of measuring match quality. *Polit. Anal.* **28**, 445–468 (2020).
10. M. E. Roberts, B. M. Stewart, R. A. Nielsen, Adjusting for confounding with text matching. *Am. J. Polit. Sci.* **64**, 887–903 (2020).

11. M. Ahrens, J. Ashwin, J. P. Calliess, V. Nguyen, Bayesian topic regression for causal inference. arXiv [Preprint] (2021). https://arxiv.org/abs/2109.05317 (Accessed 15 June 2023).
12. N. Egami, C. J. Fong, J. Grimmer, M. E. Roberts, B. M. Stewart, How to make causal inferences using texts. *Sci. Adv.* **8**, eabg2652 (2022).
13. J. Blumenau, B. E. Lauderdale, The variable persuasiveness of political rhetoric. *Am. J. Polit. Sci.* **68**, 255–270 (2022).
14. D. Bamman, B. O'Connor, N. Smith, Censorship and deletion practices in Chinese social media. *First Monday* **17**, 3–5 (2012).
15. G. King, J. Pan, M. E. Roberts, How censorship in China allows government criticism but silences collective expression. *Am. Polit. Sci. Rev.* **107**, 326–343 (2013).
16. Fu. Kw, Ch. Chan, M. Chau, Assessing censorship on microblogs in China: Discriminatory keyword analysis and the real-name registration policy. *IEEE Int. Comput.* **17**, 42–50 (2013).
17. A. Grattafiori *et al*., The llama 3 herd of models. arXiv [Preprint] (2024). https://arxiv.org/abs/2407.21783 (Accessed 21 August 2026).
18. Gemma Team *et al*., Gemma: Open models based on Gemini research and technology. arXiv [Preprint] (2024). https://arxiv.org/abs/2403.08295 (Accessed 21 August 2026).
19. T. Akiba *et al*., "A next-generation hyperparameter optimization framework" in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19*, A. Teredesai *et al*., Eds. (Association for Computing Machinery, New York, NY, USA, 2019), pp. 2623–2631.
20. V. Chernozhukov *et al*., Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21**, C1–C68 (2018).
21. J. Dorn, How much weak overlap can doubly robust t-statistics handle? arXiv [Preprint] (2025). https://arxiv.org/abs/2504.13273 (Accessed 27 February 2025).
22. S. M. Iacus, G. King, G. Porro, Causal inference without balance checking: Coarsened exact matching. *Polit. Anal.* **20**, 1–24 (2011).
23. M. E. Roberts, B. M. Stewart, E. M. Airoldi, A model of text for experimentation in the social sciences. *J. Am. Stat. Assoc.* **111**, 988–1003 (2016).
24. N. Reimers, I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks" in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, X. Wan, Eds. (Association for Computational Linguistics, 2019), pp. 3982–3992.
25. Y. Zhang *et al*., Qwen3 embedding: Advancing text embedding and reranking through foundation models. arXiv [Preprint] (2025). https://arxiv.org/abs/2506.05176. (Accessed 21 August 2026).
26. D. Won, Z. C. Steinert-Threlkeld, J. Joo, "Protest activity detection and perceived violence estimation from social media images" in *Proceedings of the 25th ACM International Conference on Multimedia*, Q. Liu *et al*., Eds. (Association for Computing Machinery, 2017), pp. 786–794.
27. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-Resolution Image Synthesis With Latent Diffusion Models (IEEE, 2022), pp. 10684–10695.
28. J. Jerit, How predictive appeals affect policy opinions. *Am. J. Polit. Sci.* **53**, 411–426 (2009).
29. L. Bos, W. Van Der Brug, C. H. De Vreese, An experimental test of the impact of style and rhetoric on the perception of right-wing populist and mainstream party leaders. *Acta Polit.* **48**, 192–208 (2013).
30. R. A. Bradley, M. E. Terry, Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* **39**, 324–345 (1952).
31. K. Nakamura, A. Breuer, M. H. Crespin, B. J. Dietrich, K. Imai, Causal inference with video features as treatments. arXiv [preprint] (2026). https://arxiv.org/abs/2607.06126 (Accessed 21 August 2026).
32. K. Nakamura, K. Imai, Genai powered dynamic causal inference with unstructured data. arXiv [Preprint] (2026). https://arxiv.org/abs/2605.07834 (Accessed 21 August 2026).
33. K. Nakamura, K. Imai, Replication data for: Leveraging generative artificial intelligence for causal inference with unstructured data. Harvard Dataverse. https://doi.org/10.7910/DVN/DX2PEX. Deposited 22 August 2026.