Political Analysis (2026), **0**, 1–9
doi:10.1017/pan.2026.10038

PA

LETTER

From Faces to Politics: Vision-Language Models (Sometimes) Link Visual Demographic Characteristics to Ideological Labels

Soyeon Jeon[1], Messi H. J. Lee[2], Jacob M. Montgomery[1] and Calvin K. Lai[3]

[1]Political Science, Washington University in St Louis, USA; [2]Division of Computational and Data Sciences, Washington University in St Louis, USA; [3]Department of Psychology, Rutgers University New Brunswick, USA

Corresponding author: Jacob M. Montgomery; Email: jacob.montgomery@wustl.edu

(Received 4 April 2025; revised 19 December 2025; accepted 20 December 2025)

Abstract

When foundation models analyze political content, do they use demographic characteristics as shortcuts for ideological attribution? We conducted detailed experiments with `GPT-4o-mini` and validated key findings across `GPT-4o` and `LLaVA`, using identical, ideologically neutral campaign advertisements with systematically varied candidate demographics. All models consistently attributed more liberal ideologies to women than men. These effects exceeded real-world gender differences from a nationally representative survey. However, racial associations differed by model: strong in `GPT-4o-mini` (where Black candidates received substantially more liberal attributions), attenuated in `GPT-4o`, and insignificant in `LLaVA`. These demographic effects persisted across temperature settings, prompt variations, and even explicit debiasing instructions in `GPT-4o-mini`. Our findings reveal that visual demographic features can shape AI outputs in ways that vary across models, with implications for applications such as content classification.

Keywords: vision-language model; artificial intelligence; stereotyping

Edited by: Adeline Lo

1. Introduction and Summary

Foundation models (FMs) are rapidly transforming political science research, offering unprecedented capabilities for analyzing political content (Heseltine and Clemm von Hohenberg 2024; Törnberg 2023; Waight *et al.* 2025; Weidmann, Faulborn, and García 2025). Given their widespread adoption, understanding how they process and create political information is crucial.

Extensive research documents that FMs perpetuate gender (e.g., Hofmann and Mõttus 2025; Rudinger *et al.* 2018; Zhao *et al.* 2018) and group-based stereotypes (e.g., Abid, Farooqi, and Zou 2021; Lee, Montgomery, and Lai 2024; Lucy and Bamman 2021).[1] These systems also exhibit liberal leanings when evaluated through standardized political orientation tests (e.g., Feng *et al.* 2023; Rozado 2023; Rozado 2024). However, these two streams of research have developed largely in isolation, and a critical gap remains: we do not understand whether or how demographic characteristics influence the way FMs attribute political characteristics. This gap has significant implications for political science research. For example, when using FMs for labeling tasks, demographic cues could systematically influence political coding in ways that researchers do not anticipate or desire, potentially introducing confounders into measures (Knox, Lucas, and Cho 2022).

[1]For a comprehensive survey, see Gallegos *et al.* (2024).

Unlike text-only models, vision-language models (VLMs) can process visual cues of race and gender in naturalistic stimuli and enable researchers to test political inferences of these stimuli in realistic contexts. By presenting identical political content while varying only the demographic features of individuals in images, we can isolate the effect of visual demographic information on political attributions.

Human perception research provides clear expectations for how demographic cues might influence political judgments. Decades of research demonstrate that people use demographic characteristics as heuristics for inferring political ideology: Black Americans are often assumed to be liberal, White Americans conservative, and women more left-leaning than men (Dolan 2014; Lerman and Sadin 2016). These associations significantly influence how political candidates are perceived and evaluated, even when their actual policy positions are identical (Crowder-Meyer *et al.* 2020; McDermott 1997; Sanbonmatsu 2002). Research on prototypicality, the stereotypical representation of features of their social group, suggests that more prototypical features strengthen category-based judgments (Burge, Wamble, and Cuomo 2020; Lemi 2021; Livingston and Brewer 2002; Ma, Correll, and Wittenbrink 2018; Maddox and Gray 2002). If FMs learn from human-generated data, we expect similar demographic-based ideology attributions.

To test whether and how VLMs link demographic characteristics to political ideology, we conducted detailed experiments with `GPT-4o-mini` and validated key findings across `GPT-4o` and `LLaVA` (Liu *et al.* 2023).[2] Using identical campaign advertisements that varied only in the race and gender of the featured candidate, we found that demographic-political associations were both systematic and model-dependent. In our primary analysis with `GPT-4o-mini`, Black candidates and women were consistently attributed more liberal ideologies than White candidates and men—patterns that persisted across temperature settings and prompt variations, and exceeded both random allocation baselines and real-world political identification patterns from nationally representative survey data. Validation tests revealed an important distinction: while the gender effect replicated across all three models, racial associations varied, weakening in `GPT-4o` and disappearing in `LLaVA`. Racial prototypicality showed no significant effect in any model, suggesting that when racial associations occur, models may process racial categories as discrete rather than continuous features.

2. Methods

2.1. Main Study Design

Figure 1 shows a visualization of the study design. The complete set of faces we used in this study is shown in Appendix I of the Supplementary Material. We conducted our primary analysis using `GPT-4o-mini`, with systematic replication in `GPT-4o` and `LLaVA`.[3] For all models, we presented identical campaign advertisements that varied only in the race and gender of the featured candidate. Each advertisement consisted of a facial image overlaid on a standardized Facebook campaign template containing neutral text ("Please endorse me for my election campaign, and I will send you a free window sticker. You can vote today!"). We prompted each model to generate captions or answer questions about the candidates' political views, specifically asking it to categorize each candidate as very conservative, conservative, moderate, liberal, or very liberal.[4]

To do this, we used 40 images from the GAN Face Database (GANFD; Marsden *et al.* 2024) representing four intersectional groups: Black women, Black men, White women, and White men. Within each racial group, we selected images with both high and low racial prototypicality ratings to examine whether the strength of racial features influenced political attributions. The GAN Face Database was generated using the StyleGAN2 model (Karras *et al.* 2020), which was trained on the

[2] Accessed from https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf.

[3] Several other popular VLMs, including Gemini, refused to answer questions about human faces, making them unsuitable for this research design.

[4] We provide additional details on our model selection criteria and prompt structure in Appendix B of the Supplementary Material.

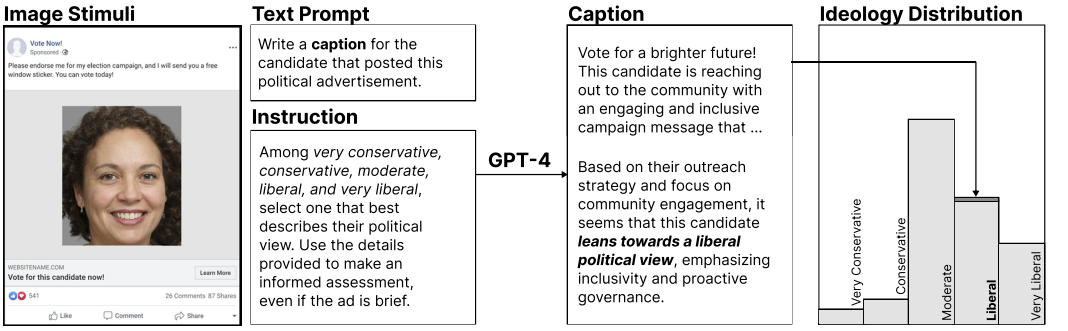


Figure 1. Visualization of the study design.

Flickr-Faces-HQ (FFHQ) dataset (Karras, Laine, and Aila 2019). A necessary limitation of this study is its reliance on synthetic images, which, despite allowing experimental control and overcoming important ethical and licensing challenges, may not fully capture the complex visual cues that AI models use to attribute ideology to real individuals.[5] We generated 50 completions for each image/prompt combination. This resulted in 8,000 total responses (50 iterations, 40 images, and 4 prompts).

After generating responses, we used `GPT-4o` to code political ideology from the generated text, achieving 100% accuracy compared to human coding in our pilot study. We assigned each response to a five-point ideological scale ranging from −2 (most conservative) to 2 (most liberal), excluding responses labeled as "refused to answer" (refusal rates varied by model; see Appendix C.3 of the Supplementary Material for details).

2.2. Baselines

To provide context to our results, we compared model outputs to five benchmarks. None represents a "true" or correct baseline. Instead, they offer different reference points to help interpret the magnitude and direction of any demographic-ideology associations in model outputs (additional details are in Appendix D of the Supplementary Material). Appendix J of the Supplementary Material briefly considers a sixth baseline of human labels.

1. *No-Picture Baseline:* The distribution of attributions assigned when no images were included to isolate the contribution of visual demographic cues.
2. *Equal Distribution:* Testing whether outputs deviate from uniform random assignment.
3. *Normal Distribution*: Testing whether outputs deviate from Gaussian random assignment centered at the moderate category.
4. *U.S. Population Data:* Overall political identification from the Weidenbaum Center Survey (WCS), a nationally representative sample collected with YouGov.
5. *U.S. Subgroup Data:* Race- and gender-specific distributions from the WCS, which included a large oversample of Black Americans.

2.3. Statistical Approach

We compared the distribution of attributed ideologies across the five categories (very conservative to very liberal) to our baselines using both descriptive statistics and χ^2 goodness-of-fit tests. These tests evaluated whether observed ideological distributions differed significantly from expected baseline

[5]One limitation of this work is potential confounding from non-demographic image features. Previous work shows that head orientation and expression predict ideology (Kosinski 2021); the GAN Face Database provides uniform direction, pose, and expression to control this, but other features may remain.

distributions, both for the overall sample and for specific demographic subgroups. We also examined *differences-in-differences*, comparing how the gap between demographic groups in model outputs differed from the corresponding gaps in baseline data. (Note that expected *group differences* are by construction zero except in the WCS benchmark; we do not expect different ideology by race or gender for the other benchmarks.)

Second, we fit linear models with random effects for each image. These models examined how race, gender, and their interaction influenced political ideology attributions on average while accounting for random variation between individual images. Again, we compare these reported differences to the WCS baseline, noting that the expected coefficients for the other baselines are zero.

2.4. Model Comparison and Robustness

We replicated our main analysis using `GPT-4o-2024-11-20` (henceforth `GPT-4o`) and the open-source `LLaVA` model to examine whether demographic-political associations generalize across architectures (accessed July 2025). For our primary model (`GPT-4o-mini`), we also conducted additional ablation experiments: a *Debias* condition with explicit instructions to avoid demographic assumptions, an *Order* condition randomizing ideology label sequence, and *Temperature* variations (0, 0.3, 0.6). Additional details are provided in Appendix H of the Supplementary Material.[6]

3. Results

3.1. Main Analysis

Figure 2 presents the difference between `GPT-4o-mini`'s ideological attributions and each benchmark distribution. Positive values (shaded green) indicate that the model assigned more candidates to that ideological category than expected under the baseline, while negative values (shaded red) indicate fewer assignments than expected. χ^2 statistics comparing the model outputs to the baseline distributions are shown on the right.

The pattern is clear and consistent across all benchmarks: `GPT-4o-mini` systematically over-attributes moderate, liberal, and very liberal ideologies while under-attributing conservative and very conservative ideologies. This pattern appears regardless of baseline comparison and is always statistically significant. Figure E.1 in the Supplementary Material shows that while the model shifts all demographic groups leftward relative to benchmarks, it does so most dramatically for women and Black candidates, effectively eliminating conservative attributions for these groups.

Figure 3 presents "difference-in-differences" analyses comparing demographic *gaps* in model attributions versus U.S. survey data. Note again that for other benchmarks, these differences are exactly zero. Panel (a) shows that while WCS data reveal modest gender differences (women are slightly more likely to identify as liberal), the model dramatically amplifies these patterns. Panel (b) reveals even more complicated racial patterns. While Black Americans do identify as more liberal than White Americans in WCS data, the model exaggerates these differences. The model fails to capture Black respondents' lower likelihood of identifying as conservative only because it assigns conservative labels so rarely to any group. Additionally, while Black WCS respondents are *more* likely than Whites to identify as moderate, the model reverses this pattern substantially.

Table 1 compares demographic effects in `GPT-4o-mini` outputs versus WCS benchmarks using linear models. For gender, the model's effect ($b = 0.41$, $SE = 0.09$, and $p < 0.05$) is nearly three times larger than the WCS benchmark ($b = 0.15$, $SE = 0.07$, and $p < 0.05$), indicating substantial relative amplification of gender-based political associations. The racial effect shows interesting convergence: while both model

[6] We adapted debiasing techniques from prior work (Chen *et al.* 2025; Furniturewala *et al.* 2024). Other relevant approaches include few-shot debiasing (Fatemi and Hu 2023), replication with temperature adjustment (Barrie, Palmer, and Spirling 2024), fine-tuning, prompt engineering, adversarial learning, and few-shot prompting (Abid *et al.* 2021; Fatemi and Hu 2023; Gallegos *et al.* 2024; Mattern *et al.* 2022; Narayanan Venkit 2023; Yang *et al.* 2024).

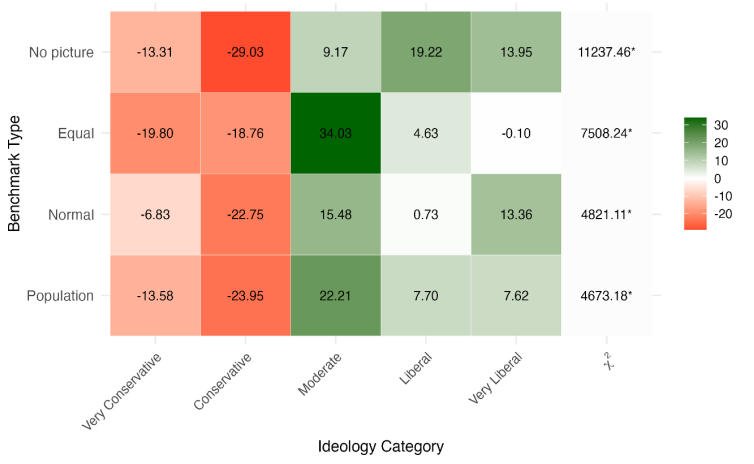


Figure 2. Percentage point differences between `GPT-4o-mini` ideological attributions and various benchmarks. Green indicates the model assigns more candidates to that ideological category relative to the benchmark; red indicates fewer. χ^2 tests compare overall distributions between model outputs and each benchmark. An asterisk denotes statistical significance at $p < 0.05$.

Table 1. Effect of predictors on ideology in `GPT-4o-mini` and WCS data.

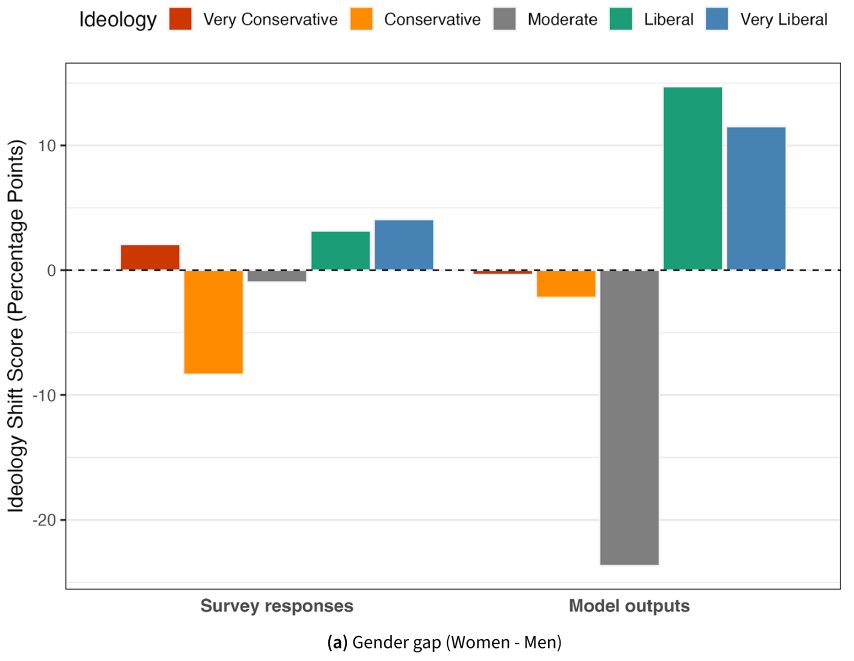
	WCS gender	GPT gender	WCS race	GPT race	WCS both	GPT both
Intercept	−0.19*	0.42*	−0.19*	0.40*	−0.25*	0.18*
	(0.04)	(0.06)	(0.04)	(0.06)	(0.05)	(0.05)
Women	0.15*	0.41*			0.13	0.46*
	(0.07)	(0.09)			(0.08)	(0.07)
Black			0.47*	0.44*	0.37*	0.49*
			(0.06)	(0.08)	(0.08)	(0.07)
Black × Women					0.21	−0.10
					(0.11)	(0.10)
N	2,657	7,809	2,657	7,809	2,657	7,809

Note: Standard errors in parentheses. An asterisk denotes statistical significance at $p < 0.05$. We used image random effects for GPT and robust standard errors for WCS models.

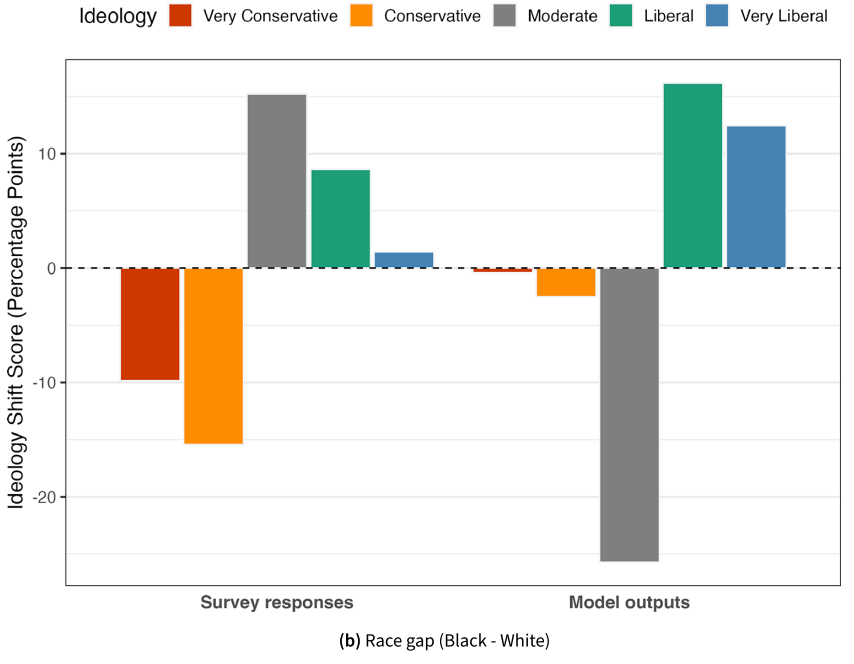
($b = 0.44$, $SE = 0.08$, and $p < 0.05$) and WCS ($b = 0.47$, $SE = 0.06$, and $p < 0.05$) data show similar magnitudes, this masks important differences in how these effects manifest across the ideological spectrum (as shown in Figure 3). When examining intersectional effects, neither model nor WCS data show significant race-by-gender interactions, suggesting these demographic effects operate additively rather than multiplicatively. Racial prototypicality showed no significant effects in any specification (see Appendix F of the Supplementary Material).

3.2. Robustness Analyses

Our replication across `GPT-4o` and `LLaVA` revealed both consistencies and important differences in how models link demographics to political ideology. All three models showed systematic liberal overattribution relative to baselines and consistent gender effects: women received more liberal attributions than men across all architectures (see Table 1 and Appendix G of the Supplementary Material). However, racial associations varied by model. While `GPT-4o-mini` showed strong racial effects (Black



(a) Gender gap (Women - Men)



(b) Race gap (Black - White)

Figure 3. Difference-in-differences analysis comparing demographic gaps in `GPT-4o-mini` outputs versus WCS survey data. Positive values indicate women (panel a) or Black candidates (panel b) receive higher percentages in that ideological category than men or White candidates, respectively. The model amplifies demographic differences relative to baselines, particularly in liberal/very liberal categories, while creating new gaps (e.g., gender differences in moderate attributions) absent from survey data.

candidates received 0.44 points more liberal on our scale), these effects attenuated in `GPT-4o` (0.18 points) and disappeared in `LLaVA`, where racial differences failed to reach statistical significance. This model-dependent pattern suggests that while gender-ideology associations appear robust across vision-language architectures, racial associations are contingent on specific model implementations.

Table 2. Liberal attribution effects across `GPT-4o-mini` ablation conditions.

Effect	Main	WCS	Debias	Order	Temp 0	Temp 0.3	Temp 0.6
Women	0.41 (0.09)	0.15 (0.07)	0.30 (0.07)	0.26 (0.05)	0.28 (0.10)	0.34 (0.08)	0.39 (0.08)
Black	0.44 (0.08)	0.47 (0.06)	0.35 (0.06)	0.24 (0.06)	0.46 (0.09)	0.37 (0.08)	0.37 (0.08)

Note: All results are significant at $p < 0.05$. We used robust standard errors for the WCS survey data. Standard errors in parentheses. Effects show an increase in liberal attribution for women vs. men and Black vs. White. Main = original results; WCS = survey benchmark; *Debias* = prompt to ignore demographics; *Order* = reversed options; *Temp* = temperature settings. Full model results are shown in Appendix H of the Supplementary Material.

Table 2 presents ablation studies examining the robustness of demographic-ideology associations in `GPT-4o-mini` across various conditions. In the *Debias* condition, we instructed the model to disregard demographic characteristics when inferring political ideology. While this instruction reduced the association, it did not eliminate it; women and Black candidates were still significantly more likely to be labeled as liberal. Similarly, in the *Order* condition, we reversed the sequence of ideological options presented in the prompt. This change, aimed at testing for ordering effects, also attenuated the associations but failed to remove them. We also tested the model's consistency across various temperature settings (0, 0.3, and 0.6), which control output randomness. The demographic associations remained remarkably stable.

4. Discussion and Implications

VLMs (sometimes) link demographics to political ideologies, even when analyzing ideologically neutral content. Persistent gender associations exceeded real-world benchmarks, whereas racial effects were model-dependent: strong in `GPT-4o-mini`, attenuated in `GPT-4o`, and insignificant in `LLaVA`. For `GPT-4o-mini`, where effects were most pronounced, demographic associations persisted even with explicit debiasing instructions, suggesting these patterns can be resistant to simple prompt-based interventions (although more successful debiasing variations surely exist).

Limitations include the use of 40 images, three models, and one political concept (ideology). Although we collected 8,000 responses for each model, we may not have sufficient power to detect small interaction effects between race and prototypicality. Perhaps more concerning, the variation we observed across models—with racial associations strong in `GPT-4o-mini`, attenuated in `GPT-4o`, and absent in `LLaVA`—reveals how demographic-political associations can vary unpredictably across architectures. This instability means researchers cannot know ex ante whether their chosen model will exhibit such associations, particularly for proprietary models where training data and architectural choices remain opaque. Future work should expand to more images, additional models, and other political outputs. As we used only English prompts and American framing of ideology, these results would also benefit from replications across cultures.

Despite limitations, our findings have important implications. Using VLMs for political content analysis may unknowingly introduce model-specific demographic confounds. To avoid this, we recommend excluding images from inputs unless appearance is central to the research question. If images are necessary, researchers should benchmark their chosen model's performance beforehand. The open-source `LLaVA` model displayed the weakest racial bias, which suggests it may work as a reasonable default. However, this finding must be interpreted cautiously given its notably high refusal rate, which suggests a potential trade-off between reduced bias and model usability. As VLM architectures evolve, ongoing validation and monitoring of demographic effects will be essential.

More broadly, as VLMs integrate into our political information infrastructure, these associations risk reshaping political communication in subtle but consequential ways. Search engines using AI-generated summaries might systematically mischaracterize candidates' ideologies based on appearance alone, while automated journalism tools could inject these patterns into political coverage. As AI becomes more fully embedded in how we create, distribute, and consume political information, researchers

and practitioners must carefully attend to how these systems encode associations between visual demographic characteristics and political attributes.

Supplementary Material. For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2026.10038.

Data Availability Statement. Replication code for this article has been published in the Harvard Dataverse at https://doi.org/10.7910/DVN/2JMT9J (Jeon *et al.* 2026).

Acknowledgements. The authors thank three anonymous reviewers and Michael Strawbridge for helpful comments. The authors used AI tools (ChatGPT, Claude, and Gemini) during writing, revision, and coding tasks. These tools were accessed in 2024 and 2025 and used with/without modification. The authors reviewed and verified all outputs. All errors are our own.

Funding Statement. The authors declare that no specific funding has been received for this article.

Competing Interests. The authors declare none.

References

Abid, A., M. Farooqi, and J. Zou. 2021. "Persistent Anti-Muslim Bias in Large Language Models." In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 298–306.

Barrie, C., A. Palmer, and A. Spirling. 2024. "Replication for Language Models Problems, Principles, and Best Practice for Political Science." URL: https://arthurspirling.org/documents/BarriePalmerSpirlingTrustMeBro.pdf.

Burge, C. D., J. J. Wamble, and R. R. Cuomo. 2020. "A Certain Type of Descriptive Representative? Understanding how the Skin Tone and Gender of Candidates Influences Black Politics." *The Journal of Politics* 82 (4): 1596–1601.

Chen, Y., V. C. Raghuram, J. Mattern, R. Mihalcea, and Z. Jin. 2025. "Causally Testing Gender Bias in LLMs: A Case Study on Occupational Bias." In *Findings of the Association for Computational Linguistics: NAACL 2025*, edited by L. Chiruzzo, A. Ritter, and L. Wang, 4984–5004. Albuquerque, New Mexico: Association for Computational Linguistics.

Crowder-Meyer, M., S. K. Gadarian, J. Trounstine, and K. Vue. 2020. "A Different Kind of Disadvantage: Candidate Race, Cognitive Complexity, and Voter Choice." *Political Behavior* 42 (2): 509–530.

Dolan, K. A. 2014. *When Does Gender Matter?: Women Candidates and Gender Stereotypes in American Elections*. Oxford University Press.

Fatemi, S., and Y. Hu. 2023. "A Comparative Analysis of Fine-Tuned LLMs and Few-Shot Learning of LLMs for Financial Sentiment Analysis." Preprint, arXiv:2312.08725.

Feng, S., C. Y. Park, Y. Liu, and Y. Tsvetkov. 2023. "From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models." In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by A. Rogers, J. Boyd-Graber, and N. Okazaki, 11737–11762. Toronto, Canada: Association for Computational Linguistics.

Furniturewala, S., et al. 2024. ""Thinking" Fair and Slow: On the Efficacy of Structured Prompts for Debiasing Language Models." In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, edited by Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, 213–227. Miami, Florida: Association for Computational Linguistics.

Gallegos, I. O., et al. 2024. "Bias and Fairness in Large Language Models: A Survey." *Computational Linguistics* 50 (3): 1097–1179.

Heseltine, M., and B. C. von Hohenberg. 2024. "Large Language Models as a Substitute for Human Experts in Annotating Political Text." *Research & Politics* 11 (1): 20531680241236239.

Hofmann, R., and R. Mõttus. 2025. "Does Speaking a Gendered Language Make You a Gendered Being? Gender Differences in Personality are Associated with Linguistic Gender Differences Across 49 Languages."

Jeon, S., M. H. J. Lee, J. M. Montgomery, and C. K. Lai. 2026. "(Replication Data For: From Faces to Politics: Vision-Language Models (Sometimes) Link Visual Demographic Characteristics to Ideological Labels)." https://doi.org/10.7910/DVN/2JMT9J

Karras, T., S. Laine, and T. Aila. 2019. "A Style-Based Generator Architecture for Generative Adversarial Networks." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4401–4410.

Karras, T., S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. 2020. "Analyzing and Improving the Image Quality of StyleGAN." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8110–8119.

Knox, D., C. Lucas, and W. K. T. Cho. 2022. "Testing Causal Theories with Learned Proxies." *Annual Review of Political Science* 25 (1): 419–441.

Kosinski, M. 2021. "Facial Recognition Technology Can Expose Political Orientation from Naturalistic Facial Images." *Scientific Reports* 11 (1): 100.

Lee, M. H. J., J. M. Montgomery, and C. K. Lai. 2024. "Large Language Models Portray Socially Subordinate Groups as more Homogeneous, Consistent with a Bias Observed in Humans." In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1321–1340.

Lemi, D. C. 2021. "Do Voters Prefer Just any Descriptive Representative? The Case of Multiracial Candidates." *Perspectives on Politics* 19 (4): 1061–1081.

Lerman, A. E., and M. L. Sadin. 2016. "Stereotyping or Projection? How White and Black Voters Estimate Black Candidates' Ideology." *Political Psychology* 37 (2): 147–163.

Liu, H., C. Li, W. Qingyang, and Y. J. Lee. 2023. "Visual Instruction Tuning." *Advances in Neural Information Processing Systems* 36: 34892–34916.

Livingston, R. W., and M. B. Brewer. 2002. "What Are we Really Priming? Cue-Based Versus Category-Based Processing of Facial Stimuli." *Journal of Personality and Social Psychology* 82 (1): 5–18.

Lucy, L., and D. Bamman. 2021. "Gender and Representation Bias in GPT-3 Generated Stories." In *Proceedings of the Third Workshop on Narrative Understanding*, edited by N. Akoury, F. Brahman, S. Chaturvedi, E. Clark, M. Iyyer, and L. J. Martin, 48–55. Virtual: Association for Computational Linguistics.

Ma, D. S., J. Correll, and B. Wittenbrink. 2018. "The Effects of Category and Physical Features on Stereotyping and Evaluation." *Journal of Experimental Social Psychology* 79: 42–50.

Maddox, K. B., and S. A. Gray. 2002. "Cognitive Representations of Black Americans: Reexploring the Role of Skin Tone." *Personality and Social Psychology Bulletin* 28 (2): 250–259.

Marsden, A. D., A. Jaurique, M. L. McDonald, and S. E. Burke. 2024. "(GAN Face Database (GANFD))."

Mattern, J., Z. Jin, M. Sachan, R. Mihalcea, and B. Schölkopf. 2022. "Understanding Stereotypes in Language Models: Towards Robust Measurement and Zero-Shot Debiasing." Preprint, arXiv:2212.10678.

McDermott, M. L. 1997. "Voting Cues in Low-Information Elections: Candidate Gender as a Social Information Variable in Contemporary United States Elections." *American Journal of Political Science* 41 (1): 270–283.

Narayanan Venkit, P. 2023. "Towards a Holistic Approach: Understanding Sociodemographic Biases in NLP Models Using an Interdisciplinary Lens." In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 1004–1005.

Rozado, D. 2023. "The Political Biases of ChatGPT." *Social Sciences* 12 (3): 148.

Rozado, D. 2024. "The Political Preferences of LLMs." *PLoS One* 19 (7): e0306621.

Rudinger, R., J. Naradowsky, B. Leonard, and B. Van Durme. 2018. "Gender Bias in Coreference Resolution." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 8–14.

Sanbonmatsu, K. 2002. "Gender Stereotypes and Vote Choice." *American Journal of Political Science* 46 (1): 20–34.

Törnberg, P. 2023. "ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning." Preprint, arXiv:2304.06588.

Waight, H., et al. 2025. "Quantifying Narrative Similarity across Languages." *Sociological Methods & Research* 00491241251340080.

Weidmann, N. B., M. Faulborn, and D. García. 2025. "Large Language Models are Democracy Coders with Attitudes." *PS: Political Science & Politics* 1–7.

Yang, N., T. Kang, J. Choi, H. Lee, and K. Jung. 2024. "Mitigating Biases for Instruction-Following Language Models via Bias Neurons Elimination." In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9061–9073.

Zhao, J., T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. 2018. "Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, edited by M. Walker, H. Ji, and A. Stent, 15–20. New Orleans, Louisiana: Association for Computational Linguistics.

Cite this article: Jeon, S., Lee, M. H. J., Montgomery, J. M. and Lai, C. K., (2026). From Faces to Politics: Vision-Language Models (Sometimes) Link Visual Demographic Characteristics to Ideological Labels. *Political Analysis*, 1–9. https://doi.org/10.1017/pan.2026.10038