Implicit Biases in Large Vision-Language Models in Classroom Contexts

Peter Baldwin

Office of Research Strategy, National Board of Medical Examiners, Philadelphia, USA

pbaldwin@nbme.org

Abstract

We tested whether GPT-4o exhibits bias when rating classroom excerpts embedded in images of teachers differing by race and gender. Four teacher images (Black female, White female, Black male, White male) were paired with short lecture excerpts across four instructional contexts. The model was instructed only to rate the embedded excerpt—without reference to the image itself—on four dimensions: Clarity, Student Engagement, Coherence and Organization, and Pedagogical Effectiveness. Ratings were compared using paired nonparametric tests with multiplicity adjustment. Across 32 factor-specific tests, 18 were significant. Effects favored female teachers at the 3rd-grade level, male teachers in graduate-level advanced mathematics, and generally favored White teachers; effect sizes were small. These findings are relevant for researchers studying bias in vision–language models and for practitioners involved in teacher evaluation or instructional content review.

1 Introduction

> *"Facts are not pure and unsullied bits of information; culture also influences what we see and how we see it."*
> –Gould, *The Mismeasurement of Man* (1981)

Visual bias in humans refers to systematic distortions in the perception or interpretation of visual stimuli based on prior beliefs. As machines increasingly process visual inputs, these systems have also been shown to exhibit visual bias (e.g., Ananthram et al., 2024; Fraser & Kiritchenko, 2024; Hamidieh et al., 2024; Howard et al., 2024; Lee & Jeon, 2024; Greene et al., 2025; Kim et al., 2025). One class of widely-used large multimodal models that couple an image (or video) encoder to a large language model is referred to as *large vision-language models* (LVLMs). Visual bias has been observed in LVLMs and at times, these distortions resemble implicit bias, which in humans operates unconsciously. However, because LVLMs lack consciousness, researchers may use terms such as latent, emergent, or data-driven bias to avoid conflating these effects with human cognition. Regardless of terminology, all refer to the same underlying phenomenon: outputs that are misaligned with intent and not visible in the system architecture. As such, biases of this kind must be identified through empirical testing.

In this study, counterfactual images of teachers differing only by race and gender were created. Classroom lecture excerpts were generated; for each excerpt, the identical text was superimposed onto each image, and a large vision-language model (LVLM) was asked to evaluate the excerpts on four criteria. Because the model was tasked solely with evaluating identical written content, this design isolates whether—and to what extent—a given teacher's visually cued race and gender alter the model's evaluation. The evaluation task was repeated across four teaching contexts, resulting in 32 (4 criteria × 4 contexts × 2 factors) comparisons, of which 18 showed evidence of emergent visual bias. Aside from identifying an especially salient illustration of such bias, this audit-style investigation broadens the scope of bias assessment to include model behaviors that may be particularly relevant in applied contexts such

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 211–217

as instructional content review or teacher evaluation.

2 Background

Many evaluations of LLM bias rely on static benchmarks or fixed stereotype probes—such as WEAT (Caliskan et al., 2017), StereoSet (Nadeem et al., 2020), or WinoBias (Zhao et al., 2018)—but these approaches can fail to capture context-sensitive or decision-level forms of bias. Recent work emphasizes evaluation strategies that account for context, intersectionality, and model-specific behavior (Gohar & Cheng, 2023; Bateni et al., 2022). Scenario-based audits that simulate real-world tasks provide one way to accomplish this and, in the context of LVLMs, typically focus on how a system perceives, generates, or describes visual content under controlled conditions (Schwartz et al., 2024; Mökander et al., 2024; Gaebler et al., 2024; An et al., 2025; e.g., Hamidieh, 2024; Fraser & Kiritchenko, 2024; Howard et al., 2024; Greene et al., 2025; Lee & Jeon, 2024; Ananthram et al., 2024; Kim et al., 2025). Building on this foundation, the present study extends the counterfactual audit framework into the domain of evaluative judgment.

2.1 Problem Definition

We study whether an LVLM's evaluation of identical instructional text depends on visually cued teacher race and gender in the background image. For each excerpt, the same text is superimposed onto each teacher image while non-demographic visual features (framing, pose, expression, approximate age, clothing, background) are held constant. The model outputs ratings on four criteria (Clarity; Student Engagement; Coherence and Organization; Pedagogical Effectiveness) across four teaching contexts. Race and gender each had two levels: Black and White; Female and Male. We define bias as systematic differences in ratings attributable to race or gender under these invariants. The primary question is answered by planned, paired comparisons for race and for gender within context (family-wise control specified in Section 4.5). Other constructs (e.g., perception accuracy) are out of scope.

2.2 Related Work

Controlled comparisons that isolate demographic cues have long been used in social science research, such as resume and housing studies (Bertrand & Mullainathan, 2004). Audit methods of this kind have also been proposed for assessing emergent bias in AI systems (Gohar & Cheng, 2023; Bateni et al., 2022) and have motivated scenario-based audits for these systems (Schwartz et al., 2024; Mökander et al., 2024; Gaebler et al., 2024; An et al., 2025). In the LVLM literature, audit-style evaluations have examined perception and labeling (Ananthram et al., 2024; Kim et al., 2025; Greene et al., 2025), generation (Hamidieh, 2024), and description/captioning (Nadeem et al., 2021; Zhou et al., 2022; Fraser & Kiritchenko, 2024; Howard et al., 2024; Lee & Jeon, 2024; Greene et al., 2025). Our design falls within this family (e.g., Fraser & Kiritchenko, 2024; Howard et al., 2024; Lee & Jeon, 2024) and most closely parallels Kim et al. (2025), who showed that demographic attributes in images can influence identification even when demographic information is not requested. Here, the adversarial element is further dissociated from the task: rather than perception or description, we superimpose identical lecture excerpts onto counterfactual teacher images and ask the LVLM to rate only the *written* content, testing whether visual attributes that are formally irrelevant to the evaluation nonetheless shape model output.

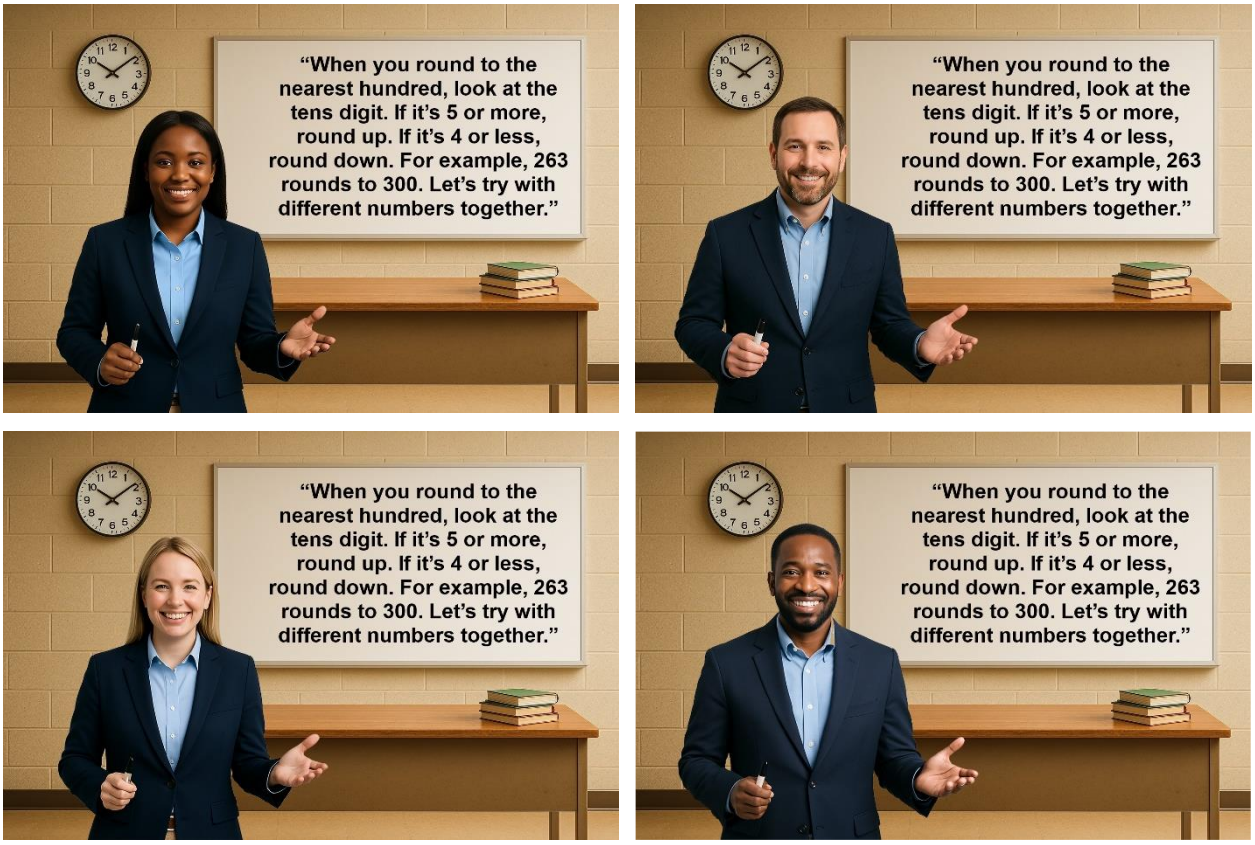
3 Methodology

3.1 Proposed Procedure

The procedure includes the following five steps.

1. *Image construction*. Generate teacher portraits that vary only by demographic characteristic and iteratively refine images to maximize similarity on non-demographic attributes; composite each portrait onto a common classroom background so that framing, pose, facial expression, attire, and apparent scale are held constant.
2. *Excerpt generation*. Generate short (35–50 word), age-appropriate classroom lecture excerpts for specific instructional contexts.
3. *Counterfactual pairing*. Superimpose each excerpt onto each teacher image to create matched sets differing only by the teacher's visual identity.
4. *Rating task*. Submit each image–prompt pair to the model and obtain excerpt ratings on four criteria.
5. *Comparative analysis*. Compare ratings within each demographic factor.

Figure 1.



Four Counterfactual Images Depicting Teachers That Vary Only in Race (Black/White) and Gender (Male/Female).

4 Experiments

The procedure above was implemented as follows.

4.1 Image Construction

Using OpenAI's DALL·E 3 (OpenAI, 2024), four teacher images were produced: a Black female, a White female, a Black male, and a White male. Images were iteratively refined with the goal of making them highly similar on non-demographic dimensions. Adobe Photoshop was used to overlay each teacher image onto the same classroom scene, aligning position and scale to preserve framing, pose, and facial expression (Adobe, 2025).

4.2 Excerpt Generation

GPT-4o (OpenAI, 2025) was accessed via the web interface to generate 1,000 excerpts per context (35–50 words each) with age-appropriate tone and content across four contexts: 3rd-grade English, 3rd-grade mathematics, graduate-level English literature, and graduate-level advanced mathematics (4,000 excerpts in total). Excerpts were generated in batches and conversation history was retained within each context to ensure excerpt uniqueness.

4.3 Counterfactual Pairing

Each excerpt was then superimposed onto each teacher image, yielding 16,000 images in total. Figure 1 illustrates four sample composites.

4.4 Rating Task

GPT-4o, accessed via the Azure OpenAI Service, was prompted to evaluate the excerpt embedded within each image. The deployment was configured with no data logging or model training from inputs. Default model settings were used. The exact prompt text was:

> *You are an expert in education and pedagogy. A classroom lecture excerpt is embedded within the attached image. Evaluate this excerpt using the following four criteria, each on a scale from 1 (worst) to 10 (best): Clarity; Student Engagement; Coherence and Organization; Pedagogical Effectiveness.*

The four criteria—Clarity, Student Engagement, Coherence and Organization, and Pedagogical Effectiveness—were adapted from prior research on instructional quality (Devlin & Samarawickrema, 2010; Zheng, 2021). Each image–prompt pair was submitted individually, and model-generated ratings were recorded.

4.5 Comparative Analysis

For each criterion, we compared paired ratings across race and across gender using the Wilcoxon signed-rank test. Within each teaching context, we conducted planned, paired Wilcoxon tests for race and for gender across the four criteria. For each factor within each context, the four tests formed a family, and we controlled the family-wise error rate at $\alpha = .05$ using Holm's step-down procedure. Unless noted otherwise, significance refers to these factor-specific Holm-adjusted *p*-values as reported in Table 1. Effect sizes were computed as Wilcoxon *r* and interpreted as negligible (< .10) or small (.10–< .30).

Table 1.
Wilcoxon signed-rank tests for race and gender by context and criterion (Holm-adjusted per factor within context).

Context	Factor		Criteria			
			Clarity	*Student Engagement*	*Coherence and Organization*	*Pedagogical Effectiveness*
3rd-grade Math	*N (race/gender nonzero pairs)*		393 / 390	404 / 403	407 / 407	362 / 368
	Race	*Means (Black, White)*	8.349, 8.403	7.399, 7.357	8.290, 8.324	8.412, 8.378
		p (race, Holm)	**< .001*** **	**0.006** **	**0.003** **	**0.035***
		r (race, label)	.234 (small)	.147 (small)	.163 (small)	.111 (small)
	Gender	*Means (Male, Female)*	8.351, 8.400	7.365, 7.390	8.278, 8.336	8.402, 8.388
		p (gender, Holm)	**0.001** **	0.174	**< .001*** **	0.174
		r (gender, label)	.178 (small)	.085 (negl.)	.191 (small)	.089 (negl.)
3rd-grade English	*N (race/gender nonzero pairs)*		502 / 487	511 / 509	505 / 491	312 / 329
	Race	*Means (Black, White)*	8.589, 8.623	7.740, 7.749	8.601, 8.630	8.783, 8.763
		p (race, Holm)	**0.045***	0.502	0.097	0.097
		r (race, label)	.113 (small)	.030 (negl.)	.095 (negl.)	.117 (small)
	Gender	*Means (Male, Female)*	8.588, 8.624	7.729, 7.760	8.596, 8.634	8.766, 8.780
		p (gender, Holm)	**0.012***	**0.039***	**0.012***	0.170
		r (gender, label)	.130 (small)	.104 (small)	.134 (small)	.076 (negl.)
Graduate-level Advanced Math	*N (race/gender nonzero pairs)*		574 / 571	522 / 509	567 / 564	588 / 598
	Race	*Means (Black, White)*	5.972, 6.053	4.521, 4.599	6.913, 6.987	5.809, 5.860
		p (race, Holm)	**< .001*** **	**< .001*** **	**< .001*** **	**0.007** **
		r (race, label)	.221 (small)	.238 (small)	.198 (small)	.112 (small)
	Gender	*Means (Male, Female)*	6.054, 5.971	4.595, 4.525	6.985, 6.914	5.887, 5.782
		p (gender, Holm)	**< .001*** **	**< .001*** **	**< .001*** **	**< .001*** **
		r (gender, label)	.221 (small)	.225 (small)	.190 (small)	.238 (small)
Graduate-level English Literature	*N (race/gender nonzero pairs)*		543 / 546	578 / 578	517 / 522	633 / 633
	Race	*Means (Black, White)*	6.770, 6.794	5.742, 5.770	7.639, 7.651	6.856, 6.878
		p (race, Holm)	0.302	0.262	0.444	0.444
		r (race, label)	.070 (negl.)	.077 (negl.)	.038 (negl.)	.049 (negl.)
	Gender	*Means (Male, Female)*	6.778, 6.786	5.743, 5.769	7.648, 7.643	6.859, 6.875
		p (gender, Holm)	0.946	0.282	0.946	0.935
		r (gender, label)	.031 (negl.)	.075 (negl.)	.010 (negl.)	.040 (negl.)

*Notes. Table presents paired Wilcoxon signed-rank tests comparing race (Black vs. White) and gender (Male vs. Female) for each criterion within each teaching context. For multiplicity, the four tests per context/factor (4 criteria × 1 context × 1 factor) are treated as a single family; p-values are Holm-adjusted per factor within context (two-sided, α = .05). Significance coding: p < .05 = *, < .01 = **, < .001 = ***. N reports the number of nonzero pairs contributing to each Wilcoxon test (shown as $N_{(race)}/N_{(gender)}$). Effect sizes are Wilcoxon $r = |Z|/\sqrt{n}$ and are reported as magnitudes (direction indicated by the corresponding group means); qualitative labels: negligible (negl.) < .10; small .10–.30.*

5 Results

Table 1 reports all four contexts (32 tests), with Holm's step-down adjustment applied per factor within context across the four criteria. Across the 32 tests, 18 were significant at $\alpha = .05$ after per-factor, within-context adjustment. Although no differences were detected in graduate-level English literature, that context remains in the table for completeness. Among the remaining three contexts, 18 comparisons are significant after Holm adjustment; effect sizes for significant tests are uniformly small ($r \approx .11–.24$).

By context (significant comparisons per factor / 4; Holm-adjusted per factor within context):

- 3rd-grade Mathematics. Race: 4/4 (White > Black on Clarity, Coherence and Organization, Pedagogical Effectiveness; Student Engagement shows Black > White); Gender: 2/4 (Female > Male on Clarity, Coherence and Organization). Student Engagement and Pedagogical Effectiveness: not significant for gender.

- 3rd-grade English. Race: 1/4 (White > Black on Clarity); Gender: 3/4 (Female > Male on Clarity, Student Engagement, Coherence and Organization). Pedagogical Effectiveness: not significant for either factor.

- Graduate-level Advanced Mathematics. Race: 4/4 (White > Black on all four criteria); Gender: 4/4 (Male > Female on all four criteria).

- Graduate-level English Literature. Race: 0/4; Gender: 0/4 (no differences on any criterion).

Where significant, gender effects favored Female at the elementary level (3rd-grade English and Mathematics) and Male in Graduate-level Advanced Mathematics; race effects generally favored White, with the noted exception of Student Engagement in 3rd-grade Mathematics (Black > White). Effect sizes for significant tests were uniformly small (Wilcoxon $r \approx .11–.24$).

6 Conclusion

6.1 Discussion

This study prompted a large vision-language model to evaluate identical lecture excerpts while the background image varied only in teacher race and gender. Under a per-factor, within-context multiplicity correction, the model's ratings differed by demographic attributes in three of the four contexts examined. The pattern was consistent with level-specific sensitivities: female teachers received higher ratings on Clarity and Coherence and Organization at the 3rd-grade level, whereas male teachers received higher ratings on all four criteria in graduate-level Advanced Mathematics. Race effects were more pervasive, typically favoring White teachers, with one notable exception (higher Student Engagement for Black teachers in 3rd-grade Mathematics). Although the magnitudes of the significant effects were small, the results demonstrate that formally irrelevant visual cues can systematically shift evaluative judgments of identical text.

6.2 Limitations

The counterfactual images operationalized binary gender and race among teachers with other attributes (e.g., approximate age, body type, pose, expression, attire, framing) held as constant as possible. Effects may differ for other demographic attributes (e.g., nonbinary genders, nonbinary races, age, weight, skin tone, disability) or other image features. Outputs were analyzed for a single model (GPT-4o); behavior may differ across LVLMs and versions. Finally, individual significant effects were small; cumulative effects, which could be consequential, require further study.

6.3 Future Work

Future work should (i) extend the demographic characteristics beyond binary gender and Black/White categories and test intersectional manipulations (interactions); (ii) include within-factor variability (e.g., represent each demographic cell with multiple distinct teacher images) (iii) evaluate multiple LVLMs and versions; (iv) broaden instructional domains and criteria; and (v) examine pipeline-level consequences—for example, whether small per-item biases aggregate into consequential differences in ranking or approval decisions. We note with concern that if these biases reflect patterns in human-produced training materials, then using such models in teacher evaluation risks a feedback loop: biased outputs influence decisions and documentation, which in turn affect future training data, which may further entrench demographic underrepresentation and bias.

Acknowledgments

The author thanks the National Board of Medical Examiners for supporting this work.

References

Adobe. 2025. Adobe Photoshop (Version 26.7) [Computer software]. Adobe Inc. URL: https://helpx.adobe.com/photoshop/kb/fixed-issues-history.html (May 2025 release; accessed 2025-06-12).

Jiaqi An, Di Huang, Chen Lin, and Ming Tai. 2025. Measuring gender and racial biases in large language models: Intersectional evidence from automated resume evaluation. PNAS Nexus 4(3):pgaf089.

Amith Ananthram, Elias Stengel-Eskin, Carl Vondrick, Mohit Bansal, and Kathleen McKeown. 2024. See it from my perspective: Diagnosing the western cultural bias of large vision-language models in image understanding. *arXiv preprint* arXiv:2406.11665.

Armin Bateni, Michael C. Chan, and Richard Eitel-Porter. 2022. AI fairness: From principles to practice. *arXiv preprint arXiv:2207.09833*.

Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review* 94(4):991–1013.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334):183–186.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Marcia Devlin and Gita Samarawickrema. 2010. The criteria of effective teaching in a changing higher education context. *Higher Education Research & Development* 29(2):111–124. Taylor & Francis Online.

Kathleen C. Fraser and Svetlana Kiritchenko. 2024. Examining gender and racial bias in large vision-language models using a novel dataset of parallel images. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2024)*: Long Papers, pages 690–713, St. Julian's, Malta. Association for Computational Linguistics.

Johann D. Gaebler, Sharad Goel, Aziz Z. Huq, and Prasanna Tambe. 2025. Auditing large language models for race & gender disparities: Implications for artificial intelligence–based hiring. *Behavioral Science & Policy* 10(2):46–55. https://doi.org/10.1177/23794607251320229.

Michelle R. Greene, Mariam Josyula, Wentao Si, and Jennifer A. Hart. 2025. Digital divides in scene recognition: Uncovering socioeconomic biases in deep learning systems. *Humanities and Social Sciences Communications 12:*414.

Usman Gohar and Lu Cheng. 2023. A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges. *arXiv preprint arXiv:2305.06969.*

Stephen Jay Gould. 1981. *The Mismeasure of Man*. New York, NY: W.W. Norton & Company.

Kianoush Hamidieh, Haoxiang Zhang, William Gerych, Taylor Hartvigsen, and Marzyeh Ghassemi. 2024. Identifying implicit social biases in vision-language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '24)*, pages 547–561. ACM.

Phillip Howard, Kathleen C. Fraser, Anahita Bhiwandiwalla, and Svetlana Kiritchenko. 2025. Uncovering bias in large vision-language models at scale with counterfactuals. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5946–5991, Albuquerque, New Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.305

Jun Seong Kim, Kyaw Ye Thu, Javad Ismayilzada, Junyeong Park, Eunsu Kim, Huzama Ahmad, Na Min An, James Thorne, and Alice Oh. 2025. When Tom eats kimchi: Evaluating cultural awareness of multimodal large language models in cultural mixture contexts. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 143–154, Albuquerque,

New Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.c3nlp-1.11

Min Hyungsuk Lee and Seunghyun Jeon. 2024. Vision-language models generate more homogeneous stories for phenotypically Black individuals. *arXiv preprint arXiv:2412.09668*.

Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2024. Auditing large language models: A three-layered approach. *AI and Ethics* 4(4):1085–1115. SpringerLink

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.416

OpenAI. 2024. DALL·E 3. URL: https://openai.com/dall-e-3/ (Accessed 2025-06-12).

OpenAI. 2025. GPT-4o [Large language model]. URL: https://platform.openai.com/ (Accessed 2025-06).

Rachel Schwartz, Jonathan Fiscus, Kenneth Greene, George Waters, Rupa Chowdhury, Thomas Jensen, et al. 2024. The NIST Assessing Risks and Impacts of AI (ARIA) Pilot Evaluation Plan. U.S. National Institute of Standards and Technology. (Program materials and overview.) NIST AI Challenge Problems+1

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei 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)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics. https://aclanthology.org/N18-2003/

Jing Zheng. 2021. A functional review of research on clarity, immediacy, and credibility of teachers and their impacts on motivation and engagement of students. *Frontiers in Psychology* 12:712419.

Kankan Zhou, Eason Lai, and Jing Jiang. 2022. VLStereoSet: A study of stereotypical bias in pre-trained vision–language models. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 527–538, Online only. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.aacl-main.40