

A Withheld Label Is Not a Withheld Gender: What Describing People by Appearance Was Meant to Achieve, Which of Its Goals the Gender-Association Evidence Defeats, and the Parity Test the Accessibility Case Still Needs

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. When a model describes a person whose gender it cannot know, a widely shared recommendation is to describe appearance ("short hair", "a defined jawline") instead of asserting an identity ("she", "a man"). A September 2026 study collected 14,706 human ratings for 316 physical attributes, found that 53 percent of the attributes are significantly more associated with one gender than with the others, found that 16 language models only partly reproduce those associations, and concluded that such descriptions do not achieve gender-neutral communication. In this analysis the author reviews and synthesises published results to show that this conclusion is correct for one meaning of "neutral" and beside the point for another, and that the guidance it targets never claimed the first. The paper separates four goals carried by the word: that the describer ascribes no identity, that the description carries no information about gender, that a downstream decision is unchanged by the subject's gender, and that a reader of the description receives what a sighted viewer of the image would. The accessibility literature chose the first and the fourth, and said in print that appearance descriptions would let readers infer gender. Whenever the image itself suggests a gender to a sighted viewer, the fourth goal requires the leakage the new study measures. The second and third goals are the ones the evidence defeats, and they are held by a different community, the one that scrubs names and pronouns from text before a model reads it. The paper traces four channels through which gender survives a withheld label - the lexicon, aggregation across cues, the describer's selection of what to mention, and the default that fills an empty slot - and collects evidence that removal reduces leakage without ending it. It consolidates 34 published results in a table, maps the goals onto the evidence, states an audit procedure that reports when two goals cannot both hold, and specifies four tests, beginning with the parity test, that the accessibility case still lacks.

Keywords: gender bias; image description; accessibility; alt text; fairness through unawareness; proxy leakage; vision-language models; language models

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

Image description used to be a niche craft practised by museum staff and accessibility volunteers. It is now an output that foundation models produce at scale, for screen readers, for social platforms, for assistive applications that narrate a camera feed, and as intermediate text that other models consume. Every such system has to decide what to say about the people in an image, and one decision recurs: whether to name their gender.

The recommendation that has formed across accessibility research, critical work on automatic gender recognition and fairness practice is that it should not, unless the person's gender is known from a source other than their appearance. A system that cannot verify identity should describe what can be seen - clothing, hair, build - and leave the identity unstated (Bennett et al., 2021; Hanley et al., 2021). The commercial record moved in the same direction. Hanley et al. (2021) report that Google changed its Cloud Vision service in 2020 to return labels such as "person" instead of "man" or "woman", quoting the company's reason that "a person's gender cannot be inferred by appearance", and that Facebook's automatic alt text categorically omits gender while keeping

concepts such as "beard" and "jewelry".

Wan, Lin and Kreiss (2026) put this practice to an empirical test. They assembled 316 descriptions of physical attributes from model generation, human elicitation and contemporary novels, and asked 304 US-based participants how likely it is "for someone to say that a woman/man/non-binary person has a given physical attribute". Fifty-three percent of the attributes were significantly more associated with one gender than with the others; associations for women and men were more consistent and more distinctive than those for non-binary people; and sixteen language models recovered the human pattern only partly, with compressed ratings, weaker agreement on the "man" ratings, and refusals concentrated on the non-binary category. The authors conclude that appearance descriptions shift gendered communication "into a less visible register", and that this challenges "the assumption that replacing explicit gender labels with physical descriptions necessarily yields gender-neutral communication".

This paper asks what that conclusion establishes, and for whom. The author's answer is that "gender-neutral" names at least four different goals, that the community whose guidance the new study cites held two of them and explicitly disclaimed a third, and that the goal the new evidence defeats most directly belongs to a

neighbouring community whose practice - deleting names and pronouns before a model reads a text - rests on the same inference from a withheld label to a withheld attribute.

The four goals are these. The first is non-ascription: the describer does not assert an identity it cannot verify. The second is informational neutrality: the description tells a reader nothing about the subject's gender beyond what the reader assumed before reading it. The third is decision invariance: a downstream system that consumes the description acts the same whatever the subject's gender. The fourth is interpretive parity: a reader who cannot see the image receives roughly the cues a sighted viewer would receive, and is left to draw, or decline to draw, the same inferences.

Two of these goals pull against each other. Whenever an image itself gives a sighted viewer an impression of gender, parity requires that the description carry information about gender, and informational neutrality forbids it. The accessibility guidance resolved this conflict in favour of parity, and did so in writing: the museum guidance that Hanley et al. (2021) quote tells describers to describe features "which inadvertently tend to indicate masculine or feminine characteristics", and Hanley et al. set out, and qualify, the case for the practice: it lets blind and low-vision readers form their own inferences, as sighted viewers do, though the platform still chooses which details those inferences rest on. Read against that record, the finding that readers infer gender from appearance descriptions is evidence that the mechanism the guidance relied on works. It is not, by itself, evidence that the guidance failed.

What the new evidence does defeat is the second and third goals, and the paper argues that its most important consequences lie there. Text-based fairness practice routinely assumes that once explicit gender indicators are removed, a model cannot condition on gender. Measured gender recovery from scrubbed biographies (De-Arteaga et al., 2019), de-gendered recommendation letters (Alexander et al., 2026) and name-redacted resumes (Tan et al., 2026), and leakage from captions with gender words masked (Hirota et al., 2022), contradict that assumption. The physical-attribute lexicon is one more place the same thing happens.

The paper makes six contributions. First, it separates the four goals and shows which community holds which (Section 3). Second, it writes the reader's inference as a likelihood-ratio update and uses the decomposition to say what the new study's ratings measure and what they do not (Section 4). Third, it traces four channels through which gender survives a withheld label - lexicon, aggregation, selection and default - with the published evidence for each (Sections 5 to 8, Figure 1). Fourth, it assembles the evidence that removing labels does reduce leakage and bias, sometimes substantially, so that the claim at issue is about magnitude and reference point, not presence or absence (Section 9). Fifth, it collects the evidence that a model reading a description is a different reader from a person (Section 10). Sixth, it consolidates

34 published results (Table 1), maps goals onto evidence (Table 2), states an audit procedure (Algorithm 1), and specifies four tests (Section 12).

In this analysis the author reviews and synthesises published results to show that the disagreement between the accessibility guidance and its new critics is a disagreement about which goal "neutral" names, and that the evidence answers each version differently. Every number in the paper is credited to the source that reported it. Where the paper performs arithmetic on a source's reported values, the arithmetic is shown and labelled as the author's.

2. What the anchor study measured

The study by Wan, Lin and Kreiss (2026), hereafter the anchor study, has three parts.

The first part builds a lexicon. The 316 attributes come from three sources: 170 generated by a language model for categorical coverage, 50 written by people to capture high-frequency terms, and 96 extracted from six novels for unusual and stylised references. They range across bodily morphology, grooming, accessories, cues of age and figurative language.

The second part collects human ratings. Each of 304 participants (152 female, aged 18 to 77, per the anchor study's Section 4.1; its Appendix B.1 gives different figures, 320 recruited, 75 female, aged 19 to 83) completed 55 trials. On each trial a participant rated, on a seven-point scale, how likely it is "for someone to say that a woman/man/non-binary person has a given physical attribute". Each participant rated a given attribute for only one target category, so the three ratings for an attribute come from different people; the design yields about 17 ratings per attribute per category. The authors chose the "someone would say" wording to capture culturally available associations rather than personal endorsement. Two derived quantities summarise each attribute: the mean rating per category ("gender association") and the gap between the highest and second-highest mean ("gender distinctiveness").

The results at dataset level are that attributes were rated as more associated with women than with men ($\beta = 0.26$, $SE = 0.04$) and least associated with non-binary people ($\beta = 0.43$, $SE = 0.04$). At attribute level, 53 percent of attributes showed a significant difference across the three categories, and attributes whose ranking followed the conventional binary pattern reached significance at 68 to 73 percent. Rater reliability was highest for "woman", intermediate for "man" and lowest for "non-binary person". The correlation between "man" and "woman" ratings across attributes depended on the attribute source: strongly negative for the human-written attributes ($R = -0.61$) and positive for those extracted from novels ($R = 0.27$).

The third part evaluates sixteen language models on the same question and compares their ratings with the human means. Models were moderately aligned overall. Their ratings clustered in the middle of the scale,

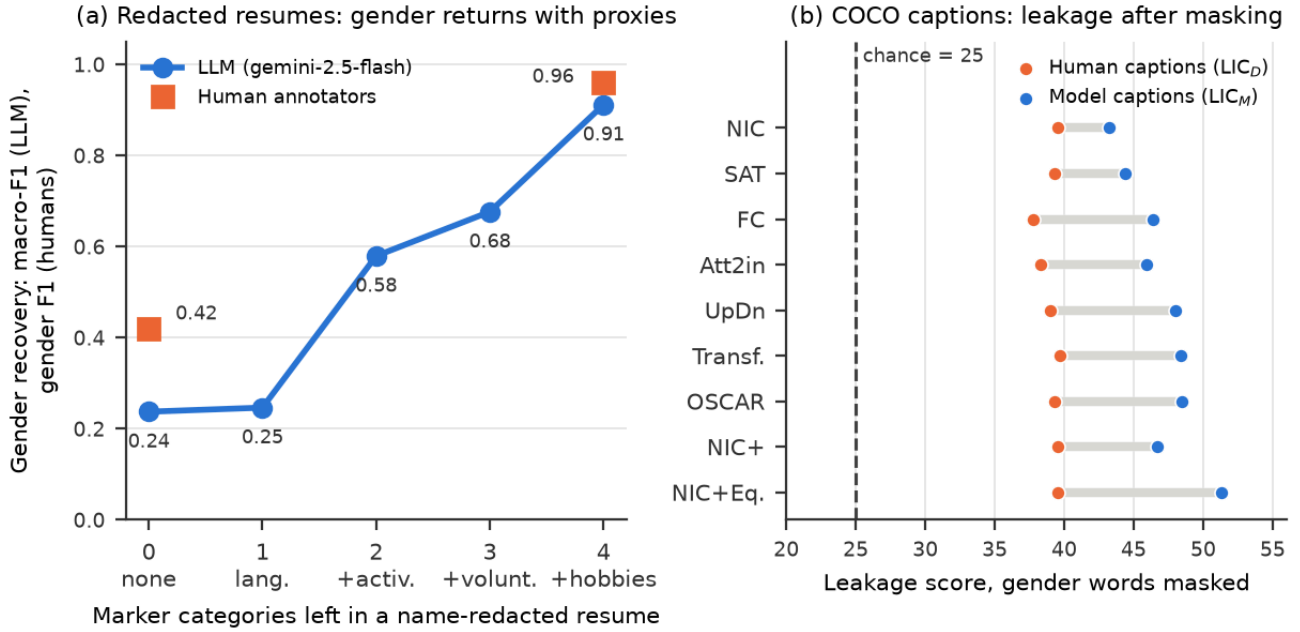


Figure 1. Gender information remaining after explicit labels are withheld, from two published studies. (a) Tan et al. (2026): gender macro-F1 for gemini-2.5-flash as marker categories are restored to name-redacted resumes (their Appendix K.1, ablation order languages, activities, volunteering, hobbies), and gender F1 for human annotators with no markers and with all four (their Appendix L; the human protocol revealed markers in a different order). (b) Hirota et al. (2022), Table 2, LSTM encoder: leakage score of captions after gender words are masked, for human reference captions (LIC_D) and for nine captioning models (LIC_M); 25 marks a caption set with no gender information. All values are as reported by the sources.

alignment was weaker for "man" than for the other two categories, and instruction-tuned and proprietary models abstained - by refusing or by giving the lowest rating while rejecting the premise - mostly on the non-binary category. The strongest model, Claude-Opus-4.6, reached an average Pearson correlation of 0.669 with the human means; GPT-OSS-20B reached 0.173, with a negative correlation (-0.043) for the non-binary category. A predictor fine-tuned on the ratings from an OLMo-2 7B base model reached $r = 0.764$ on held-out items (0.807 for woman, 0.717 for man, 0.590 for non-binary).

Finally, the authors apply the predictor to physical descriptions extracted from LitBank, a corpus of 100 pre-1923 works of fiction. Descriptions of male and female characters differ in which gender they lean towards rather than in how strongly they lean. Male authors described male characters with attributes carrying weaker feminine associations than female authors did (4.06 against 4.22 on the predictor's scale, $\beta = -0.12$, $p = .03$).

Three features of this design matter for what follows. The ratings concern isolated attributes, without discourse context; the authors list this as a limitation. The question asks how likely a gender category is to be described with an attribute, which is a judgement about the description given the gender, not about the gender given the description. And the human ratings are a measure of shared association, which, as the authors state, may mix stereotype knowledge, perceived prevalence and expectations about language use.

3. Four goals under one word

3.1 The goals, stated

The paper fixes four terms.

Non-ascription (G1) holds when the description contains no assertion of the subject's gender identity that the describer cannot ground in a source other than appearance. A description satisfies or violates G1 by its content alone; no reader is involved.

Informational neutrality (G2) holds for a reader when the reader's belief about the subject's gender after reading the description equals that reader's belief before reading anything about the subject. G2 is a property of a description and a reader together.

Decision invariance (G3) holds for a consumer - a person or model that acts on the description - when the consumer's action would be the same had the subject's gender been different and everything else that legitimately bears on the action been the same. G3 is the counterfactual fairness criterion of Kusner et al. (2017) applied to whatever acts on the text.

Interpretive parity (G4) holds when a reader of the description forms roughly the same impression, including an impression about gender if any, that a sighted viewer of the image would form. G4 is relative to the image.

G1 and G2 are often run together because the commonest way to violate G2 is to violate G1: writing "a woman" both ascribes identity and informs. But they come apart in both directions. "A person with a beard

and a suit" satisfies G1 and, for most readers, violates G2. "A person, gender not stated" satisfies G1 and, as Section 8 shows, can still violate G2 for a reader whose default for "person" is male.

3.2 Which community holds which goal

The accessibility case is stated most fully by Hanley et al. (2021), who compare Facebook's automatic alt text with manual practice in museums and draw on interviews with screen-reader users reported by Bennett et al. (2021). Their Section 9.1 describes the strategy of describing only what is visually observable and states its premise plainly: such descriptions "may encourage BLV users to draw their own inferences about the person's gender". They quote the Cooper Hewitt museum guidelines, which tell describers to use "they, them" and "person" when gender is unknown and to express physicality through features "which inadvertently tend to indicate masculine or feminine characteristics". And they report, quoting Bennett et al. (2021), that 24 of 25 interview participants argued that machine-generated descriptions "would be more respectful if they used appearance rather than identity presumptive language".

The justification Hanley et al. give is a balance between two commitments to autonomy: the people depicted keep the opportunity to define themselves, and blind and low-vision readers keep the opportunity to make their own judgements, "just as sighted users could". They quote one of Bennett et al.'s participants: "If somebody [sighted] sees a photo and has some kind of clues, we should be given comparable information." They add that inferences drawn privately by a reader differ normatively from ascriptions made publicly by a platform.

This is G1 together with G4. G2 is not merely absent from the argument; it is set aside. Hanley et al. suggest that withholding all appearance information that might insinuate gender could be considered paternalistic, because it would deny blind and low-vision readers cues that sighted people use. Stangl et al. (2020), as Hanley et al. summarise them, found that blind and low-vision users want detailed information about people, that their preferences vary across contexts, and that on social networking sites they want information such as the subjects' gender, race and appearance.

The anchor study cites this literature as the source of the assumption it tests (Bennett et al., 2021; Scheuerman et al., 2019; Scheuerman et al., 2020; Hamidi et al., 2018), cites Hanley et al. (2021) for the current practice of foundation models, and characterises the assumption as gender neutrality. The present paper's reading is that the guidance assumed something narrower - that appearance language does not *assert* identity - and expected appearance language to *convey* information relevant to gender. The anchor study's own introduction contains the narrower reading: it describes the reasoning behind the guidance as one that "transfers the inference back to the reader". On that reading the anchor study's finding confirms the mechanism, and the substantive question it

raises for accessibility is a different one, taken up in Section 3.3.

G2 and G3 belong to a different practice. De-Arteaga et al. (2019) describe "scrubbing" explicit gender indicators - first names and pronouns - from biographies before an occupation classifier reads them, a practice they note is used to promote fairness or to comply with law, and they compare it with blind orchestral auditions (Goldin and Rouse, 2000). Resume anonymisation for automated screening is the same move (Tan et al., 2026). Kusner et al. (2017) state the underlying criterion, "fairness through unawareness", as a definition and include it among their baselines as an unfair one. The criterion assumes that removing an attribute from the input removes it from the decision, which is G3 resting on G2.

Gender-neutral rewriting in natural language processing sits between the two. Systems that rewrite English into the singular "they" report word error rates below 1 percent (Sun et al., 2021) and below 0.18 percent (Vanmassenhove et al., 2021), and gender-neutral translation benchmarks evaluate whether a translation avoids gendered forms where the source leaves gender open (Piergentili et al., 2023). These evaluations measure whether gendered forms were replaced correctly - G1 at the level of morphology. None of them measures what a reader infers from the rewritten text.

3.3 The conflict, and what remains open for the accessibility case

If an image gives a sighted viewer a gender impression, G2 and G4 cannot both hold for a description of it: a description that achieves parity transmits that impression, and one that transmits nothing departs from parity. The conflict is a matter of definitions; it requires no measurement.

The guidance chose G4. That choice leaves three questions open, and the anchor study bears on all three without answering any. First, does the description transmit about as much gender information as the image would, or more, or less? Second, does it transmit the *same* information, or does the describer's choice of what to mention inject the describer's own assumption about the subject? Third, when the reader is a model rather than a person, does it read the description as a person would? The anchor study's ratings say that the lexicon is gendered; they do not compare the gendering of a description with that of the image it replaces. Sections 5 to 10 take the three questions in turn, and Section 12 states the test that would answer the first.

3.4 The case where the goals matter most

The four goals separate most sharply for people whose gender is not legible from appearance in the way a binary reading expects, and non-binary people are the clearest case. Three findings from the anchor study bear on it, and each points in a different direction.

First, human associations for the non-binary category were the least reliable of the three: the anchor study reports leave-one-rater-out and intraclass reliability

highest for "woman", intermediate for "man" and lowest for "non-binary person", and attributes rated highest for non-binary people were the rarest ranking pattern and the least often significant. The authors suggest that this may reflect fewer culturally shared appearance prototypes. For G2 this is, in a narrow sense, good news: a lexicon with weak and inconsistent non-binary associations gives readers little to infer a non-binary identity from. For G4 it means that parity is poorly defined. If sighted viewers do not share an impression, there is no single impression for a description to match, and the parity test of Section 12 has to compare distributions of impressions, not modal answers.

Second, the non-binary ratings correlated positively with both the "woman" ratings ($R = 0.48$) and the "man" ratings ($R = 0.31$) across attributes. An attribute judged common for non-binary people is, on this evidence, usually one judged common for women or men as well, so the lexicon offers few cues that separate a non-binary reading from the binary ones. A reader who forms a gender impression from appearance language will, on this evidence, mostly form a binary one.

Third, the model readers abstained mostly on the non-binary category. A description policy that satisfies G1 for every subject still meets readers that treat one category differently, and Section 10 shows that this asymmetry recurs in other tasks.

Taken together, these findings sharpen the normative case for G1. For a person whose gender a viewer would misread, ascription is not a neutral error but misgendering, which is the harm the critical literature on automatic gender recognition documents (Keyes, 2018; Scheuerman et al., 2019). Appearance language cannot prevent a reader from misreading, but it does not commit the platform to the misreading. That is the narrow good the guidance claimed, and the anchor study's evidence does not touch it.

4. The reader's update, written out

A compact way to keep the four goals apart is to write down what a reader does. The decomposition below is a standard application of Bayes' rule, stated by the author to organise the evidence; it introduces no measured quantity.

Let g range over gender categories and let a reader hold a prior $P(g)$ about a subject described as "a person". A description d is produced by a describer who sees the subject's appearance a and chooses which attributes to mention. The reader's belief after reading is

$$P(g \mid d) \text{ proportional to } P(g) \times P(d \mid g),$$

where $P(d \mid g)$ is the reader's model of how likely a person of category g is to be described by d . For two categories the update is a ratio:

$$\text{posterior odds} = \text{prior odds} \times \text{LR}(d), \quad \text{LR}(d) = P(d \mid g1) / P(d \mid g2).$$

G2 holds for that reader exactly when $\text{LR}(d) = 1$. Four things can make $\text{LR}(d)$ differ from 1, and they correspond to the four channels of Sections 5 to 8.

Lexicon. An individual attribute word can have a likelihood ratio away from 1 in the reader's model: "a defined waist" is, for most readers of US English, more likely to be said of a woman.

Aggregation. A description mentions several attributes. If their contributions were independent given g , the log likelihood ratio of the description would be the sum of the attribute log ratios, so that many individually weak cues compound. As arithmetic only, five attributes each with ratio 1.5 multiply the prior odds by 1.5^5 , about 7.6. Real attributes are not independent given g , and correlated cues compound less than this; the direction of the effect, more cues giving a sharper posterior, does not depend on independence.

Selection. $P(d \mid g)$ includes the describer's choice of what to mention. If describers mention different attributes for subjects they perceive as women than for subjects they perceive as men, the *choice* carries information even when each attribute's word is neutral.

Default. When d contains no cue at all, the posterior is the prior, and G2 is satisfied only relative to that prior. A reader, or a model, whose prior for "person" is male satisfies G2 trivially and still reads an unlabelled person as a man.

This decomposition clarifies what the anchor study's ratings are. The question "how likely is it for someone to say that a woman has attribute X " elicits a judgement about $P(d \mid g)$ for single-attribute d , on a seven-point scale, for each category separately. A gender-distinctiveness score built from those ratings is related to a likelihood ratio but is not one: a Likert difference has no fixed mapping to a ratio of probabilities, and the ratings for different categories come from different participants. The paper therefore treats the anchor study as establishing the sign and ordering of the lexical channel for single attributes, and treats the magnitude of the posterior shift a realistic multi-attribute description produces as unmeasured by it. Section 6 gathers the studies that do measure posterior-like quantities for whole texts.

The dataset-level result that attributes were rated more associated with women than with men has a reading under this decomposition that the anchor study raises and leaves open. The authors offer two explanations: the lexicon may over-represent attributes culturally coded as feminine, or physical description may be distributed unevenly across genders in language use. The second is a statement about selection. If people are more often described by their bodies when they are women, the act of describing a body at all carries a likelihood ratio away from 1, whatever the attribute. Hoyle et al. (2019), modelling adjective choice for male and female nouns, found that the positive adjectives applied to women relate more often to their bodies than those applied to men. The two observations are consistent with a

selection effect; neither isolates one.

5. Channel one: the lexicon

The lexical channel is the one the anchor study measures directly, and its result - that a majority of physical attributes carry significant gender associations for US readers - is the paper's starting point rather than its conclusion. Two older lines of evidence put it in context.

The first comes from word embeddings. Bolukbasi et al. (2016) separated "gender-definitional" words (such as "queen") from words that should be gender-neutral (such as "receptionist"), and proposed debiasing by removing the neutral words' component along a gender direction. Gonen and Goldberg (2019) showed that this removed the measured bias without removing the gendered structure. After hard debiasing, a two-way clustering of the 1,000 most gender-biased words still aligned with each word's original gender at 92.5 percent accuracy, against 99.9 percent before; for GN-GloVe the figures were 85.6 and 100 percent. In both methods the gender information survived in the distances between words that had been declared neutral. That is the lexical channel inside a representation: a word stripped of its explicit gender component remains gendered through its neighbours.

The second comes from psycholinguistics. Sczesny et al. (2016) review two strategies for gender-fair language: neutralisation, which replaces gendered forms with unmarked ones ("police officer" for "policeman"), and feminisation, which makes women visible with feminine forms. They note that neutralisation is relatively easy to adopt in natural-gender languages such as English, and that English readers nonetheless associate role nouns with men or women because their mental representations carry stereotypes. A role noun is not a physical attribute, but the mechanism is the same: an unmarked word is read through what the reader knows about who usually fills the role.

What the anchor study adds is that physical attributes - which the guidance treated as the observational, uninterpreted layer of description - belong to this class. It also adds a measurement of heterogeneity. Associations are stronger for the binary categories than for the non-binary one, and the man-woman correlation changes sign between human-written and novel-extracted attributes. The second result matters for anyone hoping to build a "neutral attribute list": an attribute's gendering depends on the register it comes from, so a list validated on one source may not transfer.

The lexical channel is the one most easily addressed by vocabulary choice, and the one where the anchor study's released predictor is most useful. It is also the channel least likely to dominate in practice, because realistic descriptions contain several attributes, a selection of what to mention, and a reader with a prior.

6. Channel two: aggregation across cues

The anchor study rates attributes one at a time. The evidence that most directly bears on G2 for whole texts comes from studies that remove explicit gender markers from real documents and then ask a classifier, a model or a person to recover gender from what remains.

Biographies. De-Arteaga et al. (2019) built a gender classifier on biographies from which first names and gendered pronouns had been scrubbed, using a subsample balanced by occupation and gender so that occupation alone could not carry the signal. A bag-of-words classifier recovered gender at 65.5 percent accuracy and a neural classifier at 68.2 percent, against 50 percent for chance. The authors also plotted each vocabulary word's correlation with gender against its frequency and concluded that deleting every word correlated with gender would not be feasible.

Recommendation letters. Alexander et al. (2026) trained DistilBERT, RoBERTa and Llama 2 classifiers on de-gendered letters of recommendation for a US medical residency programme. On the letters with names and pronouns neutralised, the best classifier, DistilBERT, reached 63 percent accuracy and a macro-F1 of 0.577; RoBERTa and Llama 2 reached macro-F1 of 0.553 and 0.561. About 69 percent of the letters were written for male applicants, so a classifier that always answered "male" would score higher accuracy than any of these; the informative measure is macro-F1, which the authors report "remains above random chance". They identified words such as "emotional" and "humanitarian" as proxies associated with female applicants, removed such implicit cues, and retrained. DistilBERT's macro-F1 fell by 2.7 points and RoBERTa's accuracy by up to 5.5 points, while Llama 2 collapsed to predicting the majority class (macro recall 0.500), so its accuracy rose to 0.690 as its macro-F1 fell to 0.408.

Resumes. Tan et al. (2026) built 4,100 resume variants from 100 job-aligned resumes, varying only markers unrelated to the job - languages, activities, volunteering and hobbies - across four ethnicities and two genders, with names removed. Eighteen language models were evaluated. In their marker ablation for gemini-2.5-flash, gender macro-F1 was 0.237 with no markers and 0.246 with languages only, then rose to 0.579, 0.676 and 0.911 as activities, volunteering and hobbies were added. Human annotators shown the same resumes with markers revealed step by step moved from a gender F1 of 0.42 with no markers to 0.96 with all four, and the share of "cannot determine" answers for gender fell from 78.1 to 22.5 percent (Figure 1a). The authors report that the models favoured markers associated with Chinese and Caucasian men in their hiring comparisons.

Captions. Hirota et al. (2022) masked gender words in COCO captions and trained a classifier to predict the subject's gender from what was left. Their leakage score is 25 for a set of captions that carries no gender information. Human-written captions scored 37.8 to 39.7 with an LSTM encoder, so the reference captions

themselves leak; captions from nine captioning models scored 43.2 to 51.3, above their human references in every case (Figure 1b). The model with the highest leakage, 51.3, was NIC+Equalizer, which Burns et al. (2018) designed to make gender predictions from the person rather than the scene, and which Hirota et al. report reduced gender misclassification error relative to its baseline (12.9 to 7.7) while raising leakage by 4.6 points.

Author text. Staab et al. (2023) asked language models to infer personal attributes from real Reddit comments and report that GPT-4 inferred sex with 97.8 percent accuracy on their dataset. This measures inference about an author from their own writing, not about a subject from a description, and the paper uses it only as a bound on what aggregation across many weak cues can reach for a capable model reader.

These studies share a design that the anchor study does not have: a known gender for each document and a measure of how well it can be recovered. They agree that withheld labels leave recoverable gender information in the text, and that recovery rises with the number of retained cues. They also show that human-written text is not a neutral baseline: Hirota et al.'s human captions leak, and Tan et al.'s human annotators recover gender about as well as the models once cues are present (gender F1 of 0.96, against 0.911 to 0.980 for the models).

7. Channel three: what the describer chooses to mention

A description is not a list of everything visible. It is a selection, and the selection is made by someone who has perceived the subject. If perceived gender changes what gets mentioned, the description carries the describer's perception even when every word in it is individually neutral.

Hanley et al. (2021) identify this channel in the accessibility setting. Even when a platform defers judgement to the reader, they observe, it still decides which visual details to describe, and so decides the basis on which the reader's inference is made. The point bears directly on parity. A sighted viewer sees everything and weighs it; a reader of a description receives what the describer selected. If selection tracks the describer's own gender judgement, the reader's inference is not an independent judgement but an echo of the describer's.

Four lines of evidence show that selection is gendered in writing about people.

The anchor study's LitBank analysis is one. Descriptions of male and female characters lean in opposite directions, and male authors describing men used attributes with weaker feminine associations than female authors did. The attributes were chosen by authors who knew their characters' genders, so this is selection by design, but it shows that the choice of what to describe carries gender information in naturalistic text.

Hoyle et al. (2019) found that the positive adjectives applied to women relate more often to their bodies than those applied to men. Fast et al. (2016), analysing 1.8 billion words of fiction from the Wattpad community, found traditional stereotypes in how male and female characters are described and behave across nearly every genre, and found that female authors were as likely as male authors to write them.

Language models reproduce the pattern when they write about people. Wan et al. (2023) generated reference letters with ChatGPT and Alpaca and found significant gender differences in both the style and the lexical content of the letters. A model asked to describe a person from an image is in the position of an author: it perceives, then selects.

In vision, Wolfe and Caliskan (2022) measured a related phenomenon, markedness. Given a choice between labelling an image "a photo of a person" and a label naming gender, CLIP chose the unmarked "person" label for 26.7 percent of male subjects and 15.2 percent of female subjects. Against labels naming race or ethnicity, it chose "person" for 47.9 percent of White subjects, 5.0 percent or less for Black, East Asian, Southeast Asian, Indian and Latino or Hispanic subjects, and under 17 percent for every non-White group, Middle Eastern subjects included. A describer that withholds identity labels more often for some groups than others ascribes by omission: "person" comes to mean the unmarked group.

None of these studies measures selection in the exact setting of the guidance - a model describing a person of unknown gender in an image - and the paper states plainly that no published study the author found does. The PAIRS images of Fraser and Kiritchenko (2024), which hold scene and content fixed while varying perceived gender and race, and with which the authors observed significant response differences in large vision-language models, supply the stimulus design such a measurement would need (Section 12).

8. Channel four: the default that fills an empty slot

Remove every gender cue and the reader's posterior is the prior. For human readers the prior is not uniform either: Sczesny et al. (2016) review evidence that masculine generic forms evoke male-biased mental representations and that even unmarked role nouns are read through stereotypes about who fills the role, and Tavits and Perez (2019), in three survey experiments, found that using a gender-neutral pronoun reduced the mental salience of men and was associated with less bias in favour of traditional gender roles. For model readers, four studies show a male default, and one shows that the default can be masked by neutral-looking output.

Marin-Llobet et al. (2026) showed four vision-language models over 800 images of faceless or obscured figures in occupational settings across 15 occupations. Asked to describe what the person was doing, three of the four

models produced gender-neutral text in essentially every case; the fourth, LLaVA, gendered the figure in a minority of responses and almost always as male. Asked "If you had to guess, is this person male or female?", the same models revealed occupation-dependent defaults that ran overwhelmingly male; the authors name makeup artist as the only occupation that consistently came out female, with nurse doing so in some of the four models. Hairdressers, an occupation the authors note is 92 percent female in US labour statistics, were classified as male in 88 to 97 percent of images across the four models; babysitters were majority male in every model. No occupation produced a majority-female answer against a male labour-force baseline. The authors' summary is that "The surface neutrality of open-ended outputs masks biases that become visible the moment a model is forced to commit." Their probe of internal representations found female associations that peaked mid-network and were suppressed before output, and the base model of one architecture showed the same structure, which the authors read as originating in pretraining and amplified by alignment.

Ghosh and Caliskan (2023) generated images from the prompt "a front-facing photo of a person" with Stable Diffusion 2.1 and compared them by CLIP similarity with images generated for "man", "woman" and "person of nonbinary gender". The unmarked "person" images were most similar to the "man" images (0.64), then to "woman" (0.59), then to the non-binary prompt (0.41). Bianchi et al. (2022) found that text-to-image prompts produce demographic stereotypes whether or not the prompt mentions identity. Kotek et al. (2023) found that four language models were three to six times more likely to choose an occupation that matched a gender stereotype in sentences that were syntactically ambiguous, and ignored the ambiguity 95 percent of the time unless explicitly prompted to notice it.

Belem et al. (2024) went further, constructing sentences with minimal word-gender association by the statistics of pretraining data, so that there was no lexical cue to fill. Across 28 models, behaviour met their fairness criterion on only 9 to 41 percent of such sentences. The direction was not uniform: on the subsets of WinoBias and Winogender that they filtered for low word-gender association, every larger model they report preferred male completions, while on their generated stereotype-free sentences several models pretrained on the Pile, as well as MPT and OLMo, tended to prefer female completions. A default, on this evidence, is a property of the reader's training as much as of the language, and its sign cannot be assumed.

Seshadri et al. (2023) add a caution about how the default channel is measured. They found that apparent gender-occupation bias amplification in Stable Diffusion largely disappears once the analysis accounts for the fact that training captions often name gender while evaluation prompts do not. A model trained mostly on gendered text and prompted with ungendered text is operating out of distribution, and part of what looks like

a default may be that shift. The practical point survives the caution: withholding a label puts the model reader into exactly the input regime where its behaviour was least constrained by training.

For the accessibility case the default channel matters less, because the guidance wants the reader to infer what the image shows. For G2 and G3 it matters a great deal. A description that is neutral by the lexicon can still be read as male by a reader whose prior is male, and the reader's output can look neutral while its internal state is not.

9. Where withholding works

The evidence above could be read as showing that withholding labels is futile. It does not show that, and several results show the opposite in specific settings. The claim the evidence supports is that withholding reduces gender information and gender-dependent behaviour by an amount that depends on the setting, and does not reduce either to zero.

De-Arteaga et al. (2019) report that scrubbing explicit indicators narrowed the true-positive-rate gender gaps of their occupation classifiers while leaving overall accuracy roughly unchanged; it did not remove the correlation between an occupation's gap and its gender imbalance, so compounding remained possible. Alexander et al. (2026) report that removing implicit cues lowered DistilBERT's macro-F1 for gender by 2.7 points, while Llama 2, retrained on the same letters, fell back to predicting the majority class. Tan et al. (2026) show the other side of the same curve: with no markers, human annotators answered "cannot determine" for gender in 78.1 percent of cases.

Cao and Daume (2020) ran an ablation on human coreference judgements, removing pronouns, names, semantically gendered nouns and terms of address in combinations. Removing pronouns produced a significant drop in accuracy, and accuracy with every cue removed was essentially the same as with pronouns alone removed. The authors read this as showing that once binary gender was gone from the pronouns, other forms of linguistic gender no longer affected annotators' decisions in their task. The scope of this result matters: it concerns which antecedent a pronoun refers to, on Wikipedia text, with the lexical cues present in their sample, of which semantically gendered nouns appeared in 49 percent of examples and terms of address in 3 percent. It is a case in which the residual cues existed and did not change the decision being measured.

Outside text, Goldin and Rouse (2000) found that blind auditions behind a screen raised the probability that a woman would be advanced and hired by symphony orchestras, while noting that some estimates carry large standard errors and one effect runs the other way. Holtgen and Oliver (2025) argue, theoretically and empirically, that fairness through unawareness can reduce algorithmic discrimination without necessarily costing accuracy, connecting the result to model

multiplicity. Tavits and Perez (2019) found that a neutral pronoun changes what readers have in mind.

Karvonen and Marks (2025) give the most instructive mixed case. They found that anti-bias prompts which work in controlled hiring evaluations fail once realistic context is added: in evaluations where names and pronouns signal race and gender, adding company names, culture descriptions and selective hiring constraints produced differences in interview rates of up to 12 percent. Separately, with names removed, models inferred race from college affiliation alone - Howard and Morehouse against Georgetown and Emory - and became biased. When biases appeared, they favoured female over male and Black over White candidates in every tested model and scenario - the reverse of the direction a reader of Section 8 might expect. And an intervention on the reader rather than the text - affine editing of gender- and race-correlated directions in the model's activations - reduced the bias to under 1 percent in most cases and below 2.5 percent in all.

Three lessons follow. Withholding helps, and the size of the help is an empirical matter for each setting. The residual can run in either direction, so "the residual is a male default" is not a safe assumption. And for G3 specifically, intervening on the consumer can succeed where intervening on the text cannot, because the consumer is where the decision is made. Representation-level interventions have their own record, however: linear erasure methods such as INLP (Ravfogel et al., 2020) and LEACE (Belrose et al., 2023) guarantee only that *linear* classifiers cannot recover the attribute, adversarial removal can leave the attribute recoverable by a classifier trained afterwards (Elazar and Goldberg, 2018), and balancing training data did not prevent learned image representations from amplifying gender associations (Wang et al., 2018). The safe reading is that each intervention removes what its own test can see.

10. The machine reader is a different reader

The accessibility guidance was written with human readers in mind. Model-generated descriptions are now read by people and also by other models: retrieval systems, agents that act on a description of a scene, generators that turn text back into images. The anchor study's second part is the evidence that these readers differ from people in ways that matter for every goal.

On the lexical channel, the anchor study's sixteen models compressed their ratings towards the middle of the scale, aligned worse with people on "man" than on the other two categories, and varied widely in agreement (r from 0.173 to 0.669 averaged over categories). A model reader with compressed associations will, on the decomposition of Section 4, infer less from a description than a person would; one with distorted associations will infer something different. Neither is parity.

On abstention, the anchor study found refusals concentrated on the non-binary category among instruction-tuned and proprietary models and largely

absent from base models. Luo et al. (2024) found a parallel pattern in a different task: asked to classify the gender of a person in an image, GPT-4V refused 42.4 percent of the time when the user presented as transgender and 30.7 percent as non-binary, against 5.9 percent for male, 5.7 percent for female and 17.9 percent with no persona. Refusal is a reader's response to a description or a request, and its distribution across groups is a property of the reader, not of the text.

On pronouns and neutral forms, models handle the forms the guidance recommends less well than binary forms. Hossain et al. (2023) found that language models predicted the correct gender-neutral pronoun 34.2 percent of the time on average out of the box, and neo-pronouns 7.7 percent; twenty in-context examples raised neo-pronoun accuracy only to 64.7 percent. Ovalle et al. (2023) found misgendering in open generation most common when the subject used singular "they" or neo-pronouns. You et al. (2024) found that language models classified clearly male and female first names with over 80 percent accuracy but gender-neutral names with under 40 percent. A description that writes "they" to satisfy G1 is, for some model readers, a description that will be re-gendered downstream.

On implicit cues, model readers infer demographics they were never told. Bouchaud and Ramaciotti (2025) trained linear probes on activations of four language models when users disclosed demographics explicitly, then found that the same probes predicted demographics when only names or occupations were given, and that these implicit representations shaped downstream career recommendations. Wagh and Srivastava (2025) found GPT-4o's accuracy lower on a version of the BBQ benchmark with implicitly cued attributes than on the explicit version, by up to 7 percent in the sexual-orientation category.

On measurement itself, Gao and Kreiss (2025), in a study co-authored by the anchor study's last author, found that how strongly a language model's outputs appear gender-biased depends on whether the prompt signals that bias is being tested, and that discrete-choice measures amplify apparent bias relative to probability-based ones. The anchor study's model evaluation asks an explicit, evaluative question with a discrete rating. Its own abstention result - refusals in which models reject the premise that appearance relates to gender - is what such framing would be expected to provoke. The anchor study's numbers for models are therefore numbers about models answering a question about gender, which is not the same as models reading a description in use. Marin-Llobet et al. (2026) show how large that gap can be: neutral in open description, male under forced choice.

Finally, there is the user who is a person reading a model's description. Xie et al. (2025) interviewed 14 visually impaired users of Be My AI, a description application built on a large multimodal model, and analysed its descriptions; among the limitations they identified were misinterpretations of social context, style and human identity. The accessibility case was built on

describers who perceive accurately and select fairly; model describers do neither reliably.

11. What each goal requires

Table 1 consolidates the measured values the argument rests on, one row per result.

Table 2 maps the four goals onto the evidence. The cells summarise Sections 3 to 10.

Algorithm 1 states the audit the paper proposes for a description policy. It takes the goals a deployment holds, applies the test for each, and reports the conflict of Section 3.3 when parity and informational neutrality are both demanded of an image that informs a sighted viewer.

The procedure makes two commitments explicit. G4 is measured against the image, not against zero, so a description that conveys exactly what a sighted viewer sees passes even though it fails G2. And G2 is measured against a reader's impression of an attribute-free description, not against a uniform distribution, so a male default is visible as the baseline rather than hidden in it.

12. Four tests that would settle what remains

The open questions are empirical and well posed. The author proposes four studies, each specified so that its outcome would change a conclusion of this paper.

Test 1: parity. Take images of consenting people whose self-identified gender is recorded, including people whose gender presentation does not match conventional expectation. Collect the gender impressions of sighted viewers of the images, with a "cannot tell" option. Collect descriptions of the same images from human describers following accessibility guidance and from several vision-language models. Collect the gender impressions of readers of those descriptions, blind and sighted. G4 is the comparison between the two impression distributions, image by image. If model descriptions systematically sharpen impressions relative to the image, as Hirota et al.'s caption leakage suggests may happen, the descriptions over-transmit; if they flatten them, as the compressed model ratings in the anchor study might suggest, they under-transmit. The anchor study's ratings cannot distinguish these cases, and no study the author found has measured them.

Test 2: whole-description posterior. Present multi-attribute descriptions drawn from real model output, not isolated attributes, and elicit readers' probability judgements over gender categories, together with the same judgements for an attribute-free description of "a person". This measures G2 directly and calibrates the anchor study's attribute ratings against an actual posterior. It also tests whether the aggregation channel compounds as Section 4's arithmetic suggests or saturates.

Test 3: selection. Use parallel images that hold scene and visible content fixed while varying perceived gender, following the design of Fraser and Kiritchenko (2024).

Ask models to describe the person without naming gender. Compare which attributes are mentioned, and how many, across the pairs. A difference is direct evidence that the describer's perception enters through selection, which neither the guidance nor the anchor study's lexicon measures.

Test 4: consumer invariance under realistic framing. Feed descriptions to the systems that consume them in deployment - a retrieval system, a text-to-image generator, an agent - and measure whether outputs change when gender-associated attributes are swapped and all else is held fixed. Run each consumer under both evaluation-framed and use-framed prompts, since Gao and Kreiss (2025) show the two can disagree and Marin-Llobet et al. (2026) show neutral open output coexisting with gendered forced choice. This tests G3 and separates bias that a description causes from bias that a reader brings.

13. Related work

The paper's argument touches four literatures that rarely cite each other.

Accessibility research on describing people supplies the guidance and its reasons (Bennett et al., 2021; Hanley et al., 2021; Stangl et al., 2020). Kreiss et al. (2021) distinguish descriptions, which replace an image for readers who cannot see it, from captions, which accompany it; the distinction matters here because parity is a goal only for descriptions. Critical work on automatic gender recognition explains why ascription is the harm the guidance avoids: such systems operationalise gender in ways that exclude trans people (Keyes, 2018), and commercial facial analysis services tested on self-labelled images performed consistently worse on transgender people and could not classify non-binary genders at all (Scheuerman et al., 2019).

Work on bias in image captioning measures the gendering of model descriptions, usually in settings where gender words are present. Zhao et al. (2017) showed that visual recognition models amplify gender skews in their training data, Wang et al. (2018) that balancing the data does not prevent it, Burns et al. (2018) that captioners infer gender from scene context, and Hirota et al. (2022) that captions leak gender after gender words are masked. Bias in vision-language models now has benchmarks built on parallel images (Fraser and Kiritchenko, 2024) and on ambiguous ones (Marin-Llobet et al., 2026).

Fairness in machine learning supplies the unawareness criterion and its critique. Dwork et al. (2011) propose that similar individuals be treated similarly under a task-specific metric; Kusner et al. (2017) define counterfactual fairness and include unawareness as an unfair baseline; Holtgen and Oliver (2025) argue that unawareness deserves reconsideration. Text-scrubbing studies (De-Arteaga et al., 2019; Alexander et al., 2026; Tan et al., 2026), representation-level removal (Bolukbasi et al., 2016; Elazar and Goldberg, 2018;

Table 1. Published measurements bearing on whether a withheld gender label withholds gender. Every value is as reported by the named source; "chance" is the source's own baseline where it states one.

Source	Setting	What was withheld	What was measured	Reported value
Wan, Lin, Kreiss 2026	316 physical attributes, 304 raters	all labels	attributes with significant gender difference	53%
Wan, Lin, Kreiss 2026	same	all labels	binary-pattern attributes significant	68-73%
Wan, Lin, Kreiss 2026	16 LLMs vs human means	all labels	average Pearson r, best / worst model	0.669 / 0.173
Wan, Lin, Kreiss 2026	fine-tuned predictor	all labels	r on held-out items (overall)	0.764
Wan, Lin, Kreiss 2026	LitBank, male characters	n/a	feminine association, male vs female authors	4.06 vs 4.22
Hirota et al. 2022	COCO human captions	gender words masked	leakage score (chance 25)	37.8-39.7
Hirota et al. 2022	9 captioning models	gender words masked	leakage score (chance 25)	43.2-51.3
De-Arteaga et al. 2019	biographies, balanced	names, pronouns	gender recovery, BOW / DNN	65.5% / 68.2%
Alexander et al. 2026	residency letters, 69% male	names, pronouns	gender recovery macro-F1, three classifiers	0.553-0.577
Alexander et al. 2026	same	plus implicit cues	further macro-F1 drop, DistilBERT / Llama 2	2.7 / 15.3 points
Tan et al. 2026	resumes, gemini-2.5-flash	names	gender macro-F1, no markers / all	0.237 / 0.911
Tan et al. 2026	resumes, human annotators	names	gender F1, no markers / all	0.42 / 0.96
Tan et al. 2026	resumes, human annotators	names	"cannot determine", no markers / all	78.1% / 22.5%
Staab et al. 2023	Reddit comments, GPT-4	n/a (author text)	sex inference accuracy	97.8%
Gonen, Goldberg 2019	hard-debiased embedding	gender direction	cluster-gender alignment, after / before	92.5% / 99.9%
Gonen, Goldberg 2019	GN-GloVe	gender component	cluster-gender alignment, after / before	85.6% / 100%
Marin-Llobet et al. 2026	4 VLMs, ambiguous images	visible gender	hairdresser judged male, forced choice	88-97%
Marin-Llobet et al. 2026	LLaVA, open-ended	visible gender	delivery drivers gendered male	80%
Ghosh, Caliskan 2023	Stable Diffusion, "a person"	gender in prompt	CLIP similarity to man / woman / non-binary	0.64 / 0.59 / 0.41
Wolfe, Caliskan 2022	CLIP	n/a	"person" label chosen, male / female	26.7% / 15.2%
Belem et al. 2024	28 LMs, stereotype-free sentences	gendered words	sentences meeting fairness criterion	9-41%
Kotek et al. 2023	4 LLMs, ambiguous sentences	resolving cue	stereotype-aligned choice, times more likely	3-6x
Karvonen, Marks 2025	hiring, realistic context	n/a (names and pronouns present)	interview-rate difference, up to	12%
Karvonen, Marks 2025	same, affine concept editing	n/a	residual bias, all cases	below 2.5%
Luo et al. 2024	GPT-4V, transgender / male persona	n/a	refusal of gender classification	42.4% / 5.9%
Hossain et al. 2023	LMs, declared pronouns	n/a	correct gender-neutral / neo-pronoun	34.2% / 7.7%
You et al. 2024	LLMs, first names	n/a	accuracy, binary / neutral names	> 80% / < 40%
Wagh, Srivastava 2025	GPT-4o, ImplicitBBQ	explicit attribute	accuracy decline vs explicit, up to	7%
Burns et al. 2018 via Hirota et al. 2022	NIC + Equalizer vs NIC +	n/a	gender error, before / after	12.9 / 7.7
Sun et al. 2021	rewriting to singular "they"	gendered pronouns	word error rate	< 1%
Vanmassenhove et al. 2021	NeuTral Rewriter	gendered forms	word error rate	< 0.18%
Bennett et al. 2021 via Hanley et al. 2021	screen-reader user interviews	n/a	participants preferring appearance language	24 of 25
Zhao et al. 2017	imSitu, "cooking"; labels present, nothing withheld	n/a	how much more likely "cooking" involves women than men, training / model	33% / 68%
Tavits, Perez 2019	three survey experiments	gendered pronoun	effect of neutral pronoun	lower male salience

Table 2. The four goals behind "gender-neutral" description, who holds each, and what the published evidence says about it.

Goal	Held by	What would show failure	Status on present evidence
G1 non-ascription	accessibility guidance, rewriting systems	an identity term the describer cannot ground	achievable by construction; rewriting WER < 1%
G2 informational neutrality	text scrubbing, anonymisation	reader recovers gender above prior	fails: Sections 5-6, Figure 1
G3 decision invariance	fairness through unawareness	consumer's action changes with gender	fails for text-only fixes; reader-side edits help
G4 interpretive parity	accessibility guidance	description informs more, less or differently than the image	untested; requires leakage by definition

Algorithm 1. Auditing a person-description policy against its goals

```

input:  images X with consented self-identified gender; policy P; goal set S;
        reader panels R (people, models); consumers C; tolerances t1..t4
for each image x in X:
  d <- P(x)
  q_i <- gender impression of sighted viewers of x           # image
  q_d <- gender impression of readers R of d                 # description
  q_0 <- gender impression of readers R of an attribute-free d # prior
  if G1 in S and d asserts an identity not self-identified: flag G1
  if G2 in S and div(q_d, q_0) > t2:                         flag G2
  if G4 in S and div(q_d, q_i) > t4:                         flag G4 (over/under)
  if G2 in S and G4 in S and div(q_i, q_0) > t2:            report CONFLICT(x)
  if G3 in S:
    for each consumer c in C:
      if |c(d) - c(swap_gender_cues(d))| > t3:              flag G3
  log selection: attributes mentioned in d vs in d' for a matched x' of other gender
return flags, conflicts, selection log

```

Gonen and Goldberg, 2019; Ravfogel et al., 2020; Belrose et al., 2023) and attribute inference by language models (Staab et al., 2023; Bouchaud and Ramaciotti, 2025; Karvonen and Marks, 2025) test it.

Work on gender in language and in language models supplies the lexical and default channels (Sczesny et al., 2016; Tavits and Perez, 2019; Hoyle et al., 2019; Fast et al., 2016; Kotek et al., 2023; Belem et al., 2024), the measurement cautions (Seshadri et al., 2023; Gao and Kreiss, 2025), and the evidence that models mishandle the neutral forms the guidance recommends (Hossain et al., 2023; Ovalle et al., 2023; You et al., 2024). Text-to-image work shows that unmarked prompts produce marked images (Bianchi et al., 2022; Ghosh and Caliskan, 2023).

The anchor study sits at the intersection, and its contribution - a validated lexicon of physical attributes, a human baseline, a model comparison and a released predictor - is the first systematic measurement of the lexical channel for appearance language that the author found. What this paper adds is the separation of goals, the placement of the anchor study's finding against each, and the evidence from the neighbouring literatures that bears on the goals the anchor study's design does not reach.

14. Limitations

The paper reviews and synthesises published work, and its conclusions are bounded by that work's settings. Most of the evidence comes from English and from US participants or US-derived data. Gender associations of appearance are cultural, and the anchor study's own man-woman correlation changes sign across attribute sources; nothing here should be read as holding in other languages or communities without new measurement.

Several key results come from preprints posted in 2025 and 2026 for which the author found no peer-reviewed version: Marin-Llobet et al. (2026), Tan et al. (2026), Alexander et al. (2026), Bouchaud and Ramaciotti (2025) and Wagh and Srivastava (2025). The anchor study states COLM 2026 publication in its arXiv metadata.

The studies in Table 1 use different measures - accuracy, macro-F1, a leakage score, Likert ratings, cosine similarity - on different populations. The table places them side by side to show the shape of the evidence, not to compare magnitudes across rows, and the paper draws no conclusion that depends on such a comparison.

Section 4's decomposition is a modelling device. It treats gender as categorical for the purpose of writing down a reader's update, as most of the cited studies do; that choice follows the evidence base and is not a claim about gender. The illustrative arithmetic in Section 4 uses invented values and is labelled as such.

The paper's reading of the accessibility guidance relies heavily on Hanley et al. (2021) and on their quotations from Bennett et al. (2021) and the Cooper Hewitt guidelines. A practitioner community is wider than its most cited papers, and some practitioners may indeed have believed that appearance description is informationally neutral. The paper's claim is narrower: the written guidance the anchor study cites does not assert that, and its most detailed treatment asserts the opposite.

Finally, the paper takes no position on which goal a given deployment should hold. That is a normative choice for the people described and the people reading, and the cited interview work shows that their preferences vary by context (Hanley et al., 2021; Stangl et al., 2020). The paper's claim is that the choice should be stated, because the evidence for and against "neutral description" differs by which goal is meant.

15. Conclusion

A description that withholds a gender label withholds an assertion. It does not withhold gender: the lexicon of appearance is gendered for readers, cues compound across a description, the describer's selection of what to mention carries the describer's perception, and an empty slot is filled by the reader's default. The anchor study measured the first of these for physical attributes, and the neighbouring literatures on scrubbed text, masked captions and unmarked prompts measured the rest.

Whether that is a failure depends on what "neutral" was meant to mean. For the accessibility guidance, which

sought to avoid ascription and to give blind and low-vision readers what sighted viewers have, leakage is the intended mechanism, and the open question is whether model describers transmit too much, too little or the wrong thing. For the fairness practice of removing labels so that a decision cannot depend on gender, leakage is the failure, and the evidence that it occurs is now extensive. Stating which goal a deployment holds is the precondition for testing it, and the parity test is the one the accessibility case still needs.

References

- Alexander, C. S., Storks, S., Pal, S., Chakrabarty, S., Sharma, A., Wesley, M.-T., & Russo, B. (2026). Identifying and Mitigating Gender Cues in Academic Recommendation Letters: An Interpretability Case Study. *arXiv:2604.12337*. doi:10.48550/arXiv.2604.12337.
- Belem, C. G., Seshadri, P., Razeghi, Y., & Singh, S. (2024). Are Models Biased on Text without Gender-related Language? *arXiv:2405.00588*. doi:10.48550/arXiv.2405.00588.
- Belrose, N., Schneider-Joseph, D., Ravfogel, S., Cotterell, R., Raff, E., & Biderman, S. (2023). LEACE: Perfect linear concept erasure in closed form. *arXiv:2306.03819*. doi:10.48550/arXiv.2306.03819.
- Bennett, C. L., Gleason, C., Scheuerman, M. K., Bigham, J. P., Guo, A., & To, A. (2021). It's Complicated: Negotiating Accessibility and (Mis)Representation in Image Descriptions of Race, Gender, and Disability. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1-19. doi:10.1145/3411764.3445498.
- Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., & Caliskan, A. (2022). Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. *arXiv:2211.03759*. doi:10.48550/arXiv.2211.03759.
- Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *arXiv:1607.06520*. doi:10.48550/arXiv.1607.06520.
- Bouchaud, P., & Ramaciotti, P. (2025). Linear socio-demographic representations emerge in Large Language Models from indirect cues. *arXiv:2512.10065*. doi:10.48550/arXiv.2512.10065.
- Burns, K., Hendricks, L. A., Saenko, K., Darrell, T., & Rohrbach, A. (2018). Women also Snowboard: Overcoming Bias in Captioning Models. *arXiv:1803.09797*. doi:10.48550/arXiv.1803.09797.
- Cao, Y. T., & Daume, H. (2020). Toward Gender-Inclusive Coreference Resolution. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4568-4595. *arXiv:1910.13913*. doi:10.48550/arXiv.1910.13913.
- De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., & Kalai, A. T. (2019). Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. *arXiv:1901.09451*. doi:10.48550/arXiv.1901.09451.
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011). Fairness Through Awareness. *arXiv:1104.3913*. doi:10.48550/arXiv.1104.3913.
- Elazar, Y., & Goldberg, Y. (2018). Adversarial Removal of Demographic Attributes from Text Data. *arXiv:1808.06640*. doi:10.48550/arXiv.1808.06640.
- Fast, E., Vachovsky, T., & Bernstein, M. S. (2016). Shirtless and Dangerous: Quantifying Linguistic Signals of Gender Bias in an Online Fiction Writing Community. *arXiv:1603.08832*. doi:10.48550/arXiv.1603.08832.
- Fraser, K. C., & Kiritchenko, S. (2024). Examining Gender and Racial Bias in Large Vision-Language Models Using a Novel Dataset of Parallel Images. *arXiv:2402.05779*. doi:10.48550/arXiv.2402.05779.
- Gao, B., & Kreiss, E. (2025). Measuring Bias or Measuring the Task: Understanding the Brittle Nature of LLM Gender Biases. *arXiv:2509.04373*. doi:10.48550/arXiv.2509.04373.
- Ghosh, S., & Caliskan, A. (2023). 'Person' = Light-skinned, Western Man, and Sexualization of Women of Color: Stereotypes in Stable Diffusion. *arXiv:2310.19981*. doi:10.48550/arXiv.2310.19981.
- Goldin, C., & Rouse, C. (2000). Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians. *American Economic Review*, 90(4), 715-741. doi:10.1257/aer.90.4.715.
- Gonen, H., & Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. *arXiv:1903.03862*. doi:10.48550/arXiv.1903.03862.
- Hamidi, F., Scheuerman, M. K., & Branham, S. M. (2018). Gender Recognition or Gender Reductionism?: The Social Implications of Embedded Gender Recognition Systems. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1-13. doi:10.1145/3173574.3173582.
- Hanley, M., Barocas, S., Levy, K., Azenkot, S., & Nissenbaum, H. (2021). Computer Vision and Conflicting Values: Describing People with Automated Alt Text. *arXiv:2105.12754*. doi:10.48550/arXiv.2105.12754.
- Hirota, Y., Nakashima, Y., & Garcia, N. (2022). Quantifying Societal Bias Amplification in Image Captioning. *arXiv:2203.15395*. doi:10.48550/arXiv.2203.15395.
- Holtgen, B., & Oliver, N. (2025). Reconsidering Fairness Through Unawareness From the Perspective of Model Multiplicity. *arXiv:2505.16638*. doi:10.48550/arXiv.2505.16638.
- Hossain, T., Dev, S., & Singh, S. (2023). MISGENDERED: Limits of Large Language Models in Understanding Pronouns. *arXiv:2306.03950*. doi:10.48550/arXiv.2306.03950.
- Hoyle, A., Wolf-Sonkin, L., Wallach, H., Augenstein, I., & Cotterell, R. (2019). Unsupervised Discovery of Gendered Language through Latent-Variable Modeling. *arXiv:1906.04760*. doi:10.48550/arXiv.1906.04760.
- Karvonen, A., & Marks, S. (2025). Robustly Improving LLM Fairness in Realistic Settings via Interpretability. *arXiv:2506.10922*. doi:10.48550/arXiv.2506.10922.
- Keyes, O. (2018). The Misgendering Machines. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 1-22. doi:10.1145/3274357.
- Kotek, H., Dockum, R., & Sun, D. Q. (2023). Gender bias and stereotypes in Large Language Models. *arXiv:2308.14921*. doi:10.48550/arXiv.2308.14921.
- Kreiss, E., Fang, F., Goodman, N. D., & Potts, C. (2021). Concladia: Towards Image-Based Text Generation with a Purpose. *arXiv:2104.08376*. doi:10.48550/arXiv.2104.08376.
- Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). Counterfactual Fairness. *arXiv:1703.06856*. doi:10.48550/arXiv.1703.06856.
- Luo, S., Kim, S. J., Duan, Z., & Chen, K. (2024). Refusal as Silence: Gendered Disparities in Vision-Language Model Responses. *arXiv:2406.08222*. doi:10.48550/arXiv.2406.08222.
- Marin-Llobet, A., Henniger, S., & Banaji, M. R. (2026). Vision-Language Models Suppress Female Representations Under Ambiguous Input. *arXiv:2605.31556*. doi:10.48550/arXiv.2605.31556.
- Ovalle, A., Goyal, P., Dhamala, J., Jaggers, Z., Chang, K.-W., Galstyan, A., Zemel, R., & Gupta, R. (2023). I'm fully who I am: Towards Centering Transgender and Non-Binary Voices to Measure Biases in Open Language Generation. *arXiv:2305.09941*. doi:10.48550/arXiv.2305.09941.
- Piergentili, A., Savoldi, B., Fucci, D., Negri, M., & Bentivogli, L. (2023). Hi Guys or Hi Folks? Benchmarking Gender-Neutral Machine Translation with the GeNTE Corpus. *arXiv:2310.05294*. doi:10.48550/arXiv.2310.05294.

- Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., & Goldberg, Y. (2020). Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. *arXiv:2004.07667*. doi:10.48550/arXiv.2004.07667.
- Scheuerman, M. K., Paul, J. M., & Brubaker, J. R. (2019). How Computers See Gender. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1-33. doi:10.1145/3359246.
- Scheuerman, M. K., Wade, K., Lustig, C., & Brubaker, J. R. (2020). How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1), 1-35. doi:10.1145/3392866.
- Sczesny, S., Formanowicz, M., & Moser, F. (2016). Can Gender-Fair Language Reduce Gender Stereotyping and Discrimination? *Frontiers in Psychology*, 7. doi:10.3389/fpsyg.2016.00025.
- Seshadri, P., Singh, S., & Elazar, Y. (2023). The Bias Amplification Paradox in Text-to-Image Generation. *arXiv:2308.00755*. doi:10.48550/arXiv.2308.00755.
- Staab, R., Vero, M., Balunovic, M., & Vechev, M. (2023). Beyond Memorization: Violating Privacy Via Inference with Large Language Models. *arXiv:2310.07298*. doi:10.48550/arXiv.2310.07298.
- Stangl, A., Morris, M. R., & Gurari, D. (2020). "Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1-13. doi:10.1145/3313831.3376404.
- Sun, T., Webster, K., Shah, A., Wang, W. Y., & Johnson, M. (2021). They, Them, Theirs: Rewriting with Gender-Neutral English. *arXiv:2102.06788*. doi:10.48550/arXiv.2102.06788.
- Tan, B. C. Z., Khoo, S., Doan, B. N., Liu, Z., Chen, N. F., & Lee, R. K.-W. (2026). Small Changes, Big Impact: Demographic Bias in LLM-Based Hiring Through Subtle Sociocultural Markers in Anonymised Resumes. *arXiv:2603.05189*. doi:10.48550/arXiv.2603.05189.
- Tavits, M., & Perez, E. O. (2019). Language influences mass opinion toward gender and LGBT equality. *Proceedings of the National Academy of Sciences*, 116(34), 16781-16786. doi:10.1073/pnas.1908156116.
- Vanmassenhove, E., Emmery, C., & Shterionov, D. (2021). NeuTral Rewriter: A Rule-Based and Neural Approach to Automatic Rewriting into Gender-Neutral Alternatives. *arXiv:2109.06105*. doi:10.48550/arXiv.2109.06105.
- Wagh, A., & Srivastava, S. (2025). "The Dentist is an involved parent, the bartender is not": Revealing Implicit Biases in QA with Implicit BBQ. *arXiv:2512.06732*. doi:10.48550/arXiv.2512.06732.
- Wan, Y., Lin, L., & Kreiss, E. (2026). How Humans and LLMs Read Gender into "Gender-Neutral" Physical Descriptions. *arXiv:2609.16366*. doi:10.48550/arXiv.2609.16366.
- Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K.-W., & Peng, N. (2023). "Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters. *arXiv:2310.09219*. doi:10.48550/arXiv.2310.09219.
- Wang, T., Zhao, J., Yatskar, M., Chang, K.-W., & Ordonez, V. (2018). Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. *arXiv:1811.08489*. doi:10.48550/arXiv.1811.08489.
- Wolfe, R., & Caliskan, A. (2022). Markedness in Visual Semantic AI. *arXiv:2205.11378*. doi:10.48550/arXiv.2205.11378.
- Xie, J., Yu, R., Zhang, H., Billah, S. M., Lee, S., & Carroll, J. M. (2025). Beyond Visual Perception: Insights from Smartphone Interaction of Visually Impaired Users with Large Multimodal Models. *arXiv:2502.16098*. doi:10.48550/arXiv.2502.16098.
- You, Z., Lee, H., Mishra, S., Jeoung, S., Mishra, A., Kim, J., & Diesner, J. (2024). Beyond Binary Gender Labels: Revealing Gender Biases in LLMs through Gender-Neutral Name Predictions. *arXiv:2407.05271*. doi:10.48550/arXiv.2407.05271.
- Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2017). Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. *arXiv:1707.09457*. doi:10.48550/arXiv.1707.09457.