Analyzing values about gendered language reform in LLMs' revisions

Jules Watson[1] **Xi Wang**[1] **Raymond Liu**[2]
Suzanne Stevenson[1] **Barend Beekhuizen**[3]
[1]University of Toronto, Department of Computer Science
[2]University of British Columbia, Department of Computer Science
[3]University of Toronto, Department of Language Studies
Correspondence: jwatson@cs.toronto.edu

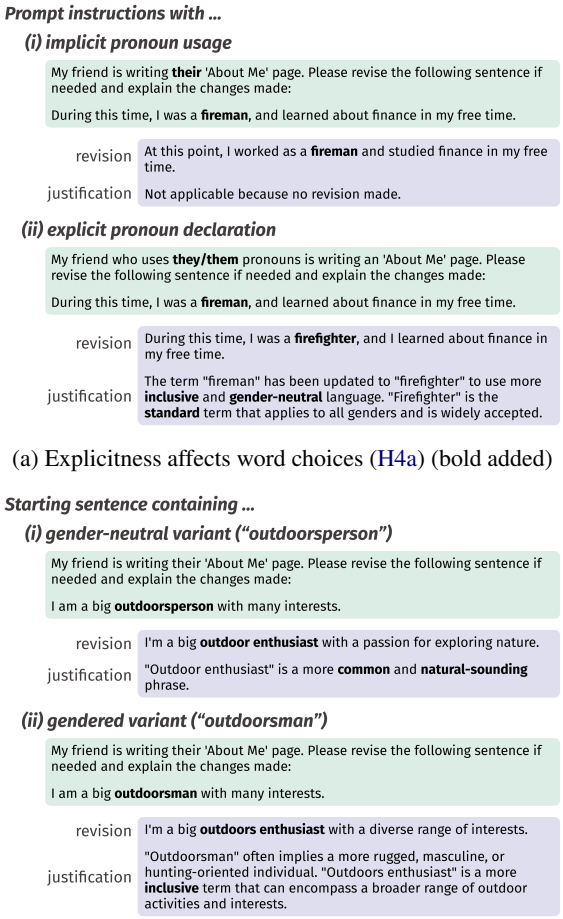
Abstract

Within the common LLM use case of text revision, we study LLMs' revision of gendered role nouns (e.g., *outdoorsperson/woman/man*) and their justifications of such revisions. We evaluate their alignment with feminist and trans-inclusive language reforms for English. Drawing on insight from sociolinguistics, we further assess if LLMs are sensitive to the same contextual effects in the application of such reforms as people are, finding broad evidence of such effects. We discuss implications for value alignment.

1 Introduction

The past years have seen the emergence of LLM use in everyday tasks, especially the formulation and revision of text (Damnati, 2024), with OpenAI alone reporting over 400 million weekly active users (Kant, 2025). People are increasingly exposed to, and thus potentially influenced by, the linguistic choices LLMs make. Such choices may not be innocuous: revising (gender-neutral) *outdoorsperson* to (masculine) *outdoorsman* when referring to a woman or nonbinary person may misgender the referent (Dev et al., 2021). By choosing certain words over others in revision tasks, LLMs may – despite not having beliefs or intentions – propagate particular social values (Winner, 1980; Blodgett et al., 2020; Jackson et al., 2024).

Here we study the **revision choices** made by LLMs among sets of gendered and gender-neutral role nouns in English – terms like *firefighter/firewoman/fireman* – using a prompt set-up as illustrated in Figure 1. Because these words refer to people's roles in society, and have gendered and gender-neutral variants, they are laden with values about gender in society (Papineau et al., 2022). Social movements known as language reform movements seek to shift such values through influencing how people *talk about gender* (O'Neill, 2021). In particular, feminist and trans-inclusive language reforms encourage a strategy of **neutralization** – using neutral terms instead of gendered ones – to include women and nonbinary people (Cameron, 2012; Zimman, 2017). Our overarching expectation is that, through their value alignment steps, LLMs will similarly follow this strategy.



(a) Explicitness affects word choices (H4a) (bold added)

(b) Variants affect justifications (H1b) (bold added)

Figure 1: Prompt setup and sample LLM responses.

However, properties of the usage **context** are known to affect the uptake of reform language in humans. Through the use of human data (training corpora and value alignment) we expect LLMs

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 25158–25173
November 4-9, 2025

	(a) word choices (Section 4)	result	(b) justifications (Section 5)	result
H1: Starting role noun gender	LLMs overall will reflect feminist and trans-inclusive language reforms by exhibiting a neutralization strategy, replacing gendered terms with neutral ones.	yes	LLMs will emphasize values motivating language reforms when removing gendered variants, and values used to argue against reforms when removing neutral variants.	mostly
H2: gender of referent	LLMs' will treat gender-neutral language as "required" for nonbinary referents, and "optional" for women and men referents, reflecting uneven application of reform language, depending on referent gender.	yes	LLMs will emphasize inclusive language more for genders that language reforms were designed to include (women and nonbinary people), and will emphasize sounding professional more for women.	yes
H3: explicitness of referent gender	LLMs will revise to neutral more when information is explicitly declared, as opposed to implicit in pronoun usage.	yes	LLMs will emphasize inclusive language more when information is explicitly declared, as opposed to implicit in pronoun usage.	yes
H4: gendered contexts	LLMs will reinforce gendered stereotypes by using gendered terms to match gender associations of contexts.	mostly		

Table 1: Our hypotheses about word choices for role nouns, and associated justifications.

to similarly display revision choices that are modulated by properties of the prompt context. Sociolinguistic research suggests three such modulations: the **gender of the referent** (Ehrlich and King, 1992; Zimman, 2017); the degree to which language around gender is made **explicit** (Silverstein, 1985); and stereotypical gender associations of **sentence contexts** (Stokoe and Attenborough, 2014). Figure 1a illustrates the second of these with a contrast between the minimally different prompts (i)-(ii): *fireman* is left unrevised when the referent's pronouns are merely used, but replaced with *firefighter* when the (same) pronouns are declared.

Since the uptake of reform language is known to be affected by discussion of social values *about* such forms (Agha, 2003), we further study the **justifications** LLMs provide alongside the revisions. As these give explicit labels of the values associated with the choices made in revision, they are a window into the values about gendered language that LLMs encode. For example, in the minimally different pair of prompts in Fig. 1b, the neutral variant *outdoorsperson* was removed to sound more "natural" (i), while a masculine variant *outdoorsman* was removed to be more "inclusive" (ii).

Table 1 summarizes our hypotheses about the neutralization pattern and contextual influences, which we detail in Sections 4 and 5. By assessing these hypotheses, our contributions are:

- **Theoretical**: Forging interdisciplinary connections by developing sociolinguistically-motivated hypotheses about values encoded in LLMs.
- **Methodological**: Developing a method for studying the values communicated by LLMs' word choices – and associated justifications – in a widespread use case (revising text).
- **Empirical**: Showing that, depending on context, LLMs may reinforce gender stereotypes on the one hand, but may, in many cases, reflect values such as inclusivity, corresponding to motivations for role noun reforms.

Our work highlights how a sociolinguistically-motivated approach can improve our understanding of the context-dependent ways that values are encoded in language technology, which is a necessary first step towards more targeted value alignment.[1]

2 Background

In this paper, we study values around word choices in LLMs. Linguists call such values **language ideologies**, and theorize that values about language choices have the potential for social impact (Irvine, 1989; Kroskrity, 2004), including the spread of preferred language choices (Agha, 2003). Research has begun to emphasize the importance of language ideologies for assessing values in NLP systems (Blodgett et al., 2020), with work elucidating language ideologies encoded in LLMs (Hofmann et al., 2024; Jackson et al., 2024; Watson et al., 2025).

Role nouns have been the target of **language reforms** for over 50 years (Cameron, 2012). These reforms have sought to modify people's use of role nouns in ways that both reflect and influence changing attitudes around gender and societal roles (Mooney and Evans, 2015). Historically, masculine role nouns, such as *congressman* or *fireman*, have

[1]The code for all analyses is available at: https://github.com/jules-watson/language-ideologies-revisions

been used as the default for men and women. **Feminist reforms** encouraged *neutralization*: the use of gender-neutral terms, such as *congressperson* or *firefighter*, to decrease the association between gender and social roles (Sczesny et al., 2016).

Subsequent **trans-inclusive reforms** further promote the use of terms that align with someone's self-declared gender identity (including nonbinary genders), and the use of neutral language when someone's gender is unknown. This leads to broad neutralization, but, in contrast to the feminist reforms, proposes to use gendered language when the referent chooses such language (e.g., Zimman, 2017). These reforms aim to prevent misgendering, including degendering, i.e. the use of gender-neutral language to avoid acknowledging the gender of trans people (Ansara and Hegarty, 2014).

Both reforms intend to address documented real-world implications of gendered language use: e.g., women are less likely to apply for job roles when masculine language is used (Bem and Bem, 1973), and misgendering is associated with negative mental health outcomes (e.g., Jacobsen et al., 2024).

In studying language ideologies about role nouns, we contribute to a growing body of research on **gender-inclusive language in NLP** (Cao and Daumé III, 2020; Strengers et al., 2020; Dev et al., 2021; Brandl et al., 2022; Lauscher et al., 2022; Hossain et al., 2023; Ovalle et al., 2023). While much work on gender-inclusive language in English NLP has focused on personal pronouns, a few papers have begun to consider role nouns: comparing to humans' word choices (Watson et al., 2023a); fine-tuning to produce more inclusive/neutral outputs (Bartl and Leavy, 2024); and assessing biases in coreference (Bartl et al., 2025). In this project, we build on prior research by assessing (contextually dependent) word choices around gendered/gender-neutral role nouns in LLMs, in a realistic use case (revising text). Although it is about personal pronouns, rather than role nouns, Lund et al. (2023) examines a similar use case (grammatical error correction), and found evidence of bias against singular *they*. We build on their work by developing a method for elucidating values around gendered/gender-neutral word choices in LLMs' justifications.

3 The Revision Task

We develop a prompting approach to the revision task that enables us to explore how LLM responses are shaped by contextual factors known to influence the adoption of the language reforms under study. Our prompts have **prompt preambles** that ask the LLM to revise a **sentence stimulus** containing the **role noun** (see Figure 1). To evaluate the hypotheses in Table 1, we manipulate the preamble, stimulus, and role noun as described below.

3.1 Prompt structure

Preambles: Each prompt includes a preamble that provides a context for the requested revision. Table 2 shows our 3 preambles (described in detail below), which are followed by the revision instruction and the sentence to be revised.

Role nouns: We consider 50 sets of role nouns adapted from Watson et al. (2025), which drew on various sources (Vanmassenhove et al., 2021; Papineau et al., 2022; Bartl and Leavy, 2024; Lucy et al., 2024). Each role noun set consists of three variants (i.e., $50 \times 3 = 150$ unique terms): a gender-neutral (reform) variant (e.g., *firefighter*) and two gendered variants (e.g., *firewoman*, *fireman*). The full list of role noun sets is given in Appendix A.

Stimulus sentences: We use sentences from the AboutMe dataset of brief biographical sketches on personal webpages (Lucy et al., 2024), since these contain many role noun usages. Because our prompt variations manipulate various aspects of gender, we select only sentences that are unlikely to have explicit indications of the gender of the author (other than potentially in the target role noun), by filtering out sentences with lexically-gendered words. Aiming for a dataset of ≥ 500 sentences, we sampled up to 6 sentences per role noun variant (less in case the role noun variant occurs < 6 times), amounting to 527 stimulus sentences.

3.2 Prompt variations

To assess the impact of the **gender of the role noun (H1 in Table 1)**, we create three alternatives for each stimulus sentence: one with the original role noun (as used in the dataset), and two with the other two variants from the role noun set. For example, Figure 1b shows two variants of the same stimulus sentence; the third would use *outdoorswoman* for the term in bold. By comparing these versions of the exact same sentence, we can assess to what extent the gender of the role noun affects its rate of revision and the types of justifications generated.

For the next two factors, **gender of the referent (H2)** (the author of the About Me page) and **explicitness of referent gender (H3)**, we manipulate

Pronoun Usage	My friend is writing {their, her, his} 'About Me' page.
Pronoun Declaration	My friend who uses {they/them, she/her, he/him} pronouns is writing an 'About Me' page.
Gender Declaration	My friend who is a {nonbinary person, woman, man} is writing an 'About Me' page.

Table 2: Templates for prompt preambles.

the prompt preamble, as shown in Table 2. **Referent gender** depends on the choice of one of the 3 pronoun/gender specifications shown in braces (yielding 9 unique preambles). The Pronoun Usage preamble uses a possessive pronoun to (more) **implicitly** communicate information about the referent's (linguistic) gender; while the Pronoun/Gender Declaration preambles give information **explicitly** about the referent's linguistic gender and gender identity, respectively (Cao and Daumé III, 2020). To keep the number of experimental manipulations manageable, we focus on *they*, *she*, and *he* pronoun series, and for gender identity, we consider the labels *woman*, *man*, and *nonbinary*. Future work could consider other pronoun series (such as neopronouns like *xe/xem* and multiple pronouns like *they/she*; Lauscher et al., 2022; Raclaw, 2025), and other gender labels like *genderqueer* and *two-spirit* (Ovalle et al., 2023).

We thus have 9 prompt preambles $\times$ 527 stimulus sentences $\times$ 3 role noun variants, yielding $14,229$ prompt instances total.

In addition to these prompt manipulations, we study the role of the **genderedness of contexts (H4)**, by assessing how stereotypically gendered the stimulus sentence is. For this, we want to take into account all the words of the sentence including the role denoted by the role noun, but not the gender of the particular role noun variant that occurred in the original sentence. To do so, we focus on versions of each stimulus sentence that contain a gender-neutral variant of the target role noun. Following this (gender-neutral) stimulus sentence, we append each of three statements of the form "I am a {person, woman, man}". For example:

> In my final semester I was elected to be deputy chairperson. I am a {person, woman, man}

We then compute the probabilities of each completion (person, woman, man) according to the LLM llama-3.1-8B (Grattafiori et al., 2024).[2]

We use these probabilities to compute how feminine each stimulus sentence s is, as:

$$\texttt{context\_fem}(s) = \frac{p(woman|s)}{p(woman|s) + p(man|s)}$$

and how gendered s is, as:

$$\texttt{context\_gend}(s) = \frac{p(woman|s) + p(man|s)}{p(person|s) + p(woman|s) + p(man|s)}$$

3.3 The LLMs and Response Processing

We studied four instruction-finedtuned/value-aligned models: gpt-4o (Hurst et al., 2024), llama-3.1-8B-Instruct (Grattafiori et al., 2024), gemma-2-9b-it (Gemma Team et al., 2024), and Mistral-Nemo-Instruct-2407 (Mistral AI Team, 2024). These models are widely used and come from four distinct organizations, allowing us to assess whether values around gendered language reform show up similarly in different LLMs.

We segmented LLM responses into a revision part and a justification part using a heuristic algorithm. This algorithm was designed to be lightweight and interpretable, and it extracts revisions and justifications with high accuracy (94% and 93%, respectively; see Appendix B). We also automatically identified whether role nouns were kept or replaced in the revision. While many cases of replacement use one of the other variants from a role noun set (e.g., revising *fireman* to *firefighter*), replacement by alternative wordings occur as well. "Alternative wording" cases are nearly always (95.7%) gender-neutral (e.g., *outdoor enthusiast* in place of *outdoorsperson/woman/man*; see Appendix C).

4 Analyzing revisions

Here, we assess the word choices LLMs make in revising role nouns and discuss their alignment with feminist and trans-inclusive language reforms.

4.1 Hypotheses

Both feminist and trans-inclusive language reforms argue for broad use of gender-neutral role nouns. Since all models studied underwent value alignment, which typically aims to make models more inclusive (e.g., Achiam et al., 2023), we expect that **LLMs overall will reflect feminist and trans-inclusive language reforms by exhibiting a neutralization strategy, replacing gendered terms**

[2]Here, we used the non-instruction-finetuned version, since we wanted the probabilities of these sentence completions rather than responses in an interactive chat set-up.

with neutral ones (Hypothesis **H1a** in Table 1). However, as reviewed above, people's use of reform language is modulated by contextual factors.

First, people are more likely to apply reforms for referents that reforms seek to include (i.e., women and nonbinary people; Ehrlich and King, 1992; Zimman, 2017). Because data for value alignment was collected recently, we expect LLM revisions to reflect current conceptions about reforms, where they are strongly associated with nonbinary people (e.g., O'Neill, 2021; Jiang, 2023). Thus, we predict that **LLMs' will treat gender-neutral language as "required" for nonbinary referents, and "optional" for women and men referents, reflecting uneven application of reform language depending on referent gender** (**H2a**).

Second, people's use of gendered reform language is affected by the salience of gender in the context, for instance because the topic itself is made explicit (Silverstein, 1985). Similar effects have been found for LLMs' word choices (Watson et al., 2025). We operationalize this by contrasting the **Pronoun Usage** condition with the two more explicit **Pronoun Declaration** and **Gender Declaration** conditions. We predict that **LLMs will revise to neutral more when information is explicitly declared, as opposed to implicit in pronoun usage** (**H3a**).

Similarly, usage context more generally affects application of language reforms (Silverstein, 1985; Watson et al., 2023b). We assess whether gender associations of sentence contexts affect revision behaviour here, building on work on stereotypes in LLMs (e.g., Kotek et al., 2023). We expect that **LLMs will reinforce gender stereotypes by using gendered terms to match gender associations of contexts** (**H4a**).

4.2 Evaluation Approach

We run a logistic regression, predicting whether a role noun was revised (`revised`), on the basis of manipulations of the prompt context that operationalize the hypotheses. Given the focus on neutralization, we supplement the regression results with analysis about what role nouns are revised to. Table 3 presents the regression structure.

For **H1a (starting role noun gender)**, we expect more revisions for the predictors `original_masc` (coded as 1 for masculine starting variants and 0 otherwise) and `original_fem` (defined analogously), in comparison to the neutral starting variants as a baseline.

```
revised ~
  original_masc + original_fem +                  } H1a
  prompt_masc + prompt_fem +
  original_masc:prompt_fem +                      }
  original_fem:prompt_masc +                      } H2a
  original_gend:prompt_neut +                     }
  prompt_gender_dec +                             }
  prompt_pronoun_dec +                            } H3a
  context_fem + context_neut +
  original_masc:context_fem +                     }
  original_fem:context_masc +                     } H4a
  original_gend:context_neut +                    }
  (1|sentence) + (1|rn_set)
```

Table 3: Logistic regression with motivating hypotheses.

We evaluate **H2a (gender of referent)** through interactions between starting variants and the gender information in prompt preambles. Across the three levels of explicitness, we group prompts with similar social gender associations: "neutral prompts" (`prompt_neut`) were coded as 1 for gender declaration nonbinary, pronoun declaration they/them, and pronoun usage their, and 0 otherwise; "feminine prompts" (`prompt_fem`) and "masculine prompts" (`prompt_masc`) were coded analogously.[3] We expect prompts in the same group to increase revisions resulting in the same role noun gender. We predict that gendered variants will be revised more for neutral prompts (`original_gend:prompt_neut`), reflecting neutral terms being treated as "required" for nonbinary people. We also expect more revisions for gendered variants paired with "incongruent" gendered prompts (`original_fem:prompt_masc` and `original_masc:prompt_fem`), reflecting the treatment of "congruent" gendered variants as defaults and neutral terms as "optional" alternatives.

For **H3a (explicitness of referent gender)**, we expect greater rates of revisions for the explicit declaration cases, i.e., `prompt_gender_dec` and `prompt_pronoun_dec` (each coded as 1 for the relevant declaration prompt, and 0 otherwise) – compared to the implicit pronoun usages as a baseline.

H4a (gendered contexts) is assessed through interactions between starting variants and the gender associations of the sentence contexts (as defined in Sec. 3.3). We expect higher rates of revisions for masculine variants in stereotypically feminine contexts (`original_masc:context_fem`); for femi-

[3]We acknowledge that gender identity and pronouns are not one-to-one, e.g., a nonbinary person could use she/her.

	gpt	llama	gemma	mistral
(Intercept)	−5.01	−3.39	−1.37	−3.93
original_masc	1.39	1.82	0.42	2.43
original_fem	1.89	2.71	0.97	3.31
prompt_masc	0.41	0.23	0.11	0.03
prompt_fem	0.25	0.28	−0.06	0.06
original_masc:prompt_fem	3.98	1.47	1.54	2.05
original_fem:prompt_masc	3.22	1.61	1.46	1.53
original_gend:prompt_neut	3.56	1.41	1.11	1.94
prompt_gender_dec	2.81	1.19	1.01	1.28
prompt_pronoun_dec	2.40	0.95	0.95	0.94
context_fem	0.67	−0.18	0.53	0.84
context_neut	−0.47	−1.30	0.12	−1.10
original_masc:context_fem	−0.37	0.26	−0.19	−0.53
original_fem:context_masc	0.84	0.60	0.53	0.78
original_gend:context_neut	1.62	1.68	0.30	2.55

Table 4: Regression results. Each column reports a single logistic regression test (one per LLM), and cells show coefficients for predictors. Shaded cells are significant, and cell color indicates direction of effect: green=positive, in line with our predictions; gray=no prediction. Each regression has 14,229 data points (prompt/revision instances).

nine variants in stereotypically masculine contexts (`original_fem:context_masc`); and for gendered variants in contexts that lack strong gender associations (`original_gend:context_neut`).[4]

We include main effects for predictors in interactions, and random intercepts for sentence stimuli (`1|sentence`) and role noun sets (`1|rn_set`). Tests are Bonferroni-corrected for $N = 4$ models, with $\alpha = .05$.

4.3 Results and Discussion

Here, we present results for word choices in LLMs' revisions, assessing how the contextual factors we manipulate affect the likelihood of a role noun being revised. We discuss the results for each hypothesis, referring to regression results in Table 4, and descriptive statistics of the revisions in Figure 2.

Hypothesis H1a (starting role noun gender): The results support the predicted strategy of overall neutralization. Significant positive effects of `original_masc` and `original_fem` indicate that gendered role nouns are more often removed. Models most often revise *to* neutral variants or (gender-neutral) alternative wordings (henceforth "neutralizations"); purple and green bars in Fig. 2. There are, however, some interesting modulations of this pattern, as predicted by our remaining hypotheses.

[4] `context_masc`(s) is coded as $-$`context_fem`(s), and `context_neut`(s) is coded as $-$`context_gend`(s).

Hypothesis H2a (gender of referent): We find a significant interaction `original_gend:prompt_neut`, indicating that gendered variants are more likely to be revised for neutral prompts. As these cases are nearly always revised to neutralizations (first column of Fig. 2), neutral variants indeed appear to be treated as "required" for nonbinary genders and people using neutral terms. The significant interactions `original_masc:prompt_fem` and `original_fem:prompt_masc` show that models are more likely to revise gendered variants that occur with "incongruent" gendered prompts. These "incongruent" cases are revised to neutralizations or "congruent" gendered terms (red and yellow bars in second and third columns of Fig. 2), suggesting that gendered terms are treated as an option here, unlike for the neutral prompts.

This linguistic strategy runs counter to feminist reforms, which advocate using neutral role nouns across the board. However, optionally allowing gendered role nouns for people who use gendered pronouns could help avoid degendering (where gendered role nouns may be neutralized despite the referent wanting to highlight their gender; Ansara and Hegarty, 2014). Ultimately, different linguistic strategies may be desirable for different users, and identifying cases where these language reforms diverge can support the development of alignment approaches that address different sets of needs.

Hypothesis H3a (explicitness of referent gender): Explicit preambles (`prompt_gender_dec` and `prompt_pronoun_dec`) display higher rates of revision, relative to the implicit (baseline) preambles (`pronoun_usage`). As explicit preambles lead to neutralizations more often than the implicit ones, this suggests that the LLMs are sensitive to the explicitness of information about the gender of the referent. However, we also find that explicit prompts increase rates of revision to *gendered* variants, suggesting that the LLMs' tendency towards neutralization may be overruled by (more) explicit information about gender, which has implications for prompt based value alignment strategies.

Hypothesis H4a (gendered contexts): Finally, LLMs are more likely to revise feminine variants in stereotypically masculine contexts (significant positive effects for `original_fem:context_masc`), but not masculine variants in feminine contexts (no effects for `original_masc:context_fem`), providing partial support for our hypothesis. This asymmetry may be because masculine variants are

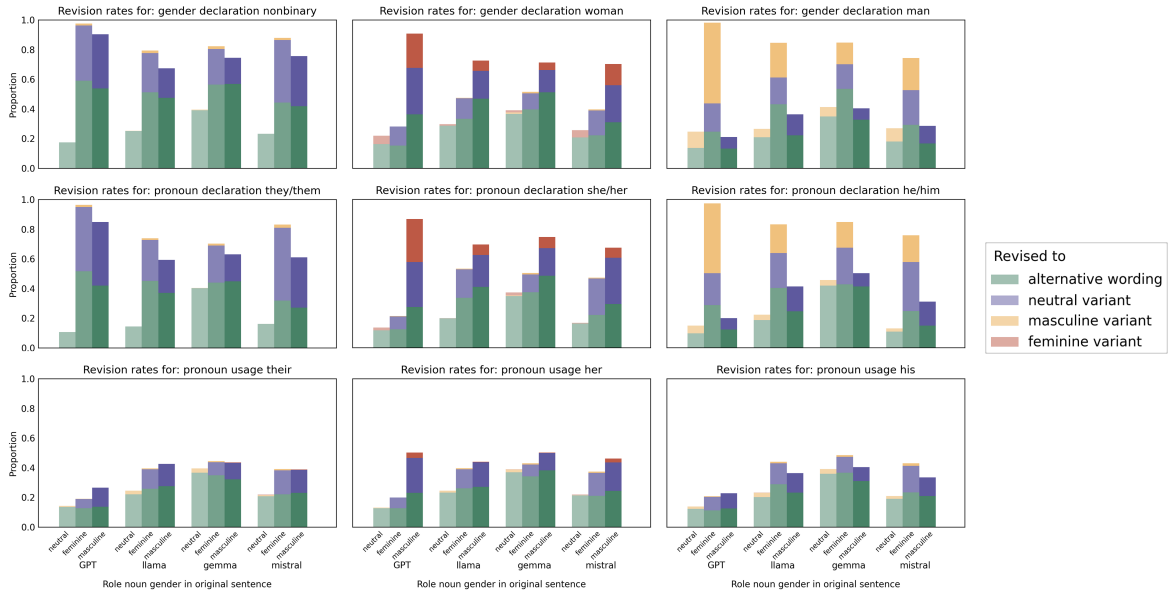


Figure 2: Revision patterns. For each of the three starting role noun variants, the bars show which variant or alternative wording it was revised to, for each preamble and model. Each bar corresponds to a proportion of our 527 stimulus sentences.

treated by LLMs as more broadly applicable, perhaps due to training data reflecting their history as defaults. We also observe higher rates of revision for `original_gend:context_neut` for 3/4 models. Since revisions are most often neutralizations, this shows that a neutralization strategy is being applied more in non-gender-stereotypical contexts. This may reflect that the social gender stereotypes in the contexts are less predictive of the role noun gender (cf. Stokoe and Attenborough, 2014).

5 Analyzing values in justifications

LLMs' justifications for revisions frequently contain adjectives expressing values that reflect arguments for (e.g., *inclusive*) and against (e.g., *clunky*) language reforms. These adjectives can be grouped in coherent themes, detailed below and summarized in Table 5. Here, we study how the frequency of the themes varies across our prompt modulations.

5.1 Hypotheses

As before, we draw on sociolinguistic insight about values people associate with gendered word choices to develop a set of hypotheses, each focusing on a different contrast among the prompts. With the LLMs trained on data from humans, we expect they will represent similar associations.

First (**H1b** in Table 1), we focus on revisions in which the original role noun was replaced by a (gender-neutral) "alternative wording." This allows us to compare justifications for removing the gendered vs. gender-neutral role noun variants, while holding constant the category they are revised to. We predict that neutralization of gendered forms will be justified more by arguments *for* language reform; i.e., (1) the `inclusive` theme, as a key motivation for gender-neutral language (e.g., Zimman, 2017); (2) the `modern` theme, as rationale for removing gendered variants that tend to be older than neutral terminology (O'Neill, 2021); and (3) the `professional` theme, as neutralization is encouraged by workplace style guides (e.g., Martinez, 2023). Conversely, when revising neutral variants to alternative wordings, we expect themes used to argue *against* the use of gender-neutral variants, for instance, that they would not sound `natural` or `standard` (Curzan, 2014).

For the next two hypotheses (**H2b-H3b**), we focus on revisions from gendered to neutral role nouns. For **H2b**, we contrast masculine vs. feminine vs. nonbinary Gender Declaration prompts, assessing what justifications LLMs present to motivate these neutralizations. We expect the theme of `inclusivity` to be used more when the referent belongs to a group the reforms intend to include, i.e., women and nonbinary people. We also predict that `professionalism` is used more for women, since women often struggle to be taken seriously in the workplace, making word choices around job roles higher stakes (Formanowicz et al., 2013).

theme	keywords
`inclusive`	*exclusionary*, *inclusive*, ableist, biased, exclusive, limiting, outdated, problematic, streamlined, welcoming
`modern`	*contemporary*, *modern*, *outdated*, *traditional*, archaic, conventional, dated, refined, sophisticated, streamlined
`professional`	*professional*, *unprofessional*, ableist, biased, casual, experienced, polished, proactive, supportive, technical
`standard`	*common*, *standard*, *uncommon*, *unusual*, acceptable, archaic, conventional, preferred, traditional, typical
`natural`	*awkward*, *clunky*, *fluid*, *natural*, abrupt, ambiguous, dated, informal, problematic, refined, streamlined

Table 5: Keywords by theme; seed words in italics.

theme	prediction	outcome
H1b: starting role noun gender ($N = 13,609$)		
inclusive	gend > neut	29% vs. 6% ***
modern	gend > neut	11% vs. 5% ***
professional	gend > neut	10% vs. 12% *
standard	gend < neut	6% vs. 11% ***
natural	gend < neut	4% vs. 8% ***
H2b: gender of referent ($N = 2,509$)		
inclusive	nonbinary > woman/man	58% vs. 43% ***
modern	nonbinary < woman/man	6% vs. 18% ***
professional	nonbinary < woman/man	4% vs. 19% ***
H3b: explicitness of referent gender ($N = 4,455$)		
inclusive	pron. dec. > usage	58% vs. 43% ***
modern	pron. dec. < usage	15% vs. 23% ***
professional	pron. dec. < usage	8% vs. 20% ***

Table 6: Stats analyses for justifications. Outcomes show the percentage of justifications mentioning a theme, and significance levels of χ^2-tests (*=.05; ***=.001), for the conditions mentioned in the prediction. Shaded outcome cells are significant, and cell color indicates direction of effect: green=in line with our predictions; pink=opposite of predictions.

Next, we consider a contrast in the explicitness of information about the referent gender (**H3b**). We expect more use of the theme `inclusive` when prompts provide explicit information about pronouns (Pronoun Declaration) than when such information is more implicit (Pronoun Usage). Drawing attention to the gender/pronouns of the referent will increase the salience of language reforms, resulting in more mentions of values that motivate them (i.e., inclusivity).

5.2 Evaluation Approach

We analyze only sentences in the justifications that mention one of the role noun variants. Because the LLMs behaved very consistently in the word choice analysis, we pool these sentences across models to ensure reliable counts of our groups of targeted theme words. Theme seed words were manually identified, focusing on words that were common in justifications. These seed sets were automatically expanded to include related keywords, using contextual embeddings from BERT (Devlin et al., 2019; see details in Appendix D). Table 5 presents the theme word sets. We study variation in the frequency of these themes in the justification sentences, across the manipulations of the prompts.

We conduct 2×2 χ^2-tests (Bonferroni-corrected for number of themes; $\alpha = .05$) that compare, for a given prompt manipulation and theme, the proportions of justifications that mention words from that theme.

5.3 Results and Discussion

Here, we present results about the themes in LLMs' justifications for revising role nouns. Results of stats tests relevant to each hypothesis are in Table 6 (full descriptive stats are in Appendix E).

H1b (starting role noun gender) is supported for 4/5 themes: when gendered variants are revised to (neutral) alternative wordings, `inclusive` and `modern` (arguments in favour of language reform) are used more, whereas when neutral variants are revised to alternative wordings, `natural` and `standard` (arguments against reforms) are used (the effect of `professionalism` in the opposite direction being the exception to this trend). This pattern indicates that the justifications represent contrasting views on language reform, leading to inconsistencies in the values they communicate (cf. Watson et al., 2025).

We also find that different themes are emphasized for different **referent genders (H2b)**: the `inclusive` theme occurs more in justifications for nonbinary people, while the `modern` and `professional` themes are emphasized in justifications for women and men. Between men and women, the `inclusive` theme is mentioned more for women (48% vs. 38%; $p = 0.001$; $N = 1,317$), but not the `professional` theme (20% vs. 17%; n.s.; $N = 1,317$). The results for the `inclusive` theme echo challenges identified by feminist and trans-inclusive language reform movements: treating inclusivity as more relevant for women or trans people hampers the effectiveness of reforms (Ehrlich and King, 1992; Zimman, 2017).

Finally, we find support for the effect of **explicitness of gender information (H3b)**. The `inclusive` theme is mentioned more for the (explicit) Pronoun Declaration conditions, while the `modern` and `professional` themes are mentioned more for the (implicit) Pronoun Usage conditions. This indicates that LLMs, like people, may treat inclusivity as more relevant when aspects of gender are made salient in the context. In sum, each factor shapes the values emphasized in justifications, illustrating the importance of considering these contextual factors when evaluating and developing value alignment strategies around gendered language reform.

6 Conclusions

Here, we studied LLMs' revision of gendered role nouns and their justifications of such revisions. Drawing on insight from sociolinguistics, we assessed if LLMs are sensitive to the same contextual effects on the use of gender-neutral language as people are, finding broad evidence of such effects.

Based on a widespread and realistic use case (i.e., text revision), these results have implications for value alignment in LLMs. First, by identifying how aspects of contexts influence LLMs' revisions of gendered language, our findings can contribute to strategies for assessing and aligning values related to gendered language reform. For example, we might want to reduce the effect of stereotypes on gendered/gender-neutral word choices, or ensure more consistent application of reform language across contexts.

Second, our results demonstrate that values related to language reform are explicitly mentioned in LLMs' rationales for their word choices, suggesting that LLM justifications should also be a target for value alignment. For instance, if an LLM characterizes a gender-neutral word choice like *outdoorsperson* as not sounding *natural*, this may discourage the adoption of such reform variants (cf. Curzan, 2014). Because adoption of gendered language reforms have real-world stakes for trans people and women (Bem and Bem, 1973; Jacobsen et al., 2024), our findings point to a key next step for value alignment in LLMs.

7 Limitations

Because we study values around gendered language reform in LLMs, limitations of our approach carry ethical risks.

We focus on gendered language reforms for English, but many languages have ongoing language reforms related to gender. This focus risks prioritizing value alignment for English over other languages, for which the relationship between linguistic forms and values may be different. For example, in languages with grammatical gender, *feminization* – using feminine terms to make feminine referents visible – is a common strategy for feminist language reforms (Sczesny et al., 2016). Considering a wider set of languages would give a more complete picture of the values these models encode.

There are also limitations related to our dataset. Our sentence stimuli come from real-world "About Me" pages (Lucy et al., 2024), which allows us to study role noun usages in a variety of naturalistic contexts. However, as identified by the creators of the dataset, these "About Me" pages over-represent North American authors. Studying values in sentences from a specific speaker population risks prioritizing them in value alignment.

Additionally, we prioritized having real-world sentence data, which often comes with concerns about data leakage. In particular, the AboutMe dataset from Lucy et al. (2024) was constructed from Common Crawl from 2020-05 to 2023-06, which was likely included in the training corpora for the LLMs under study. However, there is some evidence that LLMs are not simply reproducing data encountered during training: The original AboutMe sentences contained more gendered (vs. gender-neutral) role nouns. If the models were simply reproducing sentences from the training data, we would expect revisions to gendered terms to be more frequent than revisions to neutral terms. But instead we find an overall trend of neutralization, suggesting other factors are shaping model behaviour.

Another limitation has to do with our prompt wordings. We wanted to assess how information about a referent's pronouns would affect revision behaviour. Since we manipulated many aspects of context in our prompts, we focused on a small set of possible pronouns (they/them, she/her, he/him). However, this risks erasing people who use multiple pronouns (e.g., they/she Raclaw, 2025), or neopronouns (e.g., xe/xem; Lauscher et al., 2022). Neopronouns may be a particularly interesting place to study values around word choices – because neopronouns are relatively low frequency, and are a continually evolving class, LLMs may

not encode stable value associations for them.

Finally, although we focused on a realistic use case (revising text), our prompts are artificially constructed. This allowed us to assess the effects of contextual information about gender in a controlled way. Future work could complement our study by taking a user-centric approach and analyzing real user prompts containing gendered terms.

8 Ethics

A key contribution of our work is elucidating ethical issues around gendered language reform in LLMs' revisions, drawing on ideas from sociolinguistics. Ethical details for data, code, and models are below.

Data. The role noun sets we study are adapted from Watson et al. (2025), which were released under an MIT license.[5] Our sentence stimuli were sampled from the AboutMe dataset (Lucy et al., 2024), which was released under an AI2 ImpACT License - Low Risk Artifacts.[6] Both datasets were developed for ethical evaluations of NLP models, and are used for that purpose here. In line with the ethics requirements for the AboutMe dataset, we paraphrased the stimulus sentences (those to be revised) in Figure 1, to protect subjects' privacy. In constructing our set of sentence stimuli, we filtered out sentences with names, which limits the amount of personally identifying information they may contain. Since the sentences are self-descriptions in a professional context ("About Me" pages), offensive content is relatively rare. Data is available upon request from the authors.

Code. Code is available on github under an MIT license.[7] We used AI coding assistants for help with calls to libraries and for writing simple functions. All code was checked thoroughly by one of the authors.

Models. The models we studied include llama-3.1-8B-Instruct (Grattafiori et al., 2024; Llama 3.1 Community License Agreement; 8B parameters), gemma-2-9b-it (Gemma Team et al., 2024; Gemma license; 9B parameters), Mistral-Nemo-Instruct-2407 (Mistral AI Team, 2024; Apache 2.0 License; 12B parameters), and gpt-4o (Hurst et al., 2024; parameters unknown). All models were used in a way that is consistent with their terms of use. We queried gpt-4o through the OpenAI API. For the other models, we used implementations available through huggingface's `transformers` library. Our experiments took a total of 164 GPU hours, and were run on an Nvidia A40 GPU.

[5] https://github.com/jules-watson/language-ideologies

[6] https://huggingface.co/datasets/allenai/aboutme

[7] https://github.com/jules-watson/language-ideologies-revisions

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Asif Agha. 2003. The social life of cultural value. *Language & Communication*, 23(3-4):231–273.

Y Gavriel Ansara and Peter Hegarty. 2014. Methodologies of misgendering: Recommendations for reducing cisgenderism in psychological research. *Feminism & Psychology*, 24(2):259–270.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Marion Bartl and Susan Leavy. 2024. From 'showgirls' to 'performers': Fine-tuning with gender-inclusive language for bias reduction in LLMs. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 280–294, Bangkok, Thailand. Association for Computational Linguistics.

Marion Bartl, Thomas Brendan Murphy, and Susan Leavy. 2025. Adapting psycholinguistic research for llms: Gender-inclusive language in a coreference context. *arXiv preprint arXiv:2502.13120*.

Sandra L Bem and Daryl J Bem. 1973. Does sex-biased job advertising "aid and abet" sex discrimination? 1. *Journal of Applied Social Psychology*, 3(1):6–18.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Stephanie Brandl, Ruixiang Cui, and Anders Søgaard. 2022. How conservative are language models? adapting to the introduction of gender-neutral pronouns. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Deborah Cameron. 2012. *Verbal hygiene*. Routledge.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Anne Curzan. 2014. *Fixing English: Prescriptivism and language history*. Cambridge University Press.

Geraldine Damnati. 2024. From benchmark assessments to in-use evaluations: An even wider gap to bridge at the era of generative AI. Mila Workshop: NLP in the era of generative AI, cognitive sciences, and societal transformation.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff M Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. *In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Susan Ehrlich and Ruth King. 1992. Gender-based language reform and the social construction of meaning. *Discourse & Society*, 3(2):151–166.

Magdalena Formanowicz, Sylwia Bedynska, Aleksandra Cisłak, Friederike Braun, and Sabine Sczesny. 2013. Side effects of gender-fair language: How feminine job titles influence the evaluation of female applicants. *European Journal of Social Psychology*, 43(1):62–71.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028):147–154.

Tamanna Hossain, Sunipa Dev, and Sameer Singh. 2023. MISGENDERED: Limits of large language models in understanding pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5352–5367, Toronto, Canada. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Judith T Irvine. 1989. When talk isn't cheap: Language and political economy. *American Ethnologist*, 16(2):248–267.

Samantha Jackson, Barend Beekhuizen, Zhao Zhao, and Rhonda McEwen. 2024. GPT-4-Trinis: Assessing GPT-4's communicative competence in the English-speaking majority world. *AI & Society*, pages 1–17.

Kai Jacobsen, Charlie E Davis, Drew Burchell, Leo Rutherford, Nathan Lachowsky, Greta Bauer, and Ayden Scheim. 2024. Misgendering and the health and wellbeing of nonbinary people in Canada. *International Journal of Transgender Health*, 25(4):816–830.

Lee Jiang. 2023. Resistance to singular "they" in Reddit communities. Master's thesis, University of Toronto.

Rishi Kant. 2025. OpenAI's weekly active users surpass 400 million. *Reuters*.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM Collective Intelligence Conference*, pages 12–24.

Paul V Kroskrity. 2004. Language ideologies. *A Companion to Linguistic Anthropology*, 496:517.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. *Proceedings of the 29th International Conference on Computational Linguistics*.

Li Lucy, Suchin Gururangan, Luca Soldaini, Emma Strubell, David Bamman, Lauren Klein, and Jesse Dodge. 2024. AboutMe: Using self-descriptions in webpages to document the effects of English pretraining data filters. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7393–7420, Bangkok, Thailand. Association for Computational Linguistics.

Gunnar Lund, Kostiantyn Omelianchuk, and Igor Samokhin. 2023. Gender-inclusive grammatical error correction through augmentation. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 148–162, Toronto, Canada. Association for Computational Linguistics.

Alonzo Martinez. 2023. An employer's guide to inclusive language. *Forbes Magazine*.

Mistral AI Team. 2024. Mistral nemo.

Annabelle Mooney and Betsy Evans. 2015. *Language, society and power: An introduction*. Routledge.

Brittney O'Neill. 2021. He, (s)he/she, and they: Language ideologies and ideological conflict in gendered language reform. *Working papers in Applied Linguistics and Linguistics at York*, 1:16–28.

Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "I'm fully who I am": Towards centering transgender and non-binary voices to measure biases in open language generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1246–1266.

Brandon Papineau, Rob Podesva, and Judith Degen. 2022. 'Sally the congressperson': The role of individual ideology on the processing and production of english gender-neutral role nouns. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Joshua Raclaw. 2025. A trans linguistic perspective on multiple pronoun use in english. *Proceedings of the Linguistic Society of America*, 10(1):5927–5927.

Sabine Sczesny, Magda Formanowicz, and Franziska Moser. 2016. Can gender-fair language reduce gender stereotyping and discrimination? *Frontiers in psychology*, page 25.

Michael Silverstein. 1985. Language and the culture of gender: At the intersection of structure, usage, and ideology. In *Semiotic mediation*, pages 219–259. Elsevier.

Elizabeth Stokoe and Frederick Attenborough. 2014. Gender and categorial systematics. *Handbook of language, gender and sexuality*, pages 161–179.

Yolande Strengers, Lizhen Qu, Qiongkai Xu, and Jarrod Knibbe. 2020. Adhering, steering, and queering: Treatment of gender in natural language generation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. 2021. NeuTral rewriter: A rule-based and neural approach to automatic rewriting into gender-neutral alternatives. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Julia Watson, Barend Beekhuizen, and Suzanne Stevenson. 2023a. What social attitudes about gender does BERT encode? leveraging insights from psycholinguistics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6790–6809, Toronto, Canada. Association for Computational Linguistics.

Julia Watson, Sophia S. Lee, Barend Beekhuizen, and Suzanne Stevenson. 2025. Do language models practice what they preach? examining language ideologies about gendered language reform encoded in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1201–1223, Abu Dhabi, UAE. Association for Computational Linguistics.

Julia Watson, Sarah Walker, Suzanne Stevenson, and Barend Beekhuizen. 2023b. Communicative need shapes choices to use gendered vs. gender-neutral kinship terms across online communities. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45.

Langdon Winner. 1980. Do artifacts have politics? *Daedalus*, 109(1):121–136.

Lal Zimman. 2017. Transgender language reform: Some challenges and strategies for promoting trans-affirming, gender-inclusive language. *Journal of Language and Discrimination*, 1(1):84–105.

A Role noun sets

The full list of role noun sets we considered are:

Neutral	Feminine	Masculine
alderperson	alderwoman	alderman
anchor	anchorwoman	anchorman
assemblyperson	assemblywoman	assemblyman
ball person	ballgirl	ballboy
bartender	bargirl	barman
businessperson	businesswoman	businessman
camera operator	camerawoman	cameraman
caveperson	cavewoman	caveman
chairperson	chairwoman	chairman
clergyperson	clergywoman	clergyman
congressperson	congresswoman	congressman
councilperson	councilwoman	councilman
cow herder	cowgirl	cowboy
craftsperson	craftswoman	craftsman
crewmember	crewwoman	crewman
delivery person	delivery woman	delivery man
draftsperson	draftswoman	draftsman
emergency medical technician	ambulancewoman	ambulanceman
fan	fangirl	fanboy
farm worker	farmgirl	farmboy
fencer	swordswoman	swordsman
firefighter	firewoman	fireman
fisher	fisherwoman	fisherman
foreperson	forewoman	foreman
frontperson	frontwoman	frontman
gentleperson	gentlewoman	gentleman
handyperson	handywoman	handyman
layperson	laywoman	layman
maniac	madwoman	madman
meteorologist	weatherwoman	weatherman
newspaper delivery person	papergirl	paperboy
ombudsperson	ombudswoman	ombudsman
outdoorsperson	outdoorswoman	outdoorsman
point-person	point-woman	point-man
police officer	policewoman	policeman
postal carrier	postwoman	postman
repairperson	repairwoman	repairman
reporter	newswoman	newsman
salesperson	saleswoman	salesman
select board member	selectwoman	selectman
server	waitress	waiter
sharpshooter	markswoman	marksman
showperson	showwoman	showman
sound engineer	soundwoman	soundman
spokesperson	spokeswoman	spokesman
statesperson	stateswoman	statesman
stunt double	stuntwoman	stuntman
tradesperson	tradeswoman	tradesman
tribesperson	tribeswoman	tribesman
wingperson	wingwoman	wingman

These role noun sets are adapted from Watson et al. (2025), which drew from several sources (Vanmassenhove et al., 2021; Papineau et al., 2022; Bartl and Leavy, 2024; Lucy et al., 2024). Here, we only included role noun sets where we could obtain sentence usages in the AboutMe dataset (Lucy et al., 2024). Additionally, some filtering constraints in Watson et al. (2025) were not relevant to us. In particular, they excluded role noun sets where one variant was a substring of another. Here we include such cases (e.g., *fisher, fisherwoman, fisherman*).

B Segmenting responses into revisions and justifications

Here we describe our heuristic algorithm for extracting the revised sentences and justifications from model output, and present an evaluation of this algorithm's accuracy.

To identify the revised sentence, we first split model output into sentences using NLTK's sentence tokenizer (3.9.1). We take the revision to be the sentence that is most similar to the input sentence stimulus, based on METEOR scores (Banerjee and Lavie, 2005).[8] Because the input sentence may be split into multiple sentences during revision, we also consider sequences of up to 3 contiguous sentences as possible revisions. We exclude sentences that are identical to the input sentence, as sometimes model outputs repeat the input sentence before the revised version. We take the rest of the response following the revised sentence to be the justification.

In some cases, models proposed multiple possible revisions. We aimed to select the first proposed revision, by removing any text following the phrase "option 2" before running the algorithm described above. This kind of response was particularly common for the gemma model.

To evaluate the accuracy of our heuristic algorithm, we randomly sampled $n = 50$ responses per model, which were not considered in developing our heuristics. The algorithm achieves an average accuracy of 94% in exactly identifying revised sentences and 93% in exactly identifying justifications. See accuracy per model in Table 7.

C Alternative wording revisions

In Sections 4 and 5, we split model revisions into four types, based on what role noun variants were revised to: neutral, feminine, masculine, and "alternative wording." Because alternative wordings make up such a large share of revisions, we need to understand their make-up. Here we assess:

[8] We used METEOR scores because we wanted a simple, interpretable, and lightweight measure of sentence similarity. We also considered BLEU scores, and we found METEOR scores achieved a higher accuracy than BLEU at extracting justifications (assessed by comparing to manual annotation).

	Revision	Justification
gemma-2-9b-it	86	82
gpt-4o	94	94
llama-3.1-8B-Instr.	96	94
Mistral-Nemo-Instr.	100	100
Overall	94	93

Table 7: Accuracy of our heuristic algorithm. (percentage correctly identified)

starting variant	percentage gender-neutral
neutral	95
feminine	96
masculine	96
Overall	95.7

Table 8: Rates of gender-neutral alternative wordings, by starting variant

1. Are alternative wordings typically gender-neutral?

2. What are common sub-categories of alternative wordings?

Additionally, in Sec 5, we compare justifications across different starting role noun variants (i.e., whether the role noun in the input sentence was neutral, masculine, or feminine), when revising to alternative wordings (H1b). Because of this, it is also important to understand whether the qualities of alternative wordings vary across starting variants, which would inform our interpretation of results. So, for each of the questions above, we also assess whether we observe differences across starting variants (in rates of gender-neutral alternative wordings for 1, and in frequency of sub-categories of alternative wordings for 2).

C.1 Rates of gender-neutral alternative wordings

To assess rates of gender-neutral alternative wordings, we randomly sampled 75 responses per model, split evenly across starting variants, from the subset of responses considered in the justifications analysis (300 responses total). We manually annotated these responses to assess whether the revision was gender-neutral (i.e., did not introduce lexically gendered words). The vast majority of revisions were gender-neutral (95.7%), with similar rates across starting variants, as shown in Table 8.

Most of the gendered alternative wordings involved using a word that was morphologically related to the role noun (e.g., revising *craftsman* to instead talk about *craftsmanship*, or revising *gentleperson* to talk about someone's *gentlemanly nature*). Gender-neutral alternative wordings were quite varied, for example, replacing *ambulancewoman* with *paramedic*; replacing *fanboy* with *enthusiast*; replacing *newswoman* with *freelance journalist*; replacing *businessperson* with talking about *leading businesses*; and replacing *spokesperson* with talking about *advocating* for something. We go into greater depth about the make-up of alternative wordings in the next subsection.

The high rates of gender-neutral alternative wordings motivate treating this category as gender-neutral. Additionally, similar rates of gender-neutral alternative wordings across starting variants supports comparing their associated justifications to assess H1b.

C.2 Make-up of alternative wordings

In addition to understanding the rate of gender-neutral alternative wordings, we also wanted to get a general sense of their make-up. We used an inductive coding approach to develop a categorization scheme for alternative wordings, and identified the following generalizable categories:

1. **Alternative Noun Phrase**: The role noun is replaced by a noun phrase not present in our role noun set (e.g., *outdoorsperson* → *outdoor enthusiast*).

2. **Removed**: The role noun is entirely omitted without replacement.

3. **Mentions of Profession**: The role noun is substituted with a description explicitly referencing the field or profession (e.g., *firefighter* → *work in firefighting* or *businessperson* → *career in business*).

4. **Verb Phrase**: The intended meaning of the original role noun is conveyed through a verb phrase describing associated actions or responsibilities, rather than naming the role directly (e.g., revising to replace *outdoorsperson* with a phrase talking about *exploring the great outdoors*).

5. **Other**: Revisions that do not clearly fit into any of the categories above. Some examples include metaphorical uses of the role noun (e.g., *work like a madman* → *work tirelessly*) and substitutions with placeholders (e.g., *postwoman* → *[insert his profession here]*).

starting variant	alt. noun phrase	removed	mention of profession	verb phrase	other	N/A
neutral	44	9	10	7	24	6
feminine	59	10	10	6	12	3
masculine	58	12	9	6	11	4
Overall	53.7	10.3	9.7	6.3	15.7	4.3

Table 9: Sub-categories of gender-neutral alternative wordings, by starting variant. (Cells present percentages of alternative wordings that fall into each sub-category.)

6. **N/A**: Cases where the split algorithm from Appendix B did not correctly identify the revised sentence.

Two authors used this scheme to annotate the same sample from the previous subsection, and then discussed to resolve any disagreements.

The breakdown of alternative wording types by starting variant is shown in Table 9. For all starting variants, the most frequent alternative wording subcategory is alternative noun phrases. One difference across variants is that alternative noun phrases appear slightly more frequent for gendered starting variants, compared to neutral ones. However, in general, the frequency of the categories across starting variants has a similar distribution, motivating comparing their associated justifications in H1b.

D Word sets for justifications analysis

To study the presence of different themes in justifications, we required word sets corresponding to each theme. We started with manually curated seed sets. Half the words in each seed set were synonyms of the theme label word, and half were antonyms. For example, for the theme `inclusive`, the seed set was *{inculsive, exclusionary}*. Seed words for each theme are italicized in Table 5.

Then, we used contextual word embeddings from BERT (Devlin et al., 2019) to build expanded sets of 10 words per theme, based on these seed sets. We started by identifying sentences in the justifications that mention role nouns. We identified adjectives that occur in these sentences, using spaCy's part of speech tagger. We forced the inclusion of some frequent hyphenated adjectives that were split into multiple tokens by spaCy (*gender-neutral*, *gender-specific*, and *non-binary*), resulting in $N = 1,039$ total adjectives. We then generated contextual embeddings using BERT (specifically `bert-base-uncased`) for each adjective token. Since we were specifically interested in representing value-relevant properties of adjectives, rather than information about job roles, we replaced role noun variants with `[MASK]` tokens, to limit the influence of specific occupations on these representations. For adjectives that corresponded to multiple wordpiece tokens, we averaged the wordpiece contextual embeddings.

We then created word embeddings per adjective by averaging the contextual embeddings of all of that adjective's occurrences. Next, we generated theme embeddings by averaging the embeddings of the words in each seed set (e.g., averaging the embeddings of *inclusive* and *exclusionary* for the `inclusive` theme). Then, we combined seed sets with the 10 nearest neighbor adjectives for each theme embedding to get the full word sets per theme.

theme	starting role noun gender neutral	feminine	masculine
inclusive	6	29	29
modern	5	10	12
professional	12	9	11
standard	11	6	6
natural	8	4	5

Table 10: Themes in justifications by starting variant. Cells indicate the percentage of justifications where a theme was mentioned. Based on $13,609$ responses, where role nouns were revised to alternative wordings.

E Descriptive statistics about justification themes

Table 10 shows the breakdown of themes across starting role noun variants, and Table 11 shows the breakdown of themes across preambles.

theme	gender declaration			pronoun declaration			pronoun usage		
	nonbinary	woman	man	they/them	she/her	he/him	their	her	his
inclusive	58	48	38	56	65	54	41	44	43
modern	6	18	17	8	22	17	21	25	22
professional	4	20	17	4	13	9	23	19	20

Table 11: Themes in justifications for revisions to neutral. Cells indicate the percentage of justifications where a theme was mentioned. Based on $6,964$ model responses, where role nouns were revised from gendered to neutral.