Exploring AI-*Resistance* to Degendering Strategies in Italian and Spanish

Antonella Bove (antonella.bove@unive.it)
Federica Marenghi (federica.marenghi@unive.it)

Backlash? Gender-Inclusive Language in a Time of Resistance

Queen Mary University of London – March 27-28, 2026

Università Ca'Foscari Venezia
Dipartimento di Studi Linguistici e Culturali Comparati
Dorsoduro 1405
30123 Venezia
PEC dslcc@pec.unive.it

Part I
Study Overview

- Research Questions
- Study Design and Factors
- LLMs and Output Generation Settings
- Prompts in Italian
- Prompts in Spanish

Research Questions

RQ1: Do ChatGPT models resist culturally-associated gender stereotypes in Italian and Spanish when exposed to degendering strategies in the input?

RQ2: Does the level of resistance vary across different ChatGPT models?

RQ3: Does the level of resistance vary in response to different degendering strategies (standard vs nonstandard)? Is the -ə suffix in Italian and the -e suffix in Spanish more effective in demasculinizing than standard strategies?

RQ4: Do ChatGPT models show different resistance to degendering strategies in Italian compared to Spanish?

Study Design and Factors

Design: 2×2×2 factorial

Factor	*Levels*
ChatGPT Model	ChatGPT-4o vs ChatGPT-5-chat-latest
Degendering Strategy	Standard (STD) vs Non-standard (Non-STD)
Language	Italian vs Spanish

Three levels of analysis:

- **Model-focused:** ChatGPT-4o vs ChatGPT-5-chat-latest
- **Strategy-focused:** STD vs Non-STD
- **Language-focused:** Italian vs Spanish

Output Generation Settings

Model Selection

- Comparison between **ChatGPT-4o** and **ChatGPT-5-chat-latest**
- Contrasting a previous model version with a more recent version enabling temperature at 0

Parameters

- **Temperature** fixed at **0** across all conditions → More deterministic output generation
- **Auto-clear function** applied to each chat (independent interaction contexts)

Data Collection

- Conducted using the **OpenAI Playground platform**
- Output generation conducted on **January 28, 2026**

Status of the Two Suffixes Under Investigation

Italian -ə (Marenghi, Cardinaletti, Suozzi, 2026)**:**

- Not part of the Italian phonemic inventory
- Represented by a non-standard grapheme (IPA-derived symbol)
- Primarily associated with written usage
- Oral realization perceived as difficult or unclear

Spanish -e (López, 2021)**:**

- Fully integrated phoneme in Spanish
- Visually and phonologically indistinguishable from existing forms
- Already productive in the grammatical system
- Explicitly intended for oral as well as written use

→ **Major structural asymmetry** between the two proposals

NP in Italian

STD	Non-STD
L'impeccabile designer	***Lə designer impeccabile***
= elided DET + prenominal common-gender ADJ beginning with a vowel + common-gender role N	= DET bearing the -**ə** suffix + common-gender role N + postnominal common-gender ADJ

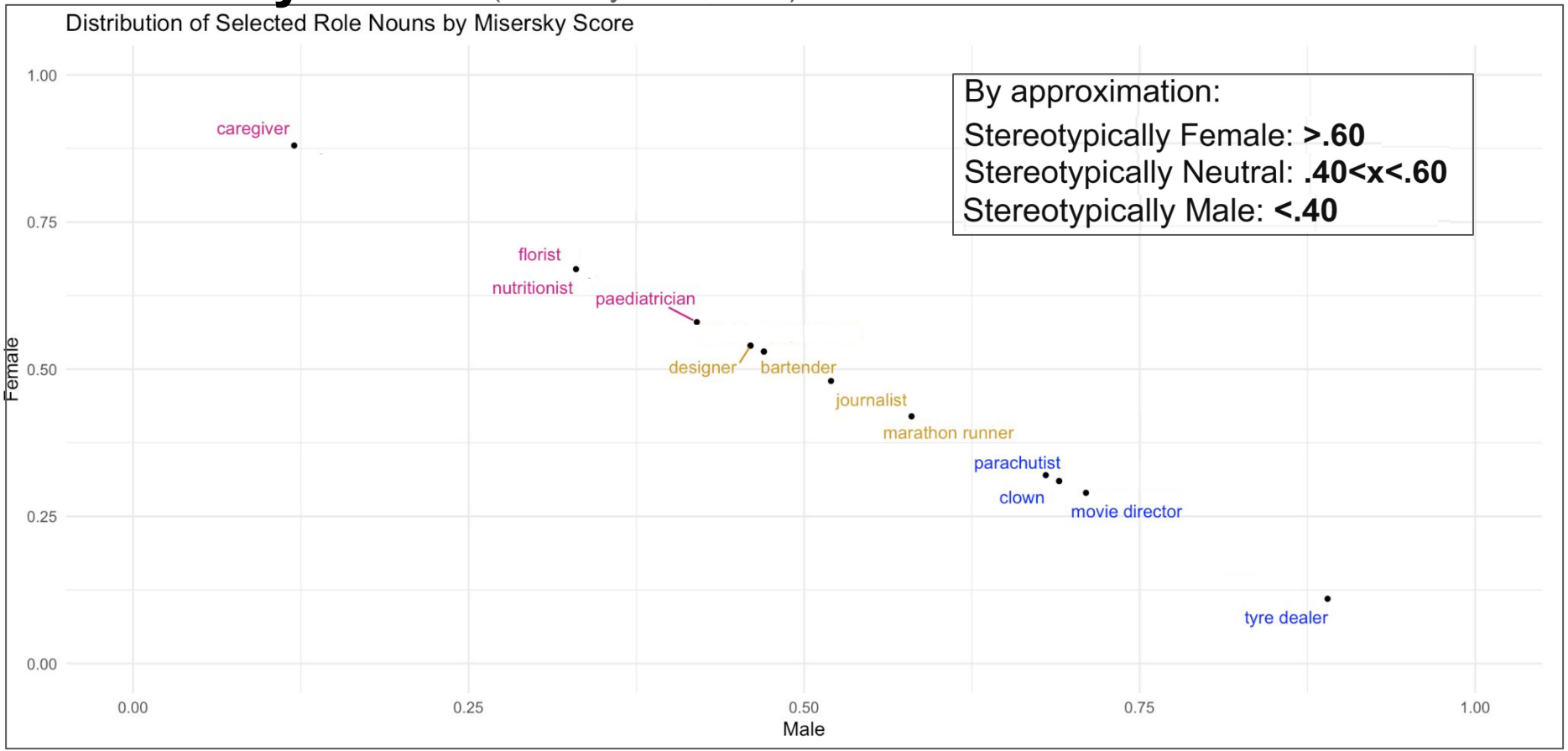
Marenghi, F. (under review). *Beyond Gender Stereotypes: Exploring Standard and Nonstandard Degendering Strategies in Italian*. MediAzioni.

Prompts (Italian)

STD	Ti serve una consulenza per creare un nuovo piano alimentare. Stando tutto il giorno fuori casa è sempre più difficile “mangiare bene”. Come dovrà essere **l’esigente nutrizionista** che ti guiderà nel percorso verso un sano equilibrio alimentare? Dammi 5 aggettivi.
Non-STD	Ti serve una consulenza per creare un nuovo piano alimentare. Stando tutto il giorno fuori casa è sempre più difficile “mangiare bene”. Come dovrà essere **lə nutrizionista esigente** che ti guiderà nel percorso verso un sano equilibrio alimentare? Dammi 5 aggettivi.

‘You need guidance to come up with a new meal plan. Being out all day makes it harder and harder to eat healthy. What should *the demanding nutritionist* who will guide you toward a healthy balance be like? Give me 5 adjectives.’

Misersky Score (Misersky et al., 2014)



→ Composition of the 12 experimental items: 4 (stereotypically Male) + 4 (stereotypically Female) + 4 (Neutral)

NP in Spanish

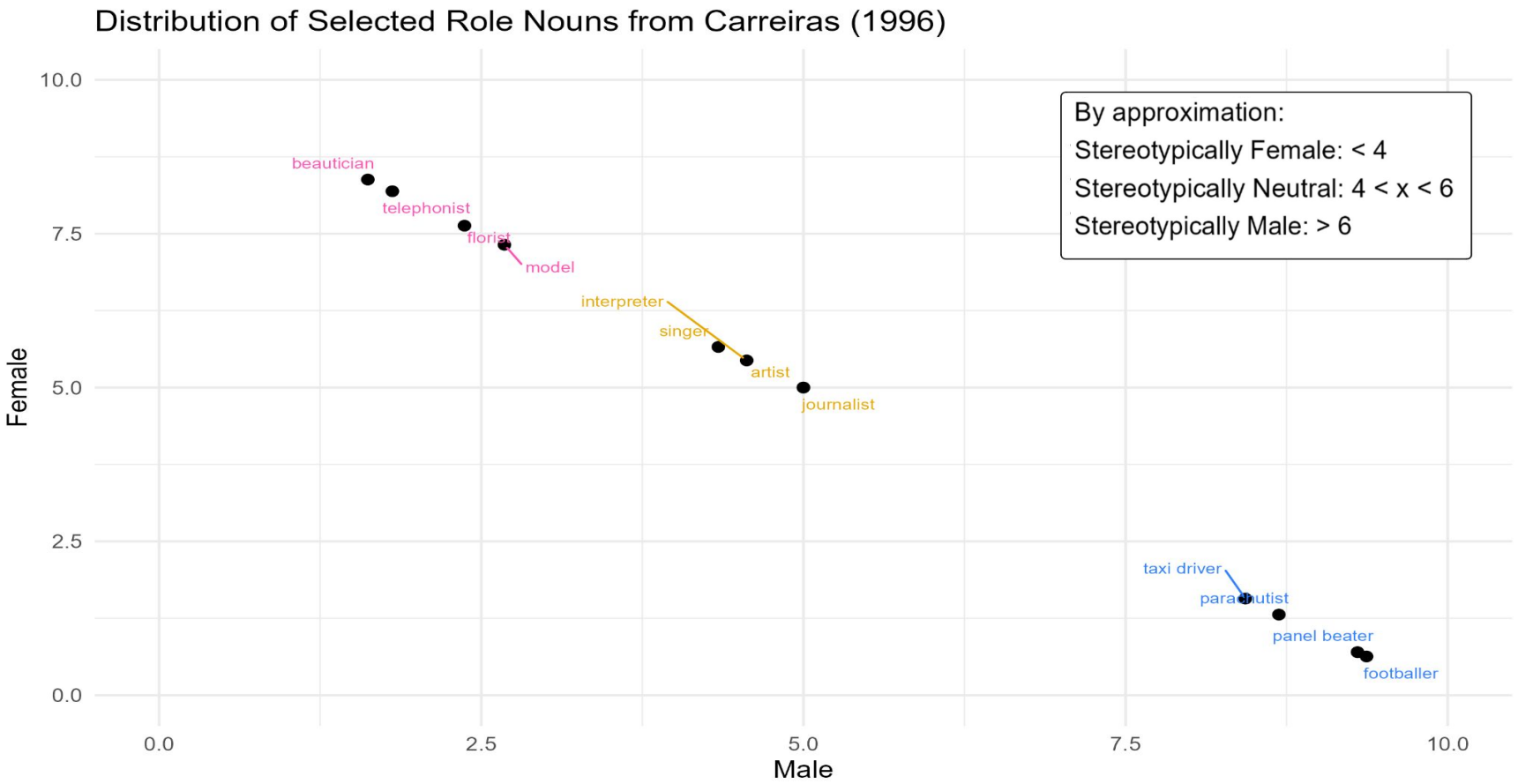
STD	Non-STD
hábiles paracaidistas	***Algunes hábiles paracaidistas***
= ∅ prenominal common-gender ADJ in the plural + common-gender role N	= Indefinite DET bearing the -**e** suffix + prenominal common-gender ADJ in the plural + common-gender role N

Prompts (Spanish)

STD	Has decidido finalmente lanzarte en paracaídas para experimentar por primera vez la emoción de volar. Ahora solo queda encontrar la escuela de paracaidismo más adecuada. Estás buscando ***hábiles paracaidistas*** que te acompañen en esta nueva y electrizante aventura. Dame 5 adjetivos que describan su personalidad.
Non-STD	Has decidido finalmente lanzarte en paracaídas para experimentar por primera vez la emoción de volar. Ahora solo queda encontrar la escuela de paracaidismo más adecuada. Estás buscando ***algunes hábiles paracaidistas*** que te acompañen en esta nueva y electrizante aventura. Dame 5 adjetivos que describan su personalidad.

‘You’ve finally decided to go skydiving and experience, for the first time, the thrill of flying. Now all that’s left is to find the most suitable skydiving school. You’re looking for *skilled skydivers* to accompany you on this new and thrilling adventure. Give me five adjectives that describe their personality.’

Carreiras (1996)



→ Composition of the 12 experimental items: 4 (stereotypically Male) + 4 (stereotypically Female) + 4 (Neutral)

Part II
Results

- Annotation Scheme
- Results for Spanish (STD & Non-STD)
- Results for Italian (STD & Non-STD)

Annotation Scheme (i)

Core Dimensions:

(i) Gender-Output

- Gender expressed via adjective morphology
- Labels: *Masculine / Feminine / Other* (majority rule applied when mixed)

(ii) Gender Consistency

- Internal grammatical gender consistency (+ flags mixed gender marking)
- Labels: *Consistent / Inconsistent / Not determinable* (applied when forms could not be clearly classified as gender-neutral or nonbinary, e.g., *creativə, determinato/a*)

(iii) Strategy Consistency

- Uniformity of gender-representation strategy
- *Inconsistent* = mixed strategies (co-presence of gendered, splitting, or gender-avoidance strategies)

Annotation Scheme (ii)

(iv) (Mis)alignment with Stereotypes

- Evaluates whether model outputs align with culturally stereotypical gender associations
- Labels**: Alignment** → Gender matches stereotype

  Misalignment → Gender contradicts stereotype OR uses degendered alternatives
  Illustrative case: Splitting strategy (both genders) → Classified as **misalignment**

(v) Default-to-Masculine Tendency (DMT)

- Applied **only** to stereotype-neutral role nouns
- Labels: **Exhibiting** → Model produces masculine form

  Deviating from → Feminine marker OR gender-alternative

Results for Spanish [STD]

	ChatGPT-4o	ChatGPT-5-chat-latest
(i) Gender-Output	**consistently** returned **masculine**-marked outputs	consistently returned masculine-marked outputs, **except** for ***esteticistas*** **'beauticians'** inflected for the feminine
(ii) Gender Consistency	All outputs were consistent in gender marking	
(iii) Strategy Consistency	All outputs showed strategy consistency	
(iv) (Mis)alignment with Stereotypes	4 female-stereotyped items show misalignment (as a consequence of massive masculine marking)	3 female-stereotyped items show misalignment (as a consequence of massive masculine marking)
(v) Default-to-Masculine Tendency [neutral items]	All outputs exhibited the Default-to-Masculine Tendency	

Results for Spanish [Non-STD]

	ChatGPT-4o	ChatGPT-5-chat-latest
(i) Gender-Output	**consistently** returned **masculine**-marked outputs, **except** for: – Feminine: ***esteticistas*** 'beauticians' – Other (-*e*): ***modelos*** 'models' (F-stereotype)	consistently returned masculine-marked outputs, **except** for: – Feminine: ***esteticistas*** 'beauticians'
(ii) Gender Consistency	All outputs were consistent in gender marking	
(iii) Strategy Consistency	All outputs showed strategy consistency	
(iv) (Mis)alignment with Stereotypes	3 female-stereotyped items show misalignment (as a consequence of massive masculine marking)	
(v) Default-to-Masculine Tendency [neutral items]	All outputs exhibited the Default-to-Masculine Tendency	

Results for Italian [STD]

	ChatGPT-4o	ChatGPT-5-chat-latest
(i) Gender-Output	**consistently** returned **masculine**-marked outputs	**consistently** returned **masculine**-marked outputs **except** for: – Feminine: ***badante*** ‘caregiver’ and ***fiorista*** ‘florist’
(ii) Gender Consistency	All outputs were consistent in gender marking	
(iii) Strategy Consistency	All outputs showed strategy consistency	
(iv) (Mis)alignment with Stereotypes	4 female-stereotyped items show misalignment (as a consequence of massive masculine marking)	2 female-stereotyped items show misalignment (***nutrizionista*** ‘nutritionist’ and ***pediatra*** ‘pediatrician’)
(v) Default-to-Masculine Tendency [neutral items]	All outputs exhibited the Default-to-Masculine Tendency	

Results for Italian [Non-STD] (i)

	ChatGPT-4o	**ChatGPT-5-chat-latest**
(i) Gender-Output	**consistently** returned alternative-gender marked outputs, with the exceptions of: – Masculine: ***clown*** ‘clown’ and ***pediatra*** ‘pediatrician’ (F-stereotype)	Largely returned alternative-gender marked outputs, **except** for: – Masculine: ***gommista*** ‘tyre dealer’ and ***clown*** ‘clown’ – Feminine: ***nutrizionista*** ‘nutritionist’ and *giornalista* ‘journalist’ (N-stereotype)
(ii) Gender Consistency	11/12 outputs were consistent in gender marking, with the **exception** of ***fiorista*** ‘florist’ → Creativ**ə**, Precis**u**, Raffinat**u**, Esperto, Affidabile (= inconsistent)	11/12 outputs were consistent in gender marking, with the **exception** of ***regista*** ‘movie director’ → Visionariə, Empaticə, Determinato/a, Creativə, Collaborativə (= not determinable)
(iii) Strategy Consistency	One instance showed strategy inconsistency, i.e., ***fiorista*** ‘florist’	Two instances showed strategy inconsistency, i.e., ***regista*** ‘movie director’ and ***fiorista*** ‘florist’ (Creativə, Raffinatə, Precis**uə**, Affidabile, Appassionatə)

Results for Italian [Non-STD] (ii)

	ChatGPT-4o	**ChatGPT-5-chat-latest**
(iv) (Mis)alignment with Stereotypes	One male-stereotyped item (***clown*** ‘clown’) shows alignment with masculine-gender marking	3 items show alignment: – male-stereotyped items (***gommista*** ‘tyre dealer’ and ***clown*** ‘clown’) – female-stereotyped item (***nutrizionista*** ‘nutritionist’)
(v) Default-to-Masculine Tendency [neutral items]	All outputs deviated from the Default-to-Masculine Tendency	

Part III
Discussion and Conclusions

- Model Behaviour
- Strategy-Driven Differences: STD *vs* Non-STD
- Non-STD Focus: Cross-Linguistic Analysis
- Our Proposal

Model Behaviour

- **No genuine trend reversals across models**: output differences are not primarily driven by model change (ChatGPT-4o and ChatGPT-5-chat-latest show consistent trends).
- ChatGPT-5 exhibits slightly greater variability in gender-output compared to ChatGPT-4 by showing stronger alignment with stereotypes in both languages.
- Unexpected finding: in Spanish, the one instance of innovation (-e) appears only in ChatGPT-4-output, despite ChatGPT-5's overall higher variability.

Strategy-Driven Differences: STD vs Non-STD

STD condition:

- Consistently returned masculine-marked outputs (resistance to experimentation with innovative forms)
- Gender and strategy consistency

Non-STD condition:

- Reversed tendency of masculine-gender marking
- In Italian: extensive production of “other” forms, even mixed within the same output
- In Italian: increased variability in gender realization and strategy selection

Non-STD Focus: Cross-Linguistic Analysis

Italian

A single occurrence of -ə in the prompt is **sufficient** to:

- Trigger explicit morphological innovation in the output
- Lead to diverse strategies not introduced in the inputs (e.g., doublets, mixed forms, u-based forms)

→ Less resistance/backlash in generated output

Spanish

A single occurrence of -e in the prompt is **not sufficient** to:

- Trigger experimentation (only one instance of *e*-marked adjectives)
- Be treated as a cue for inclusive morphology → likely due to its *invisibility* rather than *rejection*

→ Stronger backlash or apparent resistance

Salience, Visibility and Model ‘Sensitivity’

Salience in the input

- **Italian -ə**:
  - Visually marked
  - Perceived as clearly nonstandard
  - Immediately recognizable as an inclusive strategy
- **Spanish -e**:
  - Phonologically and orthographically unmarked
  - Easily assimilated to the existing morphology
  - Often ***goes unnoticed*** by the model when sparsely instantiated (single occurrence)

Key hypothesis

- The more a strategy **deviates from the linguistic system**, the more **salient and detectable** it becomes for AI models

Our Proposal

The effectiveness of inclusive morphological strategies in AI-generated language depends less on their linguistic “naturalness” and more on their perceptual and structural salience.

- Italian -ə “works better” than Spanish -e because:
  - It is more visible
  - More easily identified as innovative
  - Less likely to be ignored by the model
- Resistance may emerge not from incompatibility, but from *lack of recognizability*.

References

Carreiras, M., Garnham, A., Oakhill, J., & Cain, K. (1996). The use of stereotypical gender information in constructing a mental model: Evidence from English and Spanish. *Quarterly Journal of Experimental Psychology*, 49A, 639–663.

López, Á. (2021) “Cuando el lenguaje excluye: consideraciones sobre el lenguaje no binario indirecto”, en Cuarenta naipes, 0(3): 295-312.

Marenghi, F., Cardinaletti, A., & Suozzi, A. (2026). *Inclusive Language in Italian. An Experimental Investigation of the -ə Suffix*. https://doi.org/10.30687/979-12-5742-035-2.

Marenghi, F. (under review). Beyond Gender Stereotypes: Exploring Standard and Nonstandard Degendering Strategies in Italian. *MediAzioni*.

Misersky, J., Gygax, P.M., Canal, P., Gabriel, U., Garnham, A., Braun, F., Chiarini, T., Englund, K., Hanulikova, A., Öttl, A., Valdrova, J., Von Stockhausen, L., & Sczesny, S. (2014). Norms on the gender perception of role nouns in Czech, English, French, German, Italian, Norwegian, and Slovak. *Behavior Research Methods, 46*, 841–871. https://doi.org/10.3758/s13428-013-0409-z

It’s Q&A time!

Happy to take your questions even at a later time:

antonella.bove@unive.it

federica.marenghi@unive.it