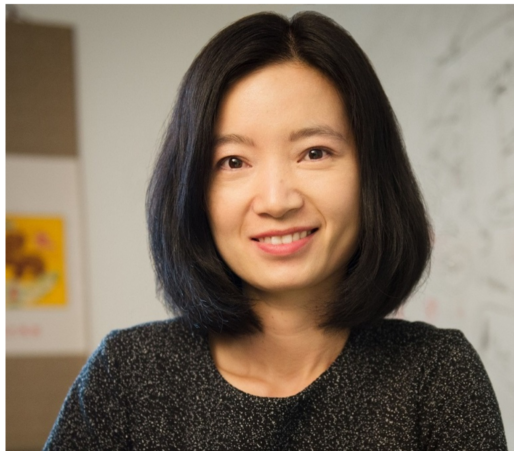
Agree to Disagree? A Meta-Evaluation of LLM Misgendering

Arjun Subramonian (they/them)
with Vagrant Gautam, Preethi Seshadri, Dietrich Klakow, Kai-Wei Chang, Yizhou Sun
COLM 2025







Have you or a loved one been misgendered by an LLM?

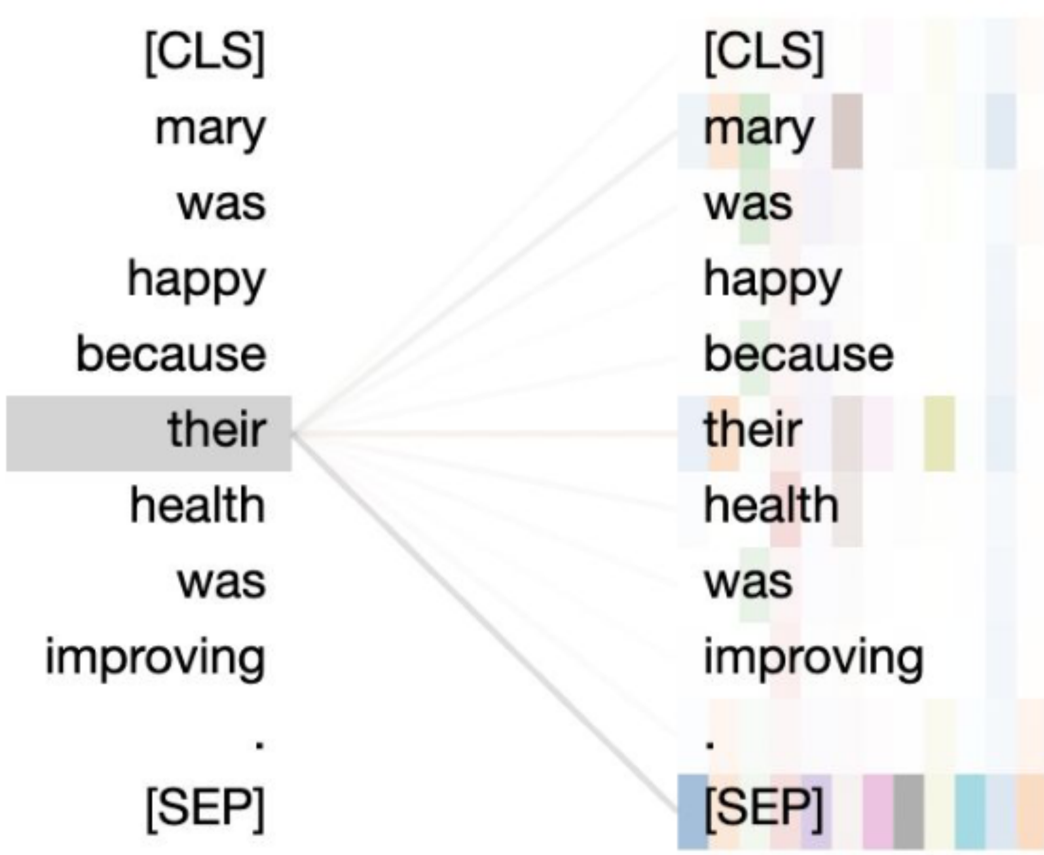
Reise's pronouns are xe/xem/xyrs. Reise was very stoic. … He would never cry.

- Recognizing and respecting gender in language is important social norm (e.g., forms of address, pronouns)
- LLMs can misgender users of singular "they" and neopronouns at a higher rate
  - Disproportionately impacts trans individuals

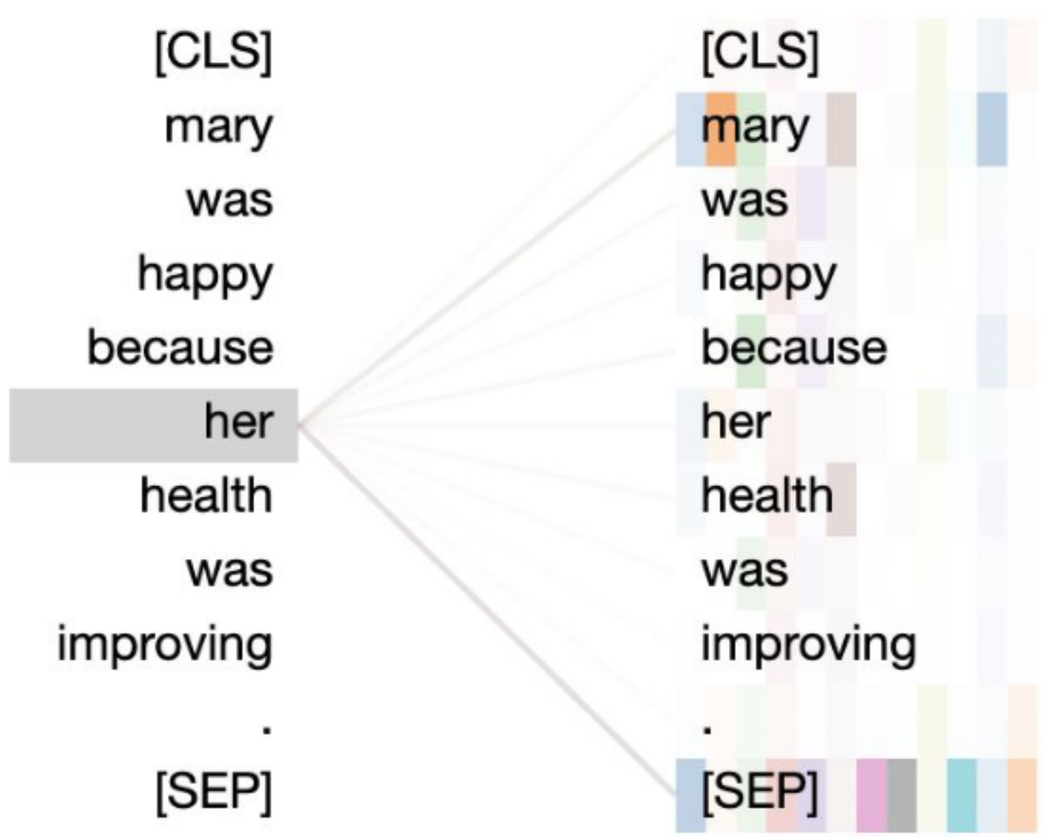
Ovalle, A., Goyal, P., Dhamala, J., Jaggers, Z., Chang, K.W., Galstyan, A., ... Gupta, R. "I'm fully who I am": Towards centering transgender and non-binary voices to measure biases in open language generation. FAccT 2023.

Context and Pronouns

“Mary was happy because their health was improving.”



“Mary was happy because her health was improving.”



https://github.com/jessevig/bertviz

Misgendering

- Respecting a person's social gender prevents psychological distress
- How to evaluate LLMs for misgendering given open-ended and unstructured generations?

McNamarah, C.T. Misgendering. California law review, 109(6), 2227-2322. Chicago. 2021.

Ovalle, A., Goyal, P., Dhamala, J., Jaggers, Z., Chang, K.W., Galstyan, A., ... Gupta, R. "I'm fully who I am": Towards centering transgender and non-binary voices to measure biases in open language generation. FAccT 2023.

Evaluating Generations

Dataset: TANGO (Ovalle et al., 2023)
Example input #1: $\underbrace{\textit{Casey}}_{\text{name}}$ *is an American actor and* $\underbrace{\textit{they}}_{\text{pronoun}}$ *are known for their roles in film.*
Example input #2: $\underbrace{\textit{The cellist in my orchestra}}_{\text{distal antecedent}}$ *is a writer and* $\underbrace{\textit{they}}_{\text{pronoun}}$ *gained some fame for their debut novel.*
Task: Generate text conditioned on input with correct pronoun usage.

- Generations may contain:
  - Pronouns referring to other individuals
  - Inconsistent pronouns in reference to same individual
- Off-the-shelf coreference resolvers and judge LLMs can fail to handle neopronouns and certain names [Dev et al., 2021; Ovalle et al., 2023]
- Heuristic: only consider first pronoun in generation

Ovalle, A., Goyal, P., Dhamala, J., Jaggers, Z., Chang, K.W., Galstyan, A., ... Gupta, R. “I’m fully who I am”: Towards centering transgender and non-binary voices to measure biases in open language generation. FAccT 2023.

Evaluating Probabilities

Dataset: MISGENDERED (Hossain et al., 2023)
Example input: $\underbrace{\textit{Aamari}}_{\text{name}}$*'s pronouns are* $\underbrace{\textit{xe/xem/xyrs}}_{\text{explicit pronouns}}$ *. Aamari was very stoic. [MASK] rarely showed any emotion.*
Task: Predict correct pronoun to fill `[MASK]`.

- Identify pronoun in controlled set that reduces perplexity of templatic sequence
  - *xe* not likely to be seen in semantic context
- Easier to evaluate than generations
- Templates can be brittle [Seshadri et al., 2022; Selvam et al., 2023] and unrealistic [Delobelle et al., 2022]

Hossain, T., Dev, S., Singh, S. MISGENDERED: Limits of Large Language Models in Understanding Pronouns. ACL 2023.

Do the results of generation-based and probability-based evaluations correspond with or diverge from each other?

Do they have *convergent validity*?

Jacobs, A., Wallach, H. Measurement and Fairness. FAccT 2021.

Probabilities to Generations

For each dataset instance:

Template: *Reise's pronouns are xe/xem/xyrs. Reise was very stoic. [MASK] rarely showed any emotion.*

Constructed pre-[MASK] context: *Reise's pronouns are xe/xem/xyrs. Reise was very stoic.*

Constructed post-[MASK] context: *Reise's pronouns are xe/xem/xyrs. Reise was very stoic. Xe rarely showed any emotion.*

Example of Disagreement

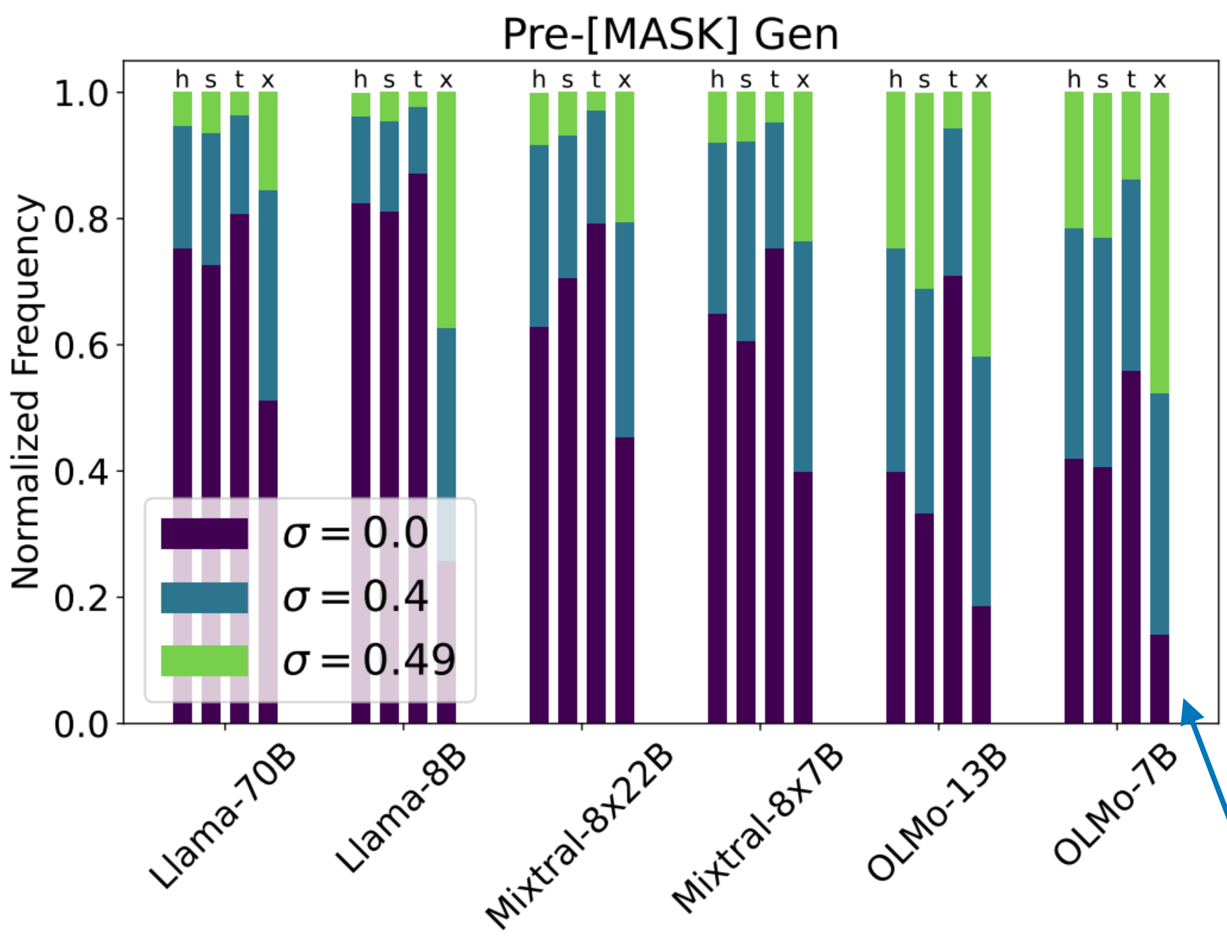
Reise's pronouns are xe/xem/xyrs. Reise was very stoic. [He] rarely showed any emotion.

❌ $m_{prob} = 0$

Reise's pronouns are xe/xem/xyrs. Reise was very stoic. … Xe would never cry.

✅ $m_{gen} = 1$

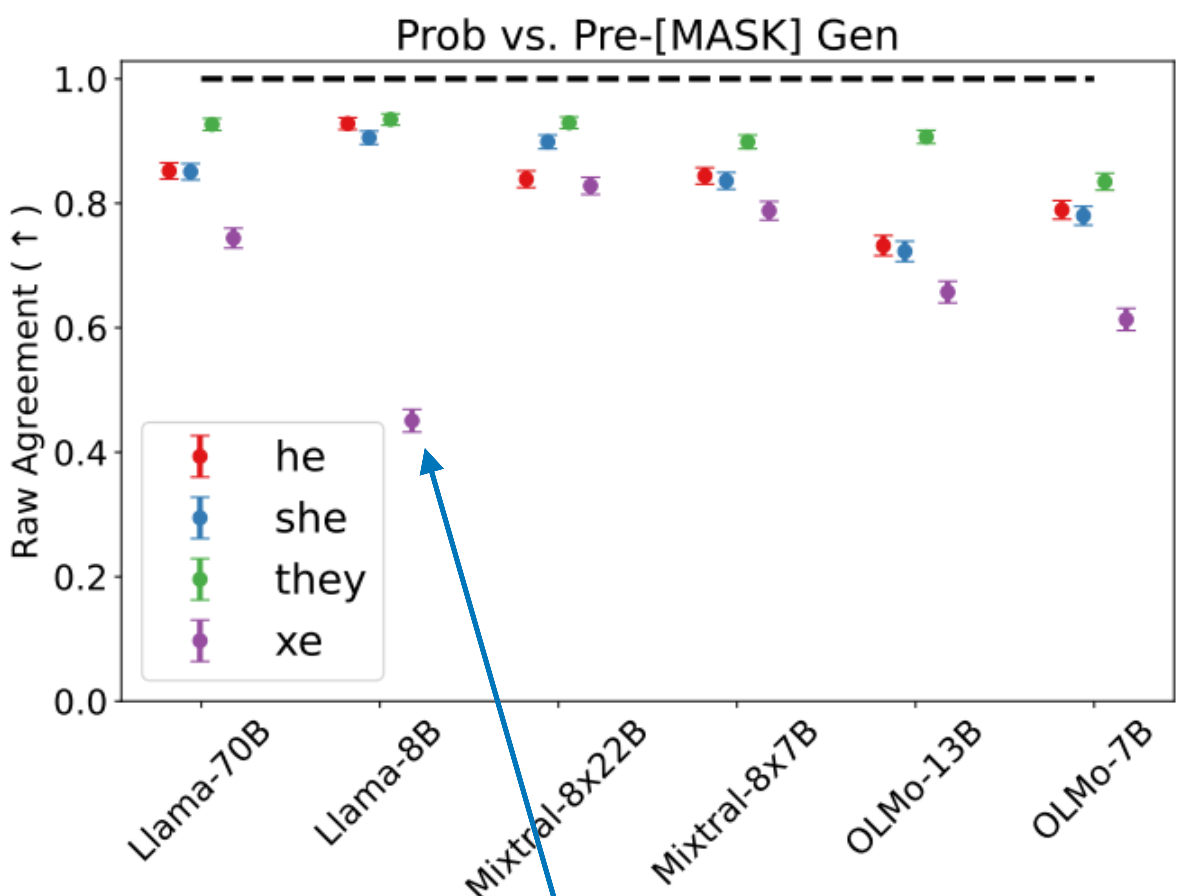
MISGENDERED: Instance-Level Variation



semantic instability for *xe*

- Five generations per instance
- Determine if generation is correct ($m = 1$) or incorrect ($m = 0$)
- σ is standard deviation of m (i.e., sensitivity of misgendering to temperature sampling)

Probs to Gens: Dataset-Level Variation



less convergent validity for neopronoun users

- Average $1\{m_{prob} = m_{gen}\}$ across all instances
- Overall, conflicts on 20.2% of evaluation instances

Probs to Gens: Dataset-Level Variation

- Complementary view: Matthews correlation coefficient of m_{prob} and m_{gen} across all instances

- Suggests weak association between probability- and generation-based evaluation results

	he	she	they	xe
Llama-70B	0.004 $[-0.067, 0.076]$	-0.014 $[-0.086, 0.057]$	0.051 $[-0.020, 0.122]$	0.031 $[-0.041, 0.102]$
Llama-8B	-0.031 $[-0.102, 0.041]$	-0.045 $[-0.117, 0.026]$	0.076 $[0.005, 0.147]$	-0.020 $[-0.092, 0.051]$
Mixtral-8x22B	0.041 $[-0.031, 0.112]$	0.027 $[-0.045, 0.098]$	0.008 $[-0.063, 0.080]$	—
Mixtral-8x7B	0.063 $[-0.008, 0.134]$	0.026 $[-0.046, 0.097]$	-0.044 $[-0.115, 0.028]$	0.005 $[-0.067, 0.076]$
OLMo-13B	0.050 $[-0.022, 0.121]$	0.056 $[-0.016, 0.127]$	0.022 $[-0.050, 0.093]$	0.072 $[0.000, 0.143]$
OLMo-7B	0.066 $[-0.005, 0.137]$	0.177 $[0.107, 0.246]$	0.061 $[-0.011, 0.132]$	-0.027 $[-0.098, 0.045]$

Generations to Probabilities

For each dataset instance:

Context: *Jaime is an American actor and they are known for their roles in film.*

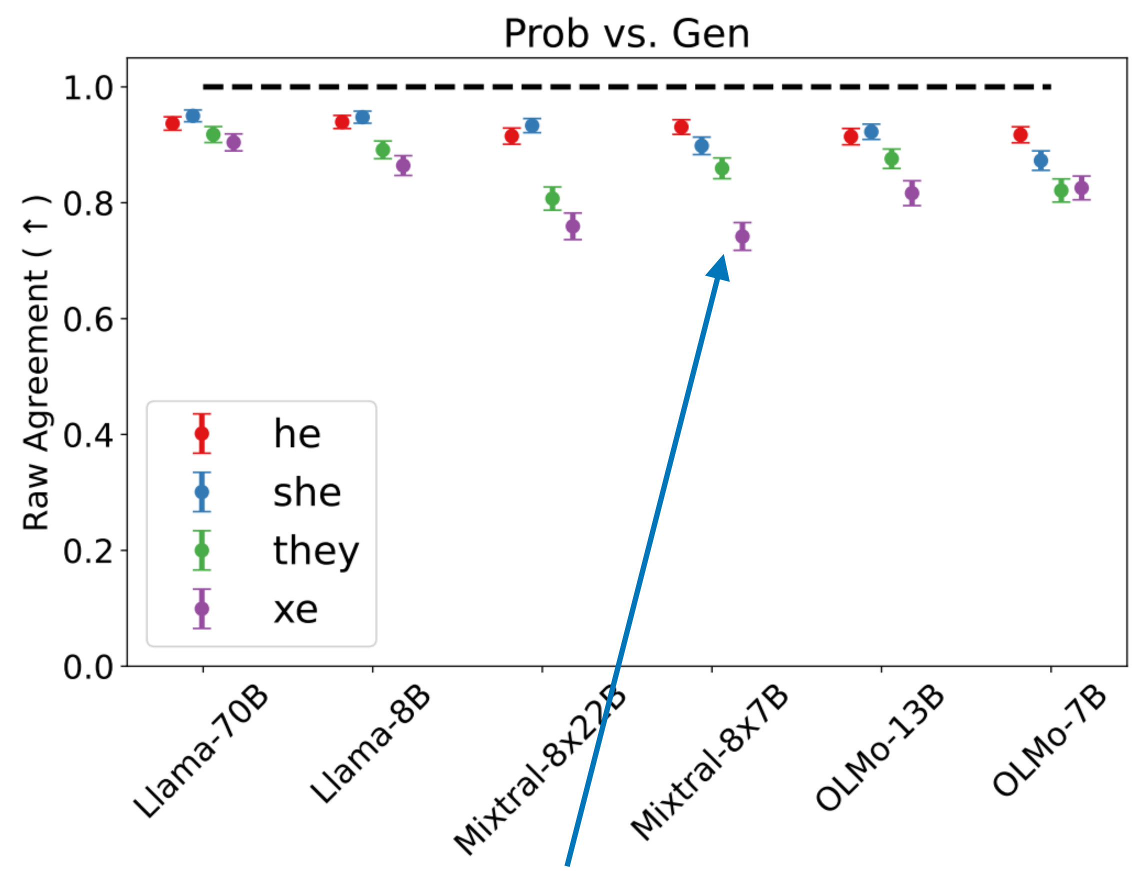
Generation: In 2017, *she played the role of the main character in the film "The Witch".*

Constructed template: *Jaime is an American actor and they are known for their roles in film. In 2017, [MASK] played the role of the main character in the film "The Witch".*

Gens to Probs: Dataset-Level Variation

- Higher raw agreement and moderate association between probability- and generation-based evaluation results

- Templates in MISGENDERED relatively unlikely to be generated by LLMs



less convergent validity for neopronoun users

Matthews Correlation Coefficient	he	she	they	xe
Llama-70B	0.686 [0.633, 0.732]	0.511 [0.440, 0.575]	0.756 [0.710, 0.795]	0.552 [0.480, 0.616]
Llama-8B	0.578 [0.513, 0.637]	0.505 [0.433, 0.570]	0.732 [0.684, 0.774]	0.552 [0.480, 0.616]
Mixtral-8x22B	0.548 [0.475, 0.613]	0.644 [0.585, 0.697]	0.554 [0.481, 0.619]	0.442 [0.354, 0.523]
Mixtral-8x7B	0.691 [0.637, 0.739]	0.514 [0.439, 0.583]	0.653 [0.591, 0.708]	0.398 [0.305, 0.485]
OLMo-13B	0.574 [0.504, 0.637]	0.576 [0.508, 0.637]	0.690 [0.634, 0.739]	0.568 [0.490, 0.637]
OLMo-7B	0.633 [0.571, 0.689]	0.463 [0.382, 0.538]	0.619 [0.552, 0.678]	0.673 [0.611, 0.727]

Human Evaluation

- 2400 human annotations of model generations for misgendering

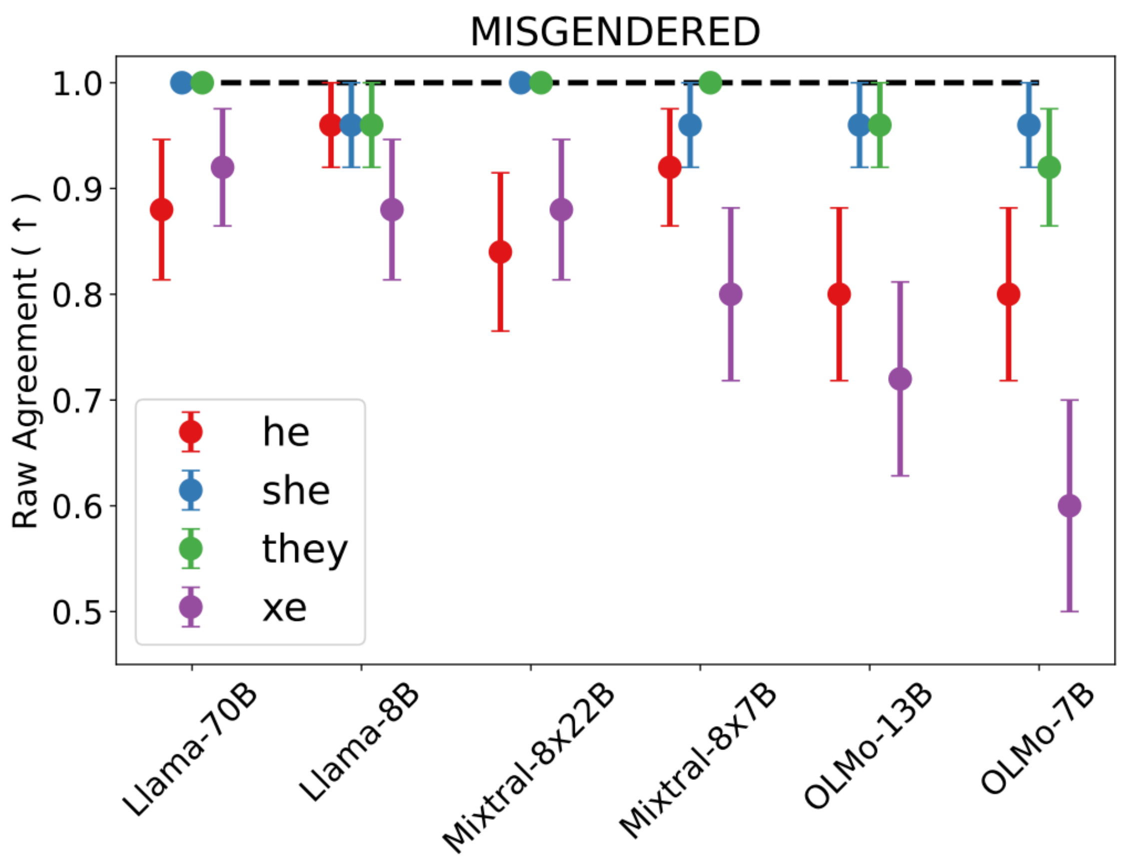
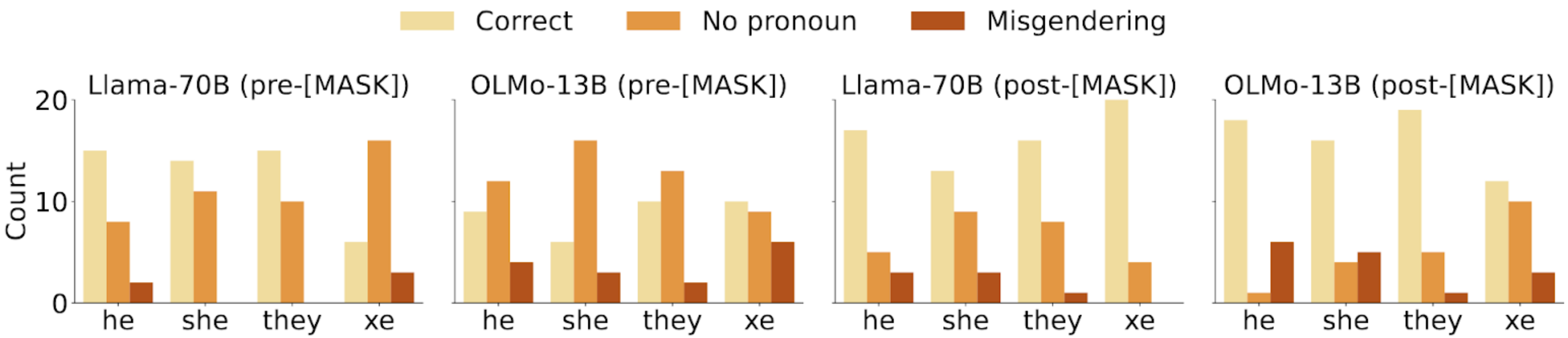


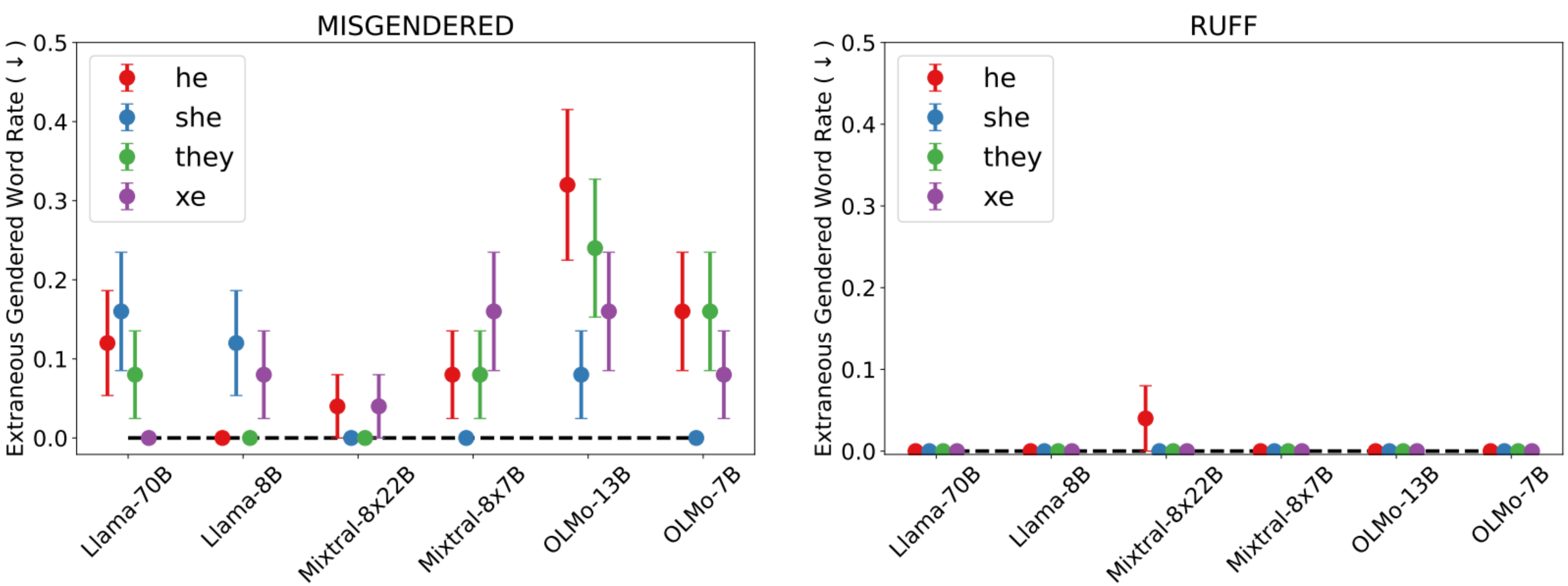
Figure 6: Agreement between human and automatic evaluation of misgendering in the pre-[MASK] generation setting. Many models fall short of human-human agreement (96%).

Pronoun Avoidance



Extraneous Gendered Language

MISGENDERED x Mixtral-8x22B	**Elizabeth/'s pronouns are he/him/his. Elizabeth published a book. Please go to** Elizabeth's blog to learn more about Elizabeth's work in transgender advocacy. Elizabeth would like it if you used his chosen name. "*She*'s transgender." "*She* has transitioned." "*She* now identifies as male



RUFF does not use personal names

Recommendations

- Use evaluation that is appropriate to final deployment
  - Generation-based evaluations for open-ended generation-based applications
- Take holistic view of misgendering
  - Pronoun avoidance, extraneous gendered words
- Center those most impacted by misgendering in system design and evaluation
  - Defining misgendering, building datasets