

Whose Toxicity? Persona Conditioning and Model–Annotator Label Agreement by Gender in Indonesian Toxicity Classification

Salsabila Zahirah Pranida^{*1,4} Antonia Karamolegkou^{*2,4}

Andrea Morales-Garzón³ Daniel Hershcovich⁴

¹MBZUAI ²INRIA

³Universidad de Granada ⁴University of Copenhagen

Abstract

Toxicity judgments can depend on who is judging. We study whether persona prompting captures this variation by evaluating ten Indonesian-capable LLMs under four persona conditions on 7,524 IndoDiscourse posts, each labeled by at least one female and one male annotator, comprising 24,655 individual judgments. Persona cues substantially affect toxicity predictions, but their effects are model-specific, show no common direction, and are small between the two gender personas. Yet this steerability does not amount to perspective adoption: switching between gender personas does not selectively increase label agreement with the corresponding annotator group. Annotation-level mixed-effects models and annotator-level permutation tests likewise find no reliable female–male agreement gap. The apparent gap obtained from aggregated group labels instead follows from annotator label rates and aggregation choices.

1 Introduction

Toxicity judgments depend on perspective: annotators’ identities, beliefs, and lived experiences shape what they consider toxic (Sap et al., 2022; Kumar et al., 2021; Al Kuwatly et al., 2020). Accordingly, recent work treats annotator disagreement not merely as noise, but as evidence of legitimately different viewpoints (Basile et al., 2021; Cabitza et al., 2023; Plank, 2022; Uma et al., 2021b). This perspective has also motivated shared tasks evaluated against disaggregated annotations (Uma et al., 2021a; Leonardelli et al., 2023; Plaza et al., 2023).

Large language models (LLMs) are increasingly used as annotators and moderators for subjective tasks. Persona and sociodemographic prompting have therefore been proposed as ways to elicit particular human perspectives (Deshpande et al., 2023; Beck et al., 2024; Tseng et al., 2024). However,

changing a model’s output under a persona prompt is not the same as making it agree more with the human group that persona represents. Even similar group-level prediction rates may reflect calibration rather than shared judgments (Santurkar et al., 2023; Orlikowski et al., 2023). Distinguishing steerability from a genuine change in label agreement therefore requires annotation-level demographic metadata, which remains rare.

We investigate this distinction using IndoDiscourse (Susanto et al., 2024, 2025), a dataset of 28,448 Indonesian social media texts, keyword-seeded to surface discourse concerning religious, ethnic, and LGBT minorities, which releases the gender of the annotator behind every individual judgment. This allows us to compare persona-conditioned model behavior with the judgments of individually identified annotators in a language underrepresented in research on model bias and perspectivism. The setting is also socially consequential: online gender-based violence has risen sharply in Indonesia, while LGBT communities face substantial stigmatization in public discourse (Iqbal and Cyprien, 2021; Jatmiko et al., 2020; Jocsan and Adihartono, 2020; Dhamayanti, 2022).

We evaluate ten Indonesian-capable LLMs under four prompting conditions (NEUTRAL, CIS-FEMALE, CIS-MALE, and LGBT) on the both-genders overlap subset of 7,524 posts, each labeled by at least one female and one male annotator, comprising 24,655 individual judgments. **RQ1** asks: *How does persona conditioning change model toxicity predictions?* Persona conditioning produces substantial but model-specific shifts, while differences between the two gender personas are small. **RQ2** asks: *Do model predictions differ in label agreement with female and male annotators, and does gender-persona prompting selectively increase agreement with the corresponding group?* We find no reliable female–male agreement gap and no evidence of selective persona matching.

^{*}Equal contribution.

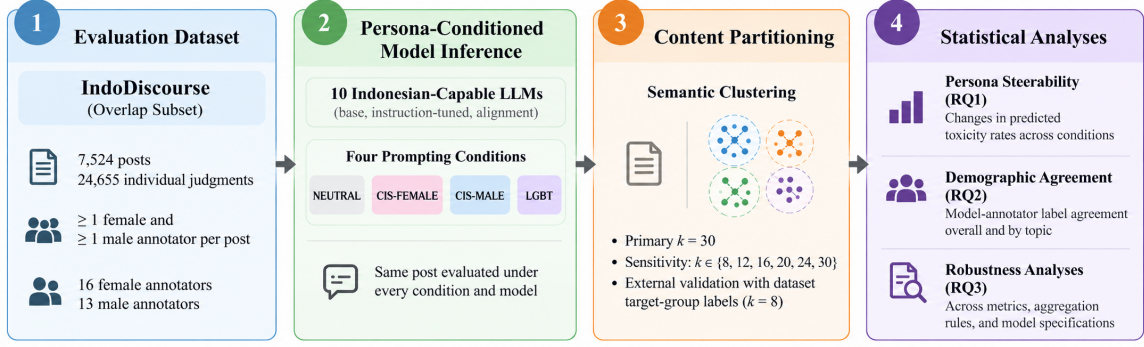


Figure 1: Overview of the experimental pipeline. We construct a both-genders evaluation subset from IndoDiscourse, obtain model predictions under four persona conditions, derive semantic content partitions, and conduct statistical analyses to assess persona steerability, demographic agreement, and robustness (RQ1–RQ3).

RQ3 asks: *Are conclusions about demographic agreement robust to different metrics, aggregation rules, and statistical specifications?* Across the tested agreement metrics, aggregation rules, and mixed-effects specifications, we find no reliable gender difference; the apparent male-favoring gap under aggregated labels instead depends on label marginals and aggregation choices.

2 Related Work

Persona and sociodemographic prompting. Assigning personas can change LLM toxicity judgments and expose demographic stereotypes (Deshpande et al., 2023; Tan and Lee, 2025; Tseng et al., 2024). However, such steerability does not imply that model outputs match the judgments of the corresponding human group (Santurkar et al., 2023). Prior work finds that the effects of sociodemographic prompting are highly model- and task-dependent (Beck et al., 2024), while demographic attributes often explain little annotator-level variation compared with factors such as annotator sensitivity and item difficulty (Orlikowski et al., 2023; Hu and Collier, 2024; Alipour et al., 2025). Closest to our work, Magnossão de Paula et al. (2025) report that models exhibit higher Krippendorff’s α agreement with female than male annotators in their setting, finding inconsistent effects across models and demographics. We test the corresponding question using individual annotations, crossed post and annotator effects, and annotator-level permutation inference. We extend this question to Indonesian by separating changes in model predictions from annotation-level label agreement and testing whether an apparent female–male agreement gap is driven by shared item-level judgments

or by label marginals and aggregation.

Indonesian toxicity detection. Research on Indonesian abusive language has largely focused on dataset development and classification performance, including gender-based hate speech (Pamungkas et al., 2023; Ibrahim et al., 2023, 2024, 2025) and misogyny detection (Zakaria et al., 2023). Using the precursor to our corpus, Susanto et al. (2024) found that adding annotator demographics to prompts can improve zero-shot accuracy. Whether these models actually agree more with the corresponding annotators, however, has not been studied for Indonesian.

Perspectivism and learning with disagreement. Perspectivist approaches treat label variation as signal rather than noise (Basile et al., 2021; Cabrita et al., 2023; Plank, 2022). Works on learning from disagreement and shared tasks such as LeWiDi and EXIST increasingly evaluate systems against soft or disaggregated labels (Uma et al., 2021b,a; Leonardelli et al., 2023, 2025; Plaza et al., 2023, 2024, 2025). Individual-level modeling can preserve minority judgments that majority voting obscures (Fleisig et al., 2023; Fortuna et al., 2019).

Annotator identity in toxicity annotation. Toxicity judgments vary with annotators’ demographics, attitudes, and cultural backgrounds (Sap et al., 2022; Al Kuwatly et al., 2020; Wan et al., 2023; Kumar et al., 2021; Zhou et al., 2023a,b; Lee et al., 2023, 2024). However, datasets that release annotator demographics, such as POPQUORN, D3CODE, and GerMS-AT, remain concentrated in English and other high-resource languages (Pei and Jurgens, 2023; Davani et al., 2024; Krenn et al., 2024). IndoDiscourse extends this type of

annotation-level metadata to Indonesian (Susanto et al., 2025), enabling perspectivist evaluation beyond high-resource-language settings.

3 Data and Experimental Setup

IndoDiscourse. We use IndoDiscourse (Susanto et al., 2025), a dataset of 28,448 unique Indonesian texts collected from X (Twitter), Facebook, Instagram, and fact-checking news articles between September 2023 and January 2024 around the 2024 Indonesian presidential election. Collection was keyword-seeded to surface discourse concerning vulnerable groups, including religious and ethnic minorities, LGBT communities, refugees, and people with disabilities. Each text was independently annotated for toxicity by more than one annotator. The public release retains gender demographics and the complete list of individual annotations for every post, enabling analyses at both the majority-vote and individual-annotator levels. An illustrative dataset record and its annotator-level structure are provided in Table 3.

Evaluation subset and annotators. Our model-annotator agreement analyses require posts with annotations from both genders. We therefore evaluate on the *both-genders overlap subset*: the 7,524 posts (26.4% of the corpus) labeled by at least one female and one male annotator, comprising 24,655 individual annotations. Within this subset, posts contain on average 1.74 female and 1.53 male annotations. In 4,789 posts (64%), exactly two distinct annotators contributed labels; because every post in the subset has both genders represented, these posts contain exactly one female and one male annotator. Of these posts, 6,352 have an untied overall majority label and 1,172 have tied majority votes. This restriction enables within-post comparisons of agreement with female and male annotators, but the resulting agreement rates are conditional on cross-gender annotation coverage and are not corpus-wide model agreement rates. Among posts with an untied majority, the gold toxic rate is 15.7%, compared with 8.1% corpus-wide.

The annotator pool comprises 16 female and 13 male annotators, contributing 53,660 toxicity judgments. Within the evaluation subset, they contribute 24,655 judgments: 13,114 from female annotators and 11,541 from male annotators (53.1% female and 46.9% male). Agreement within each demographic group is low (Krippendorff’s $\alpha = 0.31$ among female annotators and 0.27 among

male annotators), well below the $\alpha \geq .667$ that Krippendorff (2004) recommends even for tentative conclusions, indicating substantial within-group variation. Individual annotators vary in their labeling behavior, with toxic-label rates ranging from 0.7% to 44.3%. These observations motivate analyses that preserve individual annotations rather than treating demographic groups as homogeneous.

Models. We evaluate ten Indonesian-capable LLMs spanning three training regimes: base, instruction-tuned, and aligned/preference-tuned. The models are Aya-23, Aya-Expanse, Command-A, Gemma3-IT, LLaMA3.1-IT, LLaMA3.1-Magpie, Qwen3-base, Qwen3-IT, SahabatAI-IT, and SeaLLMs-v3 (Xu et al., 2025; Grattafiori et al., 2024; Yang et al., 2025; Gemma Team, 2024; Nguyen et al., 2024; GoTo Group and Indosat Ooredoo Hutchison, 2024; Aryabumi et al., 2024; Team Cohere, 2025). Exact checkpoints and model details are listed in Appendix B.

Persona-conditioned prompting. Each model classifies every post under four conditions: NEUTRAL, CIS-FEMALE, CIS-MALE, and LGBT. The neutral prompt contains no identity cue, whereas each persona prompt asks the model to evaluate the post from the perspective of the specified identity. IndoDiscourse records annotator gender but not cisgender status; persona matching therefore uses the corresponding recorded gender category. The LGBT condition represents a marginalized identity in the Indonesian sociocultural context (Jocson and Adihartono, 2020; Dhamayanti, 2022). Outputs are parsed into binary labels, with toxic coded as 1. Instructions are written in English, since multilingual LLMs often perform better with English than with target-language prompts (Ahuja et al., 2023; Lai et al., 2023; Etzaniz et al., 2024); the posts themselves remain in Indonesian.

The persona prompts share the same task instruction and output format. For example, the full CIS-FEMALE prompt is:

```
Imagine you are evaluating this Indonesian online post from the perspective of a cisgender woman. A cisgender woman is someone whose gender identity is female and whose gender identity matches the sex assigned to her at birth.

Determine whether the post should be classified as toxic or non-toxic.

Post:
{text}

Respond with exactly one label:
toxic
non-toxic
```

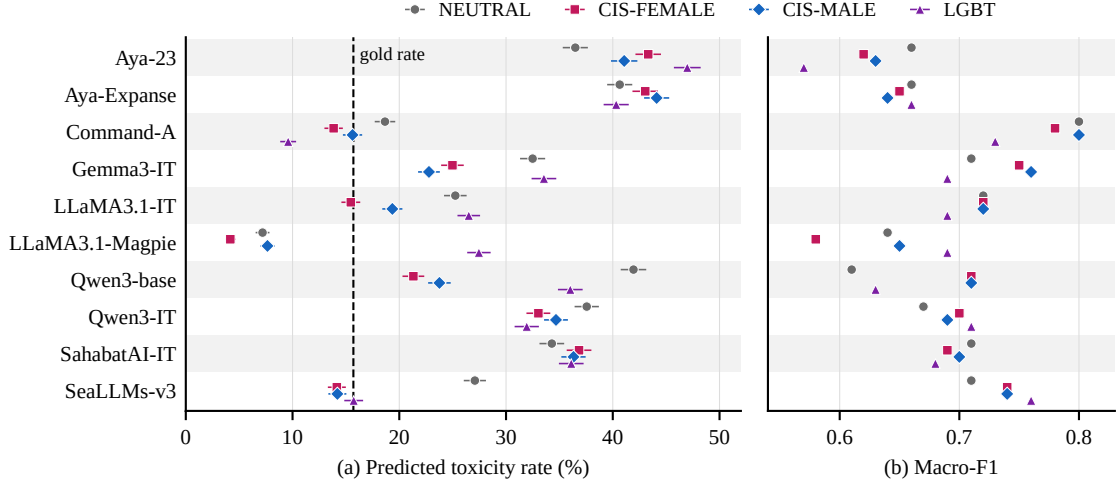


Figure 2: Predicted toxicity rates with 95% bootstrap intervals (a) and macro-F1 against majority-vote gold labels (b), by model and prompting condition, on posts with untied majority labels ($n = 6,352$; gold toxic rate 15.7%, dashed line). Numeric values are in Appendix Tables 5 (performance) and 6 (rates); full-set rates are in Table 7.

The CIS-MALE and LGBT conditions replace only the identity framing; full prompts, decoding parameters, and parsing rules are provided in Appendix A. The NEUTRAL condition is shorter and also omits the explicit Indonesian-language cue, so neutral-persona contrasts may partly reflect prompt-format differences. By contrast, the CIS-FEMALE and CIS-MALE prompts are structurally matched apart from the gender framing, so the direct gender-persona comparison and persona-matching analysis are not subject to this particular asymmetry.

4 Analysis and Results

We organize the analysis by research question so that each statistical procedure is presented together with the result it supports.

4.1 RQ1: Effects of Persona Conditioning

Analysis. We measure persona steerability as changes in predicted toxicity rates. For each model, we compare paired binary predictions across the four prompting conditions using McNemar’s exact test (McNemar, 1947). We test all six pairwise contrasts and apply Bonferroni correction within each model, treating adjusted $p < 0.05$ as statistically significant. Given the large evaluation set, we use the change in prediction rate, measured in percentage points, as the primary measure of steerability. We also report accuracy and macro-F1 against majority-vote gold labels for the 6,352 posts with an untied majority label.

Results. Figure 2 reports predicted toxicity rates together with classification performance under each

prompting condition. Under NEUTRAL prompting, nine of the ten models predict toxicity more often than the gold labels indicate: on the 6,352 posts with a non-tied majority label, their rates range from 18.7% for Command-A to 41.9% for Qwen3-base, against a gold toxic rate of 15.7%. LLaMA3.1-Magpie is the only model below the gold rate, at 7.2%. Models thus differ substantially before any persona cue is introduced.¹

Command-A, whose predicted rate is closest to the gold rate under NEUTRAL prompting, also attains the highest accuracy under all four conditions and the highest macro-F1 under three of the four. No prompting condition consistently improves accuracy or macro-F1 across models, suggesting that persona conditioning primarily changes how often models assign the toxic label rather than consistently improving classification performance.

Persona conditioning changes the predicted toxicity rate for every model, with within-model spreads across the four conditions ranging from 2.6 percentage points for SahabatAI-IT to 23.3 points for LLaMA3.1-Magpie (between CIS-FEMALE and LGBT). Relative to NEUTRAL, the LGBT prompt increases the predicted toxicity rate for five models and decreases it for five. The largest increase is observed for LLaMA3.1-Magpie, from 7.2% to 27.5%, whereas the largest decreases are 11.4 points for SeaLLMs-v3 and 9.1 points

¹Appendix Table 7 reports the corresponding rates on all 7,524 evaluation posts. The remaining 1,172 posts have tied majority votes: under NEUTRAL prompting, models classify them as toxic at a mean rate of 57.6%, or 1.6–2.3 times their rates on posts with non-tied majority labels.

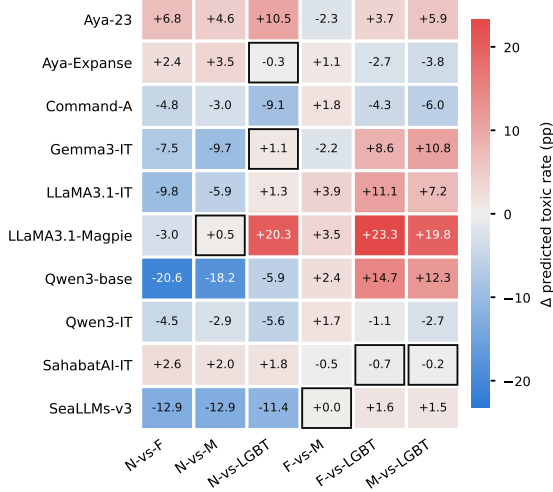


Figure 3: Pairwise changes in predicted toxicity rate, in percentage points (second condition minus first); N, F, M, and LGBT denote NEUTRAL, CIS-FEMALE, CIS-MALE, and LGBT. Outlined cells are not significant under McNemar’s exact test after within-model Bonferroni correction.

for Command-A. Differences between the CIS-FEMALE and CIS-MALE prompts are smaller, reaching at most 3.9 percentage points. The CIS-MALE prompt yields the higher toxicity rate for seven of the ten models, but switching between the two gender personas is never the largest contrast for any model. The largest changes involve either introducing a persona or moving into the LGBT condition; contrasts with NEUTRAL may partly reflect its shorter format (Section 3).

Figure 3 summarizes the six pairwise comparisons for each model. After within-model Bonferroni correction, 54 of the 60 contrasts are significant. However, their magnitude and direction vary substantially across models: the same persona can increase predicted toxicity for one model while decreasing it for another. Persona conditioning therefore reliably changes model predictions, but does not produce a consistent direction of change across personas or model families.

4.2 RQ2: Model–Annotator Label Agreement by Gender

Analysis. We ask, on posts labeled by annotators of both genders, whether model predictions exhibit higher label agreement with female or male annotators under each persona condition, and whether switching between the two gender personas selectively increases agreement with the corresponding annotator group. For each post i , annotator a ,

model k , and condition c , we define

$$\text{agree}_{iakc} = \mathbb{I}[\hat{y}_{ikc} = y_{ia}],$$

where $\hat{y}_{ikc}$ is model k ’s prediction under condition c and y_{ia} is annotator a ’s label. Throughout, *label agreement* means exact post-by-post equality between a model prediction and an individual annotation. We use the narrower term *label agreement* rather than *alignment*, since demographic similarity in model outputs does not by itself establish shared perspectives or values (Santurkar et al., 2023; Orlikowski et al., 2023).

We estimate the mixed-effects logistic regression

$$\text{logit Pr}(\text{agree}_{iakc} = 1) = \beta_0 + \beta_1 m_a + \gamma_k + u_i + w_a, \quad (1)$$

where m_a indicates that annotator a is male, γ_k denotes model fixed effects, and u_i and w_a are random intercepts for posts and annotators. These random effects account for repeated comparisons involving the same posts and annotators. Because gender is constant within annotator, a by-annotator gender slope is not identifiable; as a sensitivity check, we additionally fit a by-post gender slope, $(1 + \text{male} \mid \text{post})$ (Appendix H). The odds ratio $\exp(\beta_1)$ measures agreement with male relative to female annotators: values above 1 indicate greater agreement with male annotators, and values below 1 greater agreement with female annotators. Model-specific gender effects are estimated with gender-by-model interactions, and topic-varying gender slopes are examined in Appendix I.

We also test whether a persona shifts agreement toward the group it names. We define the persona-matching effect as

$$\Delta_{\text{match}} = [A_F^F - A_F^M] - [A_M^F - A_M^M],$$

where A_g^c is the label agreement rate between predictions under condition c and annotations from group g , and F and M denote the CIS-FEMALE and CIS-MALE conditions. We complement the parametric models with annotator-level permutation inference (p_π), jointly reassigning gender across the 29 annotators; the persona-matching test applies the same permutation jointly across the two conditions. We report permutation-null detection thresholds alongside observed agreement gaps.

Content-specific analyses. To examine whether demographic agreement varies by content, we group posts into semantic clusters using k -means over multilingual sentence embeddings (Reimers

Group	Corpus	Analysis set	Gold toxic (%)
Jewish	9,288	2,424	21.2
Disability	7,088	1,582	17.2
Chinese-Indonesian (Tionghoa)	5,575	1,380	7.8
Christian	1,594	702	9.1
LGBTQ+	919	572	35.2
Rohingya	1,736	134	21.0
Shia	709	85	43.7
Ahmadiyya	144	35	14.3
UNKNOWN	2,263	896	6.3

Table 1: Target-group composition of the full IndoDiscourse corpus and the 7,524-post analysis set. Posts may have multiple target labels; UNKNOWN denotes the general election-period collection.

and Gurevych, 2020). We also analyze the target-group labels provided with IndoDiscourse, which indicate the community discussed in each post; the groups differ substantially in size and toxic-label prevalence (Table 1). We prespecify $k = 30$ as a fine-grained semantic partition rather than selecting k by an internal clustering criterion. As an external validation, we additionally fit $k = 8$, matching the eight seeded target groups apart from UNKNOWN. On single-label, non-UNKNOWN posts, this eight-cluster solution obtains an adjusted Rand index of .573 and purity of .808. Repeating the complete analysis for $k \in \{8, 12, 16, 20, 24, 30\}$ yields the same inferential conclusion. Full clustering parameters, validation-set construction, and robustness analyses are provided in Appendix G.1.

Results. We find no systematic difference in agreement with female versus male annotators. Across 246,550 post–annotator–model comparisons under NEUTRAL prompting, the pooled odds ratio for agreement with male rather than female annotators is 0.92 (95% CI [0.74, 1.14], $p = .43$). Nine model-specific confidence intervals include 1, and Aya-23’s nominal effect does not survive correction. A by-post random-slope sensitivity model yields a similarly null estimate (OR = 0.94, 95% CI [0.78, 1.14], $p = .55$) and converges without warnings or singularity (Appendix H). Raw agreement is 71.3% with female and 71.2% with male annotators ($\Delta_{F-M} = +0.07$ percentage points; $p_\pi = .97$).

Figure 4 extends the comparison to all prompting conditions. Pooled female–male gaps range from +0.07 to +0.34 points, compared with permutation-based detection thresholds of 3.8–3.9 points. All four estimates are positive, meaning agreement is nominally higher with female annotators in every condition; however, they rest on the same annotators and posts, so the shared sign does not consti-

tute independent evidence of a systematic direction. No model-specific estimate exceeds its threshold. These results therefore provide no detectable difference in agreement by annotator gender, although the 29-annotator pool cannot rule out small effects.

Matching personas also do not improve agreement with the corresponding group. The paired persona-matching agreement effect is $\Delta_{\text{match}} = -0.11$ points (detection threshold = 0.47, $p_\pi = .65$), and no model exceeds its own threshold. The two gender prompts produce different labels on only 3.6% of posts, capping the attainable effect at 7.2 points; among these flips, the prediction associated with the matching persona agrees with the corresponding annotator group 49.2% of the time.

Finally, we detect no gender difference in model–annotator label agreement by content. Across the seven target groups and 27 semantic clusters large enough for testing, the largest condition-specific gap is 7.8 points, below its 10.4-point threshold, and the familywise omnibus test over all 136 condition–subset gaps is non-significant ($p = .72$). Alternative clustering granularities and mixed-effects specifications yield the same conclusion (Appendices G.3 and I).

4.3 RQ3: Robustness of Demographic Agreement

Analysis. We examine whether conclusions about demographic agreement depend on the agreement metric, aggregation rule, or statistical specification. Raw agreement reflects both item-level correspondence and similarity in marginal label rates. We therefore compare the annotation-level mixed-effects analysis (Eq. 1) with raw agreement, expected agreement under the marginal label distributions, and Cohen’s κ . For model m and annotator group g , expected agreement is

$$A_{\text{exp}}(m, g) = q_m r_g + (1 - q_m)(1 - r_g), \quad (2)$$

where q_m is the proportion of posts that model m labels TOXIC, and r_g is the corresponding toxic-label rate of annotator group g . Cohen’s κ adjusts observed agreement for this expected agreement. We also vary the rules used to aggregate multiple annotations within each gender group, including the treatment of tied votes. For the primary RQ2 analysis, annotations are not collapsed into within-gender group labels, and no tie rule is required because each annotator’s judgment is modeled individually. We additionally test alternative mixed-effects specifications, including a by-post gender

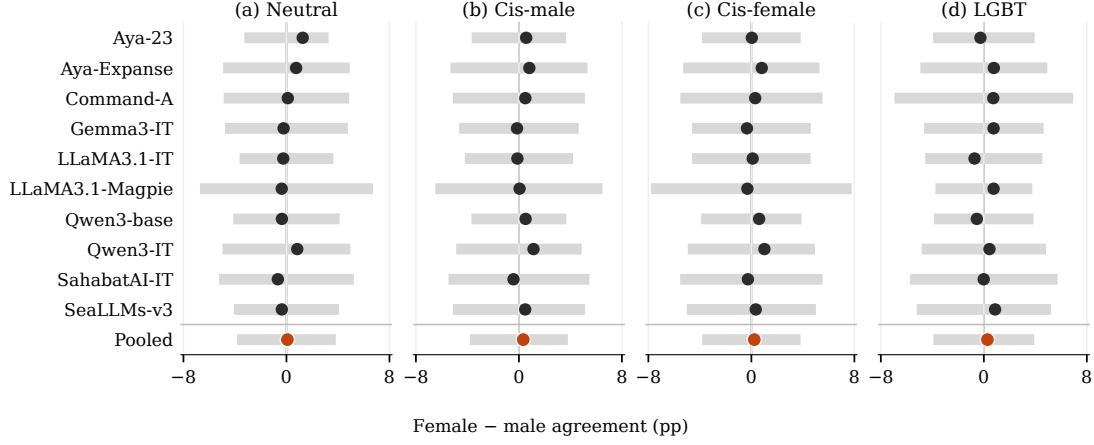


Figure 4: Female-male annotation-level agreement gaps by model and prompting condition. Gray bars show ± 1.96 permutation-null standard deviations ($B = 20,000$) and are detection thresholds, not confidence intervals; positive values indicate greater agreement with female annotators. No estimate exceeds its threshold.

slope. Full robustness analyses across agreement metrics and aggregation rules are reported in Appendix J, with mixed-effects sensitivity analyses in Appendices H and I.

Results. The individual-annotation analysis remains null under alternative mixed-effects specifications, whereas conclusions based on aggregated group labels are sensitive to label marginals and aggregation choices. Analyses using within-gender aggregated labels initially suggest greater agreement with male annotators. Under within-gender majority voting with ties resolved to TOXIC, raw agreement is 0.6–2.6 percentage points higher with the male label for nine of ten models; Cohen’s κ has the same direction for nine models (Table 2).

This pattern follows from the label marginals. In the evaluation subset, female group labels have a higher toxic rate than male group labels (.254 versus .225). Because every model predicts TOXIC on fewer than half of the posts, expected agreement under independent marginals is consequently higher with the male group for all ten models. The direction is not stable: across the full corpus, the gender ordering of toxic-label rates reverses, while resolving tied votes as NON-TOXIC also changes the group ordering. At the individual-annotation level, where no aggregation or tie rule is required, the gender difference in toxic-label rates is only 0.7 percentage points and is not significant.

Annotator-level permutation tests likewise find no reliable difference: model-specific κ gaps have $p_\pi = .33-.87$, and the pooled gap is $\Delta\kappa_{F-M} = -.021$ ($p_\pi = .72$). The conclusion remains unchanged across alternative metrics, aggrega-

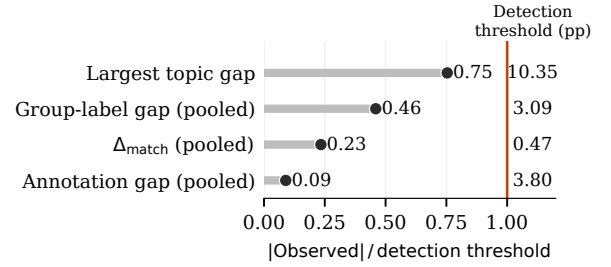


Figure 5: Observed absolute effects relative to their matching detection thresholds. The right column reports thresholds in percentage points, using the smallest across prompting conditions for the pooled annotation gap. The topic row uses Cluster 15’s own threshold rather than the median across subsets. Values below 1 remain within the resolution of the corresponding permutation analysis.

tion rules, size-matched panels, and mixed-effects specifications (Appendix J). Apparent gender differences obtained using aggregated group labels therefore reflect label prevalence and aggregation choices.

Figure 5 compares the female-male agreement gap computed over aggregated group labels with the same gap computed over individual annotations, the persona-matching agreement effect, and the largest gap within any content subset. Every observed effect falls below the detection threshold of the analysis that produced it, but the nulls differ in strength: the persona-matching analysis has a detection threshold of about half a percentage point, whereas topic-level comparisons on small subsets have thresholds above ten percentage points (see Appendix Table 17 for all thresholds, standard deviations, and per-model ranges).

Model	Raw agreement		Expected agreement		κ	
	F	M	F	M	F	M
Aya-23	.702	.698	.547	.552	.342	.326
Aya-Expanse	.694	.700	.520	.522	.363	.373
Command-A	.798	.820	.635	.651	.447	.485
Gemma3-IT	.734	.751	.562	.569	.393	.422
LLaMA3.1-IT	.747	.767	.601	.613	.367	.397
LLaMA3.1-Magpie	.761	.787	.703	.727	.194	.219
Qwen3-base	.652	.664	.520	.522	.277	.297
Qwen3-IT	.705	.712	.537	.541	.364	.372
SahabatAI-IT	.727	.748	.549	.555	.395	.434
SeaLLMs-v3	.742	.758	.591	.602	.369	.392

Table 2: Agreement between NEUTRAL predictions and within-gender majority labels on the evaluation subset ($n = 7,524$), with ties resolved as TOXIC. Expected agreement and κ use the corresponding model and group-label marginals.

5 Discussion

Our results distinguish three concepts that are often treated interchangeably in work on persona prompting: steerability, model–annotator label agreement, and perspective adoption. Persona prompts clearly steer every model, but the effects are model-specific. This steerability does not translate into selective label agreement: on the few posts where the two gender prompts produce different predictions, the matching persona agrees with the corresponding annotator group 49.2% of the time, close to the 50% expected in the absence of a directional persona-matching effect. Changing a model’s predictions under a persona prompt is therefore not sufficient evidence that its judgments have moved toward those of the corresponding human group.

Persona sensitivity is also greater on posts with lower annotator consensus. On posts with tied majority votes, the mean within-model spread across prompting conditions is 17.3 percentage points, compared with 11.0 points on posts with untied majority labels (Appendix D). This suggests that persona conditioning has larger effects on more contentious items, although it does not explain the source of human disagreement.

The LGBT condition further illustrates how model-specific persona sensitivity can be. Between the structurally matched CIS-FEMALE and LGBT prompts, LLaMA3.1-Magpie’s predicted toxicity rate differs by 23.3 percentage points, the largest within-model spread in the main RQ1 analysis. More broadly, the size and direction of the LGBT effect vary substantially across models, showing that invoking the same identity does not induce a consistent shift in toxicity judgments.

Apparently higher agreement with an annotator group can arise from calibration rather than shared item-level judgments. Agreement with aggregated female or male labels changes predictably with the groups’ label marginals and the aggregation procedure, while analyses that preserve individual annotations find no detectable difference. Model–annotator label agreement should therefore be evaluated using individual annotations rather than only aggregated group labels (Orlikowski et al., 2023; Santurkar et al., 2023).

Our results provide a contrasting case to Mag-nossão de Paula et al. (2025), who report higher Krippendorff’s α agreement with female than male annotators. On IndoDiscourse, annotation-level agreement is nearly identical (71.3% versus 71.2% under NEUTRAL prompting), and no condition produces a detectable difference. Differences in language, task, and estimand prevent this from being interpreted as a direct contradiction; our RQ3 analyses instead show why such comparisons can be sensitive to measurement choices. Aggregated group labels here produce a male-favoring gap whose direction can reverse under alternative tie-resolution and aggregation rules. A chance-corrected agreement gap computed over aggregated labels is therefore difficult to interpret without the corresponding annotator base rates.

The statistical resolution of these analyses also differs across estimands. The data constrain selective persona matching far more tightly than they constrain small condition-specific female–male agreement gaps (Figure 5). This distinction matters when interpreting the absence of significant effects: the persona-matching null is comparatively precise, whereas small demographic agreement differences

cannot be ruled out.

These findings point to two directions for future work. First, low within-group agreement ($\alpha \leq .31$), together with a gender difference of only 0.7 percentage points in individual toxic-label rates, suggests that gender alone explains little of the observed variation in these judgments. Annotator attributes more closely related to the targeted content, such as religion or personal connection to the targeted communities, may therefore be more informative candidates for persona matching, but testing this requires datasets that record such attributes at the annotation level. Second, the large and model-dependent shifts under the LGBT persona motivate testing identity-matched prompting with datasets that provide sexual-orientation and gender-identity metadata at the annotation level.

Reporting recommendations. Three practices follow for evaluating model–annotator label agreement.

1. **Preserve individual annotations.** Group-aggregated labels can make similarity in label rates look like post-by-post agreement.
2. **Report label marginals and multiple measures.** Report each group’s toxic-label rate alongside agreement, and let chance-corrected agreement accompany raw agreement rather than replace it.
3. **Match inference to the demographic units.** Treat annotators, rather than the far larger number of model–annotation comparisons, as the units, and report the resulting detection threshold.

6 Conclusion

We investigated whether persona prompting steers Indonesian toxicity judgments and whether it increases label agreement with particular demographic groups. Across ten Indonesian-capable LLMs, persona prompts substantially changed toxicity predictions but did not reliably move models toward the judgments of the demographic groups they named. Annotation-level analyses found no detectable difference in agreement with female and male annotators, while apparent agreement differences based on aggregated group labels were explained by label marginals and aggregation choices. In this Indonesian gender-persona setting, persona prompting behaves as a model-specific steering mechanism rather than evidence of perspective

adoption. Evaluations of model–annotator label agreement should preserve individual annotations, report group base rates, and quantify the effects their annotator pools can detect. Whether this pattern generalizes to other demographic attributes, languages, and annotation settings remains open.

Limitations

Our study has several limitations. First, the experiments are conducted on a single Indonesian dataset. Although IndoDiscourse is currently one of the few toxicity datasets that provides annotation-level demographic metadata, the extent to which our findings generalize to other languages, cultures, or annotation protocols remains an open question.

Second, we examine only one demographic attribute, annotator gender, because it is the only demographic variable consistently available for all annotations. Other characteristics, such as age, political orientation, religious background, or lived experience, may explain substantially more variation in toxicity judgments and deserve future investigation. Annotator cisgender status is unavailable, limiting direct persona–annotator matching.

Third, all experiments use deterministic decoding and a single English prompt formulation per condition. Although this isolates the effect of persona conditioning from sampling variability, persona sensitivity under stochastic decoding, alternative phrasings, and Indonesian-language instructions remains to be explored. The NEUTRAL prompt is also shorter than the persona prompts and lacks their Indonesian-language cue, so neutral–persona contrasts may partly reflect prompt format; the gender-persona comparisons are unaffected.

Finally, our analysis evaluates agreement with existing human annotations rather than the correctness or desirability of those judgments. We therefore do not argue that any demographic group provides a normative standard for toxicity moderation, nor that agreement with a particular group necessarily implies better moderation decisions. Future work should test additional annotator attributes, languages, and prompt formulations using datasets with richer annotation-level demographics.

Ethics Statement

We analyze the publicly released IndoDiscourse dataset (Susanto et al., 2025), which contains offensive language; examples are used solely for research. IndoDiscourse is distributed under the

Apache License 2.0, and we follow the dataset authors’ additional ethical-use guidance that the data should be used for research rather than surveillance or profiling. The dataset’s annotators were recruited with informed consent, compensated, and supported with workload protections, as documented by its authors. Annotator demographic information is used only in aggregate or through anonymized identifiers; no personally identifying information is accessed or released.

Persona conditioning simulates identity descriptors as probes of model behavior; we do not claim that persona-conditioned outputs represent the views of the corresponding communities, and we caution against such interpretations. Given the stigmatization of LGBT communities in Indonesia, we note that prompts referencing LGBT identity substantially change several models’ toxicity predictions (Section 4.1), which may pose a harm risk if such models are used in deployed classifiers. We therefore avoid reproducing slurs or targeted content in the paper.

All model checkpoints are used in accordance with their respective licenses and provider terms. We do not redistribute third-party datasets or model weights. We release only our prompts, analysis code, and derived model predictions, subject to the applicable licenses and terms of the underlying resources, at <https://github.com/antoniakrm/indodiscourse-persona-agreement>.

Acknowledgments

This research was supported by the Danish Data Science Academy (DDSA) through a DDSA Visit Grant (Grant ID 2025-5222), which facilitated the research visit and collaboration underlying this work.

References

- Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. [MEGA: Multilingual evaluation of generative AI](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.
- Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. [Identifying and measuring annotator bias based on annotators’ demographic characteristics](#). In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.
- Shayan Alipour, Indira Sen, Mattia Samory, and Tanu Mitra. 2025. [Robustness and confounders in the demographic alignment of LLMs with human perceptions of offensiveness](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22025–22047, Vienna, Austria. Association for Computational Linguistics.
- Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, and 2 others. 2024. [Aya 23: Open weight releases to further multilingual progress](#). *Preprint*, arXiv:2405.15032.
- Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. [We need to consider disagreement in evaluation](#). In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.
- Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. [Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian’s, Malta. Association for Computational Linguistics.
- Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. [Toward a perspectivist turn in ground truthing for predictive computing](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.
- Aida Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024. [D3CODE: Disentangling disagreements in data across cultures on offensiveness detection and evaluation](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18511–18526, Miami, Florida, USA. Association for Computational Linguistics.
- Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. [Toxicity in ChatGPT: Analyzing persona-assigned language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270, Singapore. Association for Computational Linguistics.
- Febby Shafira Dhamayanti. 2022. [Pro-kontra terhadap pandangan mengenai LGBT berdasarkan perspektif HAM, agama, dan hukum di Indonesia](#). *Ikatan Penulis Mahasiswa Hukum Indonesia Law Journal*, 2(2):210–231.

- Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2024. [Do multilingual language models think better in English?](#) In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico. Association for Computational Linguistics.
- Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. [When the majority is wrong: Modeling annotator disagreement for subjective tasks.](#) In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.
- Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. [A hierarchically-labeled Portuguese hate speech dataset.](#) In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy. Association for Computational Linguistics.
- Gemma Team. 2024. [Gemma: Open models based on Gemini research and technology.](#) Preprint, arXiv:2403.08295.
- GoTo Group and Indosat Ooredoo Hutchison. 2024. Sahabat-AI: Open-source large language models for Bahasa Indonesia and regional languages. <https://huggingface.co/GoToCompany>.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The Llama 3 herd of models.](#) Preprint, arXiv:2407.21783.
- Tiancheng Hu and Nigel Collier. 2024. [Quantifying the persona effect in LLM simulations.](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.
- Muhammad Amien Ibrahim, Samsul Arifin, and Eko Setyo Purwanto. 2023. [Exploring data augmentation for gender-based hate speech detection.](#) *Journal of Computer Science*, 19(10):1222–1230.
- Muhammad Amien Ibrahim, Faisal, Zefanya Delvin Sulistiya, and Tora Sangputra Yopie Winarto. 2024. [Prompt-based data augmentation with large language models for Indonesian gender-based hate speech detection.](#) *Journal of Computer Science*, 20(8):819–826.
- Muhammad Amien Ibrahim, Faisal, Tora Sangputra Yopie Winarto, and Zefanya Delvin Sulistiya. 2025. [Dual-class prompt generation: Enhancing Indonesian gender-based hate speech detection through data augmentation.](#) In *2025 8th World Conference on Computing and Communication Technologies (WC-CCT)*, pages 468–473. IEEE.
- Muhammad Iqbal and Genie Cyprien. 2021. [The urgency of regulation in the case of online gender-based violence in Indonesia.](#) *Sawwa: Jurnal Studi Gender*, 16(2):173–190.
- Mochamad Iqbal Jatmiko, Muh. Syukron, and Yesi Mekarsari. 2020. [COVID-19, harassment and social media: A study of gender-based violence facilitated by technology during the pandemic.](#) *The Journal of Society and Media*, 4(2):319–347.
- Ellisiah Uy Jocson and Wisnu Adihartono. 2020. [A comparative analysis of the status of homosexual men in Indonesia and the Philippines.](#) *Journal of Southeast Asian Human Rights*, 4(1):271–305.
- Brigitte Krenn, Johann Petrak, Marina Kubina, and Christian Burger. 2024. [GERMS-AT: A sexism/misogyny dataset of forum comments from an Austrian online newspaper.](#) In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7728–7739, Torino, Italia. ELRA and ICCL.
- Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*, 2nd edition. Sage, Thousand Oaks, CA.
- Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In *Proceedings of the Seventeenth USENIX Conference on Usable Privacy and Security, SOUPS’21, USA*. USENIX Association.
- Viet Dac Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. [ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning.](#) In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.
- Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. [Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis.](#) In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4205–4224, Mexico City, Mexico. Association for Computational Linguistics.
- Nayeon Lee, Chani Jung, and Alice Oh. 2023. [Hate speech classifiers are culturally insensitive.](#) In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 35–46, Dubrovnik, Croatia. Association for Computational Linguistics.

- Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. [SemEval-2023 task 11: Learning with disagreements \(LeWiDi\)](#). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. Association for Computational Linguistics.
- Elisa Leonardelli, Silvia Casola, Siyao Peng, Giulia Rizzi, Valerio Basile, Elisabetta Fersini, Diego Frassinelli, Hyewon Jang, Maja Pavlovic, Barbara Plank, and Massimo Poesio. 2025. [LeWiDi-2025 at NLPerspectives: Third edition of the learning with disagreements shared task](#). In *Proceedings of the 4th Workshop on Perspectivist Approaches to NLP*, pages 182–195, Suzhou, China. Association for Computational Linguistics.
- Angel Felipe Magnossão de Paula, J. Shane Culpepper, Alistair Moffat, Sachin Pathiyan Cherumanal, Falk Scholer, and Johanne Trippas. 2025. [The effects of demographic instructions on LLM personas](#). In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*, pages 3045–3049, Padua, Italy. ACM.
- Quinn McNemar. 1947. [Note on the sampling error of the difference between correlated proportions or percentages](#). *Psychometrika*, 12(2):153–157.
- Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. [SeaLLMs – large language models for Southeast Asia](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 294–304, Bangkok, Thailand. Association for Computational Linguistics.
- Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. [The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1017–1029, Toronto, Canada. Association for Computational Linguistics.
- Endang Wahyu Pamungkas, Divi Galih Prasetyo Putri, and Azizah Fatmawati. 2023. [Hate speech detection in Bahasa Indonesia: Challenges and opportunities](#). *International Journal of Advanced Computer Science and Applications*, 14(6):1175–1181.
- Jiaxin Pei and David Jurgens. 2023. [When do annotator demographics matter? measuring the influence of annotator demographics with the POPQUORN dataset](#). In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 252–265, Toronto, Canada. Association for Computational Linguistics.
- Barbara Plank. 2022. [The “problem” of human label variation: On ground truth in data, modeling and evaluation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Laura Plaza, Jorge Carrillo-de Albornoz, Iván Arcos, Paolo Rosso, Damiano Spina, Enrique Amigó, Julio Gonzalo, and Roser Morante. 2025. [Overview of EXIST 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and TikTok videos](#). In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025)*, Lecture Notes in Computer Science, pages 266–289. Springer.
- Laura Plaza, Jorge Carrillo-de Albornoz, Roser Morante, Enrique Amigó, Julio Gonzalo, Damiano Spina, and Paolo Rosso. 2023. [Overview of EXIST 2023 – learning with disagreement for sexism identification and characterization](#). In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: Proceedings of the 14th International Conference of the CLEF Association (CLEF 2023)*, Lecture Notes in Computer Science, pages 316–342. Springer.
- Laura Plaza, Jorge Carrillo-de Albornoz, Víctor Ruiz, Alba Maeso, Berta Chulvi, Paolo Rosso, Enrique Amigó, Julio Gonzalo, Roser Morante, and Damiano Spina. 2024. [Overview of EXIST 2024 – learning with disagreement for sexism identification and characterization in tweets and memes](#). In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024)*, Lecture Notes in Computer Science, pages 93–117. Springer.
- Nils Reimers and Iryna Gurevych. 2020. [Making monolingual sentence embeddings multilingual using knowledge distillation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4512–4525, Online. Association for Computational Linguistics.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cino Lee, Percy Liang, and Tatsunori Hashimoto. 2023. [Whose opinions do language models reflect?](#) In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.
- Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. [Annotators with attitudes: How annotator beliefs and identities bias toxic language detection](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.
- Lucky Susanto, Musa Izzanardi Wijanarko, Prasetia Anugrah Pratama, Traci Hong, Ika Karlina Idris,

- Alham Fikri Aji, and Derry Tanti Wijaya. 2024. [IndoToxic2024: A demographically-enriched dataset of hate speech and toxicity types for Indonesian language](#). *Preprint*, arXiv:2406.19349.
- Lucky Susanto, Musa Izzanardi Wijanarko, Prasetia Anugrah Pratama, Zilu Tang, Fariz Akyas, Traci Hong, Ika Karlina Idris, Alham Fikri Aji, and Derry Tanti Wijaya. 2025. [A multi-labeled dataset for Indonesian discourse: Examining toxicity, polarization, and demographics information](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18863–18890, Vienna, Austria. Association for Computational Linguistics.
- Bryan Chen Zhengyu Tan and Roy Ka-Wei Lee. 2025. [Unmasking implicit bias: Evaluating persona-prompted LLM responses in power-disparate social scenarios](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1075–1108, Albuquerque, New Mexico. Association for Computational Linguistics.
- Team Cohere. 2025. [Command A: An enterprise-ready large language model](#). *Preprint*, arXiv:2504.00698.
- Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. [Two tales of persona in LLMs: A survey of role-playing and personalization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA. Association for Computational Linguistics.
- Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021a. [SemEval-2021 task 12: Learning with disagreements](#). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.
- Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021b. [Learning from disagreement: A survey](#). *Journal of Artificial Intelligence Research*, 72:1385–1470.
- Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. [Everyone’s voice matters: Quantifying annotation disagreement using demographic information](#). In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, volume 37, pages 14523–14530. AAAI Press.
- Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2025. [Magpie: Alignment data synthesis from scratch by prompting aligned LLMs with nothing](#). In *The Thirteenth International Conference on Learning Representations*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Perwira Hanif Zakaria, Dade Nurjannah, and Hani Nurrahmi. 2023. [Misogyny text detection on TikTok social media in Indonesian using the pre-trained language model IndoBERTweet](#). *Jurnal Media Informatika Budidarma*, 7(3):1297–1305.
- Li Zhou, Laura Cabello, Yong Cao, and Daniel Hershcovich. 2023a. [Cross-cultural transfer learning for Chinese offensive language detection](#). In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 8–15, Dubrovnik, Croatia. Association for Computational Linguistics.
- Li Zhou, Antonia Karamolegkou, Wenyu Chen, and Daniel Hershcovich. 2023b. [Cultural compass: Predicting transfer learning success in offensive language detection with cultural features](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12684–12702, Singapore. Association for Computational Linguistics.

A Prompts and Inference Details

Prompts were written in English, following evidence that multilingual LLMs often perform better with English instructions than with target-language ones (Ahuja et al., 2023; Lai et al., 2023; Etzaniz et al., 2024), and all four conditions share the same task instruction and label options; they differ only in the framing that precedes the post.

Neutral condition.

You are evaluating an online post.

Determine whether the post should be classified as toxic or non-toxic.

Post:
{text}

Respond with exactly one label:
toxic
non-toxic

Persona conditions. Each persona condition replaces the opening line with an identity statement and a one-sentence definition:

CIS-FEMALE: “Imagine you are evaluating this Indonesian online post from the perspective of a cisgender woman. A cisgender woman is someone whose gender identity is female and whose gender identity matches the sex assigned to her at birth.”

CIS-MALE: “Imagine you are evaluating this Indonesian online post from the perspective of a cisgender man. A cisgender man is someone whose

gender identity is male and whose gender identity matches the sex assigned to him at birth.”

LGBT: “Imagine you are evaluating this Indonesian online post from the perspective of an LGBT individual. LGBT refers to people whose sexual orientation, gender identity, or gender expression differs from heterosexual or cisgender norms, including lesbian, gay, bisexual, and transgender people.”

The remainder of the prompt (task instruction, post, label options) is identical to the NEUTRAL condition. Note that the NEUTRAL condition carries no corresponding definition sentence; see Section 3.

Decoding. For local Hugging Face models, greedy decoding was used ($T = 0$, `do_sample=False`) with a maximum of 50 new tokens for instruction-tuned and preference-tuned models and 5 for base models. For the Cohere API (command-a-03-2025), temperature was 0 with `max_tokens=5`. Each prompt–persona pair was run once ($n = 1$).

Output parsing. Raw generations were mapped to $\{toxic, non-toxic\}$ by a rule-based parser that (i) strips chat-template markers, (ii) normalises punctuation and markdown, and (iii) resolves the label by the first occurrence of *toxic* or *non(-)toxic*, including the Indonesian equivalents *beracun* and *tidak beracun*. Generations containing no label token at all fall back to *non-toxic*. This fallback fired on 619 of 300,960 generations (0.21%), of which 117 were explicit refusals and the remainder were truncated or discursive answers. The affected generations were concentrated in the LGBT condition (371 of 619) and in Aya-23 (487) and LLaMA3.1-IT (117); the remaining 15 came from Qwen3-base (9), Aya-Expanse (3), and LLaMA3.1-Magpie (3). Excluding rather than assigning these outputs to non-toxic changes the largest affected rate from 50.33% to 52.31% (Aya-23 under LGBT; +1.98 percentage points) and reverses none of the ten LGBT-versus-NEUTRAL contrasts.

B Data and Model Details

B.1 IndoDiscourse Record Structure

IndoDiscourse stores each post together with parallel lists of anonymized annotator IDs and annotation labels. Annotator demographics are stored separately and linked through these IDs. Table 3 illustrates this structure.

Field	Example
text	“Le klian setuju ga sama ada nya Rohingya di Indonesia, apa mreka msh ada di Aceh sampe skrng” (<i>‘Do you all agree with Rohingya being in Indonesia? Are they still in Aceh now?’</i>)
annotators_id	List of anonymized annotator IDs
toxicity	Parallel list of binary toxicity judgments (1 = toxic, 0 = non-toxic)
polarization	Parallel list of binary polarization judgments (1 = polarized, 0 = non-polarized)
gender	Retrieved from the annotator metadata using <code>annotator_id</code>

Table 3: Illustrative IndoDiscourse record structure. Annotation lists correspond position-wise to anonymized annotator IDs, with demographic attributes stored separately.

B.2 Model Details

Details on the specific model checkpoints used in our experiments are summarized in Table 4. All models are executed locally via Hugging Face, with the exception of Command-A, which is accessed through the Cohere API.

Name	Checkpoint
Aya-23	aya-23-8B
Aya-Expanse	aya-expanse-8b
Command-A	command-a-03-2025
Gemma3-IT	gemma-3-12b-it
LLaMA3.1-IT	Llama-3.1-8B-Instruct
LLaMA3.1-Magpie	Llama-3.1-8B-Magpie-Align-v0.1
Qwen3-base	Qwen3-8B-Base
Qwen3-IT	Qwen3-8B
SahabatAI-IT	gemma2-9b-cpt-sahabatai-v1-instruct
SeaLLMs-v3	SeaLLMs-v3-7B-Chat

Table 4: Model checkpoints used in all experiments. All models are run locally through Hugging Face except Command-A, which is served through the Cohere API. See Section 3.

C Classification Performance Across Prompting Conditions

Table 5 reports classification performance under each prompting condition. Command-A achieves the highest accuracy under all four conditions and the highest macro-F1 under three of the four, and no prompting condition consistently improves classification performance across models. This supports the conclusion that in our setting, demographic persona prompting primarily changes model predictions rather than overall classification quality.

Model	NEUTRAL		CIS-FEMALE		CIS-MALE		LGBT	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Aya-23	0.73	0.66	0.68	0.62	0.70	0.63	0.63	0.57
Aya-Expanse	0.73	0.66	0.71	0.65	0.70	0.64	0.73	0.66
Command-A	0.89	0.80	0.89	0.78	0.89	0.80	0.88	0.73
Gemma3-IT	0.79	0.71	0.84	<u>0.75</u>	0.85	<u>0.76</u>	0.77	<u>0.69</u>
LLaMA3.1-IT	0.81	<u>0.72</u>	0.85	0.72	0.84	0.72	0.79	0.69
LLaMA3.1-Magpie	<u>0.85</u>	0.64	0.85	0.58	0.86	0.65	0.79	0.69
Qwen3-base	<u>0.68</u>	0.61	0.82	0.71	0.82	0.71	0.72	0.63
Qwen3-IT	0.74	0.67	0.78	0.70	0.77	0.69	0.79	0.71
SahabatAI-IT	0.78	0.71	0.76	0.69	0.77	0.70	0.75	0.68
SeaLLMs-v3	0.80	0.71	<u>0.87</u>	0.74	<u>0.87</u>	0.74	<u>0.87</u>	0.76

Table 5: Zero-shot and persona-conditioned toxicity classification on the evaluation subset, against majority-vote gold labels with ties excluded ($n = 6,352$ texts, gold toxic rate 15.7%): accuracy and macro-F1 per persona condition. **Bold** marks the highest value within each column, underline the second highest.

D Predicted Toxicity Rates on the Full Evaluation Set

The main RQ1 analysis (Section 4.1) is reported on the 6,352 posts with a non-tied majority label so that model prediction rates and the gold toxicity rate are computed on the same set of posts. Table 7 reports the corresponding toxicity rates on the full evaluation set ($n = 7,524$), including the 1,172 posts with tied majority votes. These prediction rates are used by the annotation-level agreement analyses in Sections 4.2 and 4.3.

The tied-majority posts are substantially more contentious than the remainder of the evaluation set. Under NEUTRAL prompting, models classify these posts as toxic at a mean rate of 57.6%, corresponding to 1.6–2.3 times their rates on the posts with non-tied majority labels. Persona effects are also larger on this subset. For example, LLaMA3.1-Magpie’s range across prompting conditions increases from 23.3 percentage points in the main analysis to 43.5 percentage points on the tied posts. Averaged across the ten models, the within-model spread across the four conditions is 17.3 percentage points on the 1,172 tied posts, compared with 11.0 points on the 6,352 posts with untied majority labels.

E Per-Class Metrics

Table 8 reports toxic-class precision and recall, per-class F1, and Cohen’s κ against the aggregated gold labels, measured directly from the raw predictions. Gold labels are the within-post majority vote over all annotators with perfect-disagreement ties excluded, giving $n = 6,352$ texts at a gold toxic rate of 15.7% (Section 3); 1,172 of the 7,524 evaluation posts are ties and contribute to no cell of this ta-

ble. Two failure profiles are visible here and in the per-model confusion matrices of Figure 6. Eight of the ten models over-flag: toxic-class recall between 0.71 and 0.92 at precision between 0.31 and 0.44, giving Cohen’s κ between 0.29 and 0.45. This is the source of the over-prediction noted in Section 4.1. LLaMA3.1-Magpie inverts the trade-off (precision 0.58, recall 0.27), and Command-A is the only model strong on both (0.62 / 0.73), which lifts its κ to 0.60.

F Persona-Matching Agreement Effect on Flipped Posts

Table 10 reports the persona-matching agreement effect for each model, together with the analysis restricted to posts on which the CIS-FEMALE and CIS-MALE prompts produce different predictions. On flipped posts, $\Delta_{\text{match}}^{\text{flip}} = 4 \text{ CorrWin} - 200$ when both quantities are expressed in percentage points. The flipped-post analysis is therefore an inferential reparameterisation of the corresponding-gender win rate, rather than independent evidence.

G Semantic and Topic-Level Analyses

G.1 Semantic Clustering

Independently of the target-group metadata, we embed all 28,448 post texts using sentence-transformers². We L_2 -normalize the embeddings and fit standard k -means with 10 initializations and random seed 20260722. Clustering is performed once on the full corpus, without using model predictions, persona conditions, toxicity annotations, or target-group labels.

²<https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2>

Model	NEUTRAL	CIS-FEMALE	CIS-MALE	LGBT
Aya-23	36.49 $\pm$ 0.60	43.32 $\pm$ 0.61	41.07 $\pm$ 0.63	46.98 $\pm$ 0.64
Aya-Expanse	40.65 $\pm$ 0.60	43.04 $\pm$ 0.61	44.11 $\pm$ 0.60	40.32 $\pm$ 0.60
Command-A	18.66 $\pm$ 0.49	13.85 $\pm$ 0.44	15.62 $\pm$ 0.46	9.59 $\pm$ 0.38
Gemma3-IT	32.49 $\pm$ 0.60	24.98 $\pm$ 0.54	22.78 $\pm$ 0.52	33.55 $\pm$ 0.59
LLaMA3.1-IT	25.25 $\pm$ 0.54	15.46 $\pm$ 0.45	19.35 $\pm$ 0.48	26.51 $\pm$ 0.54
LLaMA3.1-Magpie	7.19 $\pm$ 0.33	4.17 $\pm$ 0.25	7.65 $\pm$ 0.35	27.46 $\pm$ 0.56
Qwen3-base	41.94 $\pm$ 0.61	21.32 $\pm$ 0.52	23.76 $\pm$ 0.54	36.02 $\pm$ 0.59
Qwen3-IT	37.56 $\pm$ 0.58	33.03 $\pm$ 0.57	34.68 $\pm$ 0.57	31.94 $\pm$ 0.58
SahabatAI-IT	34.29 $\pm$ 0.59	36.84 $\pm$ 0.59	36.32 $\pm$ 0.58	36.11 $\pm$ 0.59
SeaLLMs-v3	27.08 $\pm$ 0.53	14.15 $\pm$ 0.43	14.20 $\pm$ 0.43	15.73 $\pm$ 0.45

Table 6: Predicted toxicity rates (%) and bootstrap standard errors by model and prompting condition on posts with untied majority labels ($n = 6,352$; gold rate 15.7%). These values are plotted in Figure 2; full-set results are in Table 7.

Model	NEUTRAL	CIS-FEMALE	CIS-MALE	LGBT
Aya-23	40.46 $\pm$ 0.55	47.43 $\pm$ 0.57	45.08 $\pm$ 0.55	50.33 $\pm$ 0.59
Aya-Expanse	45.93 $\pm$ 0.57	48.35 $\pm$ 0.58	49.43 $\pm$ 0.58	45.14 $\pm$ 0.57
Command-A	22.47 $\pm$ 0.49	16.23 $\pm$ 0.43	18.69 $\pm$ 0.45	11.04 $\pm$ 0.35
Gemma3-IT	37.41 $\pm$ 0.54	29.53 $\pm$ 0.51	27.01 $\pm$ 0.49	38.08 $\pm$ 0.55
LLaMA3.1-IT	29.52 $\pm$ 0.51	18.70 $\pm$ 0.45	23.31 $\pm$ 0.48	30.69 $\pm$ 0.52
LLaMA3.1-Magpie	8.64 $\pm$ 0.33	5.05 $\pm$ 0.25	9.24 $\pm$ 0.33	31.49 $\pm$ 0.54
Qwen3-base	46.03 $\pm$ 0.57	24.65 $\pm$ 0.50	27.39 $\pm$ 0.52	39.53 $\pm$ 0.56
Qwen3-IT	42.49 $\pm$ 0.57	38.04 $\pm$ 0.55	39.91 $\pm$ 0.55	37.03 $\pm$ 0.54
SahabatAI-IT	39.93 $\pm$ 0.56	42.70 $\pm$ 0.57	42.20 $\pm$ 0.56	41.39 $\pm$ 0.56
SeaLLMs-v3	31.41 $\pm$ 0.54	16.64 $\pm$ 0.43	16.69 $\pm$ 0.44	18.42 $\pm$ 0.45

Table 7: Percentage of posts classified as TOXIC under each prompting condition on the full evaluation set ($n = 7,524$), with bootstrap standard errors from 1,000 resamples. Unlike Table 6, these results include the 1,172 posts with tied majority votes.

Model	P_{tox}	R_{tox}	$F1_{\text{tox}}$	$F1_{\text{non}}$	κ
Aya-23	.351	.814	.490	.820	.347
Aya-Expanse	.356	.921	.514	.809	.371
Command-A	.616	.731	.668	.931	.600
Gemma3-IT	.419	.865	.564	.862	.447
LLaMA3.1-IT	.444	.713	.547	.883	.438
LLaMA3.1-Magpie	.582	.266	.365	.918	.296
Qwen3-base	.310	.827	.451	.778	.288
Qwen3-IT	.369	.882	.521	.826	.384
SahabatAI-IT	.414	.903	.568	.856	.449
SeaLLMs-v3	.425	.732	.538	.874	.423

Table 8: Toxic-class precision and recall, per-class F1, and Cohen’s κ against aggregated gold labels (NEUTRAL condition), measured on the evaluation subset with perfect-disagreement ties excluded ($n = 6,352$, gold toxic rate 15.7%). All values independently recomputed from the released predictions.

We use $k = 8$ as an external-validation setting because the dataset contains eight seeded target groups apart from UNKNOWN. Because adjusted Rand index requires a single reference partition, we exclude rather than collapse the 836 posts carrying multiple target-group labels, and also exclude the 2,263 UNKNOWN posts. This leaves 25,349

single-label seeded posts. On this subset, the $k = 8$ assignments obtain an adjusted Rand index of .573 and purity of .808, where purity is the proportion of posts assigned the majority target-group label within each cluster, summed across clusters. Multi-label posts remain members of every applicable group in the separate target-group analyses.

Our primary semantic analysis uses $k = 30$ as a prespecified finer-grained partition; $k = 8$ serves only as the cardinality-matched validation setting. We repeat the complete analysis for $k \in \{8, 12, 16, 20, 24, 30\}$ and obtain the same inferential conclusion across granularities (Appendix G.3).

G.2 Topic-Level Agreement Gaps

To test whether the pooled demographic null masks topic-specific effects, we repeated the item-level agreement analysis within each sufficiently large content subset. We considered the seven dataset-defined target groups and the 27 semantic clusters obtained from the k -means partition ($k = 30$), retaining only subsets with at least 100 evaluation posts. For each subset, we recomputed both the

Model	Agreement and annotator permutation						Crossed-effects model		
	A_F	A_M	Δ	SD_0	Thresh.	p_π	OR	95% CI	p
Aya-23	68.68	67.42	+1.26	1.66	3.26	.476	0.848	[0.724, 0.993]	.040
Aya-Expanse	68.42	67.67	+0.75	2.51	4.92	.781	0.896	[0.568, 1.415]	.639
Command-A	78.50	78.40	+0.11	2.49	4.87	.970	0.885	[0.700, 1.117]	.304
Gemma3-IT	72.01	72.24	-0.22	2.42	4.75	.932	0.929	[0.658, 1.311]	.675
LLaMA3.1-IT	73.47	73.72	-0.25	1.85	3.63	.907	1.026	[0.947, 1.112]	.524
LLaMA3.1-Magpie	75.61	75.97	-0.37	3.41	6.68	.917	0.897	[0.545, 1.479]	.671
Qwen3-base	63.63	63.98	-0.35	2.10	4.11	.879	0.929	[0.687, 1.257]	.633
Qwen3-IT	69.44	68.60	+0.84	2.52	4.94	.762	0.841	[0.559, 1.264]	.405
SahabatAI-IT	71.04	71.71	-0.67	2.66	5.22	.843	0.943	[0.661, 1.346]	.748
SeaLLMs-v3	72.26	72.62	-0.36	2.07	4.06	.875	1.035	[0.962, 1.114]	.360
Pooled	71.31	71.23	+0.07	1.96	3.84	.972	0.917	[0.739, 1.138]	.431

Table 9: Agreement with individual annotations under NEUTRAL prompting. A_F and A_M are percentages and $\Delta = A_F - A_M$ is in percentage points. SD_0 , the detection threshold ($1.96 SD_0$), and p_π come from reassigning gender jointly across the 29 annotators ($B = 20,000$). Odds ratios are from $\text{agree} \sim \text{male} + (1 | \text{post}) + (1 | \text{annotator})$; values below one indicate lower odds of agreement with male annotators. Aya-23 is nominally significant in the parametric model ($p = .040$) but not after correction for ten models, and its permutation gap remains below its threshold.

Model	Flips, n (%)	Corr. win (%)	All posts				Flipped posts only			
			Δ_{match}	SD_0	Thresh.	p_π	$\Delta_{\text{match}}^{\text{flip}}$	SD_0	Thresh.	p_π
Aya-23	477 (6.3)	47.6	-0.53	0.36	0.71	.150	-9.43	6.64	13.01	.157
Aya-Expanse	155 (2.1)	49.5	-0.00	0.18	0.35	.970	-2.15	8.60	16.86	.799
Command-A	225 (3.0)	48.3	-0.20	0.51	1.00	.700	-6.88	15.89	31.15	.676
Gemma3-IT	248 (3.3)	48.6	-0.20	0.35	0.69	.583	-5.48	9.83	19.28	.582
LLaMA3.1-IT	383 (5.1)	51.2	+0.22	0.69	1.35	.765	+4.82	14.45	28.31	.741
LLaMA3.1-Magpie	315 (4.2)	47.8	-0.36	0.85	1.67	.685	-8.83	19.16	37.55	.663
Qwen3-base	410 (5.4)	50.3	+0.08	0.52	1.02	.869	+1.20	9.64	18.90	.906
Qwen3-IT	255 (3.4)	49.6	-0.12	0.41	0.80	.764	-1.79	13.49	26.45	.904
SahabatAI-IT	130 (1.7)	52.6	+0.15	0.15	0.30	.339	+10.37	10.10	19.79	.312
SeaLLMs-v3	146 (1.9)	47.9	-0.15	0.21	0.41	.485	-8.24	11.64	22.81	.487
Pooled	2,744 (3.6)	49.2	-0.11	0.24	0.47	.648	-3.13	6.55	12.84	.643

Table 10: Persona-matching agreement effect in percentage points. $\Delta_{\text{match}} = [A_F^F - A_M^F] - [A_F^M - A_M^M]$ (Section 4.2). Positive values indicate that each gender persona selectively increases agreement with the corresponding annotator group. “Flips” counts posts on which the two gender prompts produce different predictions; “Corr. win” is the corresponding-gender win rate on those posts, with 50% indicating no directional advantage. The right panel recomputes the effect on flipped posts. SD_0 is the standard deviation of the joint annotator-permutation null, detection threshold = $1.96 SD_0$, and $B = 20,000$.

female–male agreement gap and the corresponding annotator-level permutation null using the same procedure as in the main analysis.

Table 11 summarizes the analysis, while Tables 12 and 13 report the complete target-group and semantic-cluster results under all four prompting conditions, together with the topic-specific Δ_{match} tests.

None of the 136 condition–subset agreement gaps reaches $p < .05$ even pointwise (minimum $p = .068$; joint $p_{\text{FWER}} = .72$). The largest absolute gap is 7.80 percentage points for Cluster 15 under NEUTRAL prompting, below its 10.35-point threshold ($p = .127$). Likewise, none of the 34 topic-specific Δ_{match} estimates is significant (minimum $p = .166$; joint $p_{\text{FWER}} = .93$). The largest is

−0.90 points for LGBTQ+ content, below its 1.26-point threshold. Ahmadiyya and Shia are shown descriptively but excluded from inference because they contain fewer than 100 posts; the three sub-threshold semantic clusters are omitted.

G.3 Robustness Across Clustering Granularities

The absence of topic-specific demographic effects is not specific to the choice of $k = 30$. We repeated the complete subset analysis for $k \in \{8, 12, 16, 20, 24, 30\}$.

Figure 7 visualizes every standardized cluster-level agreement gap across clustering granularities together with the corresponding annotator-level significance threshold. Table 14 summarizes the same

Analysis	Tests	Min. p	Max. $ \Delta $	Thresh. _{max}	$p_{\max}$	p_{FWER}
Condition-specific agreement gaps	136	.068	7.80	10.35	.127	.72
Topic-specific Δ_{match}	34	.166	0.90	1.26	.166	.93

Table 11: Content-specific demographic agreement across prompting conditions. The first row comprises female–male agreement gaps within each of four conditions and 34 tested content subsets ($4 \times 34 = 136$); the second comprises the paired persona-matching agreement effect within each subset. Min. p is the smallest pointwise permutation value. Thresh._{max} and $p_{\max}$ belong to the cell giving Max. $|\Delta|$, whereas p_{FWER} is the omnibus maximum-statistic test over both content partitions. Effects and thresholds are in percentage points. The 7.80-point maximum is Neutral Cluster 15—the same topic-level observation reported in Table 17—whose own threshold is 10.35; the minimum $p = .068$ comes from a different cell (CIS-FEMALE, Cluster 9).

Subset	n	NEUTRAL	CIS-MALE	CIS-FEMALE	LGBT	Δ_{match}
Ahmadiyya [†]	35	-4.22 / 5.74 / .15	-5.11 / 5.77 / .08	-4.76 / 5.81 / .11	-5.85 / 6.46 / .08	+0.35 / 0.66 / .32
Disability	1,582	-2.73 / 4.12 / .21	-3.09 / 4.13 / .15	-2.93 / 3.92 / .15	-2.19 / 3.81 / .29	+0.16 / 0.56 / .59
Jewish	2,424	-1.39 / 4.79 / .68	-1.29 / 4.48 / .64	-1.69 / 4.52 / .51	-1.69 / 4.45 / .55	-0.40 / 1.10 / .50
Christian	702	+1.47 / 4.14 / .53	+1.72 / 4.64 / .54	+1.90 / 4.64 / .48	+1.17 / 3.65 / .62	+0.18 / 0.55 / .56
LGBTQ+	572	+6.25 / 8.65 / .15	+5.87 / 9.18 / .21	+4.97 / 8.39 / .25	+6.84 / 11.48 / .27	-0.90 / 1.26 / .17
Rohingya	134	+2.27 / 5.46 / .43	+1.18 / 4.25 / .62	+1.21 / 4.07 / .57	+0.72 / 3.82 / .73	+0.03 / 2.29 / .94
Shia [†]	85	-5.47 / 10.43 / .32	-6.29 / 10.04 / .22	-5.48 / 9.13 / .25	-6.26 / 10.58 / .25	+0.81 / 1.60 / .33
Chinese-Indonesian	1,380	-3.03 / 7.02 / .45	-2.20 / 6.53 / .57	-2.13 / 6.57 / .57	-2.13 / 6.52 / .59	+0.06 / 0.56 / .87
UNKNOWN	896	+0.61 / 3.30 / .73	+1.74 / 3.96 / .41	+1.75 / 4.09 / .42	+1.09 / 3.29 / .53	+0.01 / 0.47 / .98

Table 12: Item-level agreement by prompt condition for dataset-defined target groups. Each analysis cell reports observed effect / detection threshold / pointwise p . For the four prompt conditions, the effect is the female–male agreement gap; Δ_{match} is the paired CIS-FEMALE–CIS-MALE difference-in-differences. Effects and thresholds are in percentage points; the detection threshold is 1.96 SD_0 , and p is pointwise under one joint annotator-gender permutation ($B = 20,000$). No pointwise test reaches .05; across both content partitions, familywise omnibus tests give $p = .72$ for the 136 condition–subset gaps and $p = .93$ for the 34 topic-specific persona-matching agreement effects. [†]Subsets below 100 posts are descriptive only.

analysis numerically.

Across the six clustering granularities, 102 semantic clusters were tested. Together with the seven dataset-defined target groups, this gives 109 tested content subsets in total. Only one of the 109 reached nominal significance ($p = .04$ at $k = 16$), Cluster 8 of the $k = 16$ partition—a religious-discourse cluster, consistent with chance under multiple testing. No clustering granularity produced a significant familywise maximum gap ($p \geq .52$) or dispersion statistic ($p \geq .35$).

Subset	n	NEUTRAL	CIS-MALE	CIS-FEMALE	LGBT	Δ_{match}
Cluster 0	198	+1.54 / 5.59 / .60	+1.24 / 6.35 / .71	+1.24 / 6.48 / .72	+0.95 / 6.53 / .78	+0.01 / 0.63 / 1.00
Cluster 1	149	+0.64 / 7.04 / .86	-0.59 / 7.43 / .88	-0.85 / 7.56 / .84	+0.75 / 6.07 / .82	-0.26 / 1.17 / .66
Cluster 2	259	-4.05 / 6.80 / .27	-4.54 / 5.68 / .12	-4.05 / 5.75 / .17	-2.02 / 6.73 / .59	+0.48 / 1.37 / .50
Cluster 3	303	-0.06 / 5.85 / .98	+0.08 / 6.02 / .98	+0.21 / 5.66 / .95	-0.34 / 5.85 / .91	+0.13 / 0.51 / .62
Cluster 4	194	-4.01 / 9.59 / .46	-2.70 / 8.48 / .57	-2.46 / 8.05 / .59	-2.88 / 8.23 / .58	+0.25 / 1.13 / .69
Cluster 5	194	-3.09 / 7.78 / .48	-3.04 / 10.05 / .59	-3.11 / 9.78 / .57	-2.94 / 8.35 / .52	-0.06 / 0.87 / .87
Cluster 6	327	-0.39 / 5.62 / .89	-2.10 / 4.73 / .40	-2.37 / 4.33 / .30	-0.14 / 6.34 / .96	-0.27 / 0.94 / .58
Cluster 7	279	-1.04 / 4.77 / .67	+0.92 / 5.28 / .74	+0.37 / 6.04 / .91	+0.28 / 5.49 / .92	-0.54 / 1.42 / .48
Cluster 8	203	+1.32 / 4.39 / .56	+1.88 / 4.82 / .46	+1.56 / 4.71 / .53	+1.23 / 4.21 / .58	-0.32 / 0.71 / .38
Cluster 9	283	-3.46 / 4.20 / .11	-3.43 / 3.91 / .09	-3.33 / 3.58 / .07	-4.11 / 5.04 / .11	+0.10 / 0.78 / .81
Cluster 10	228	-0.18 / 5.91 / .95	-0.22 / 5.79 / .94	-0.24 / 5.86 / .94	-0.19 / 6.28 / .95	-0.02 / 1.15 / .98
Cluster 12	213	+0.27 / 4.10 / .90	+1.58 / 4.29 / .48	+1.38 / 4.46 / .56	+1.54 / 3.75 / .44	-0.20 / 1.19 / .74
Cluster 13	299	+3.25 / 6.54 / .35	+2.39 / 6.24 / .46	+2.13 / 5.64 / .46	+2.08 / 5.82 / .50	-0.26 / 1.43 / .72
Cluster 14	349	+1.39 / 5.15 / .61	+2.34 / 4.72 / .34	+2.88 / 4.97 / .27	+2.47 / 4.50 / .29	+0.54 / 1.13 / .37
Cluster 15	242	-7.80 / 10.35 / .13	-4.73 / 8.39 / .30	-4.35 / 8.27 / .34	-4.24 / 7.26 / .28	+0.39 / 0.70 / .30
Cluster 16	170	-0.52 / 7.11 / .89	-0.73 / 6.85 / .84	-0.48 / 6.40 / .89	-1.11 / 7.31 / .79	+0.26 / 1.51 / .75
Cluster 17	398	-1.58 / 4.56 / .52	-1.62 / 6.35 / .63	-1.86 / 6.59 / .59	-1.72 / 6.86 / .64	-0.25 / 1.06 / .65
Cluster 18	388	+1.64 / 12.37 / .80	+2.06 / 9.69 / .69	+1.48 / 8.32 / .73	+0.94 / 11.47 / .87	-0.58 / 1.72 / .52
Cluster 19	273	+1.28 / 5.09 / .64	+1.15 / 4.64 / .64	+1.18 / 4.49 / .62	+1.25 / 4.31 / .58	+0.03 / 0.42 / .91
Cluster 20	231	-2.14 / 5.68 / .46	-2.09 / 5.51 / .45	-2.19 / 5.58 / .43	-1.97 / 5.56 / .49	-0.10 / 0.63 / .73
Cluster 21	549	-2.49 / 5.34 / .39	-2.10 / 6.36 / .56	-2.65 / 7.13 / .51	-2.47 / 6.56 / .51	-0.55 / 1.54 / .50
Cluster 22	405	+5.74 / 10.72 / .31	+5.59 / 10.99 / .35	+4.89 / 10.03 / .37	+6.51 / 13.65 / .42	-0.70 / 1.44 / .38
Cluster 23	316	+2.22 / 7.62 / .67	+2.97 / 6.85 / .44	+3.05 / 7.11 / .44	+1.85 / 4.43 / .45	+0.08 / 0.54 / .81
Cluster 25	275	+3.65 / 6.07 / .25	+2.93 / 5.51 / .31	+2.80 / 5.37 / .32	+2.68 / 5.57 / .35	-0.12 / 0.56 / .65
Cluster 27	139	-3.19 / 7.64 / .42	+0.10 / 6.29 / .98	+0.31 / 6.90 / .93	-0.84 / 7.04 / .82	+0.22 / 1.23 / .77
Cluster 28	206	-5.27 / 7.26 / .16	-5.93 / 8.09 / .18	-6.22 / 7.96 / .15	-5.43 / 7.22 / .15	-0.29 / 0.82 / .51
Cluster 29	235	-1.26 / 7.31 / .73	-0.82 / 6.74 / .81	-1.09 / 6.68 / .75	-0.66 / 6.61 / .85	-0.27 / 0.65 / .42

Table 13: Item-level agreement by prompt condition for the tested $k = 30$ semantic clusters. Each analysis cell reports observed effect / detection threshold / pointwise p . For the four prompt conditions, the effect is the female–male agreement gap; Δ_{match} is the paired CIS-FEMALE–CIS-MALE difference-in-differences. Effects and thresholds are in percentage points; the detection threshold is 1.96SD_0 , and p is pointwise under one joint annotator-gender permutation ($B = 20,000$). No pointwise test reaches .05; across both content partitions, familywise omnibus tests give $p = .72$ for the 136 condition–subset gaps and $p = .93$ for the 34 topic-specific persona-matching agreement effects. The three clusters below the 100-post threshold are omitted.

k	Subsets	Max $ \Delta $	SD_0	Min p	p_{fw}	p_{disp}
8	7	3.1	2.3	.25	.74	.51
12	11	3.7	2.5	.33	.82	.64
16	15	6.0	2.6	.04	.52	.35
20	19	6.1	2.8	.06	.72	.41
24	23	6.1	2.9	.06	.76	.63
30	27	7.8	3.1	.10	.53	.56

Table 14: Robustness across semantic clustering granularities. For each value of k , we report the number of tested clusters, the largest absolute agreement gap (percentage points), the median null standard deviation, the smallest per-cluster p -value, and the familywise maximum-gap and dispersion tests. The single nominally significant cluster at $k = 16$ does not survive either familywise test. These tests are computed separately for each granularity and therefore differ from the omnibus test across prompting conditions and both content partitions ($p_{\text{FWER}} = .72$; Table 11).

gold (majority vote)		Aya-23		Aya-Expanse		Command-A					
		non-tox.	72% 3,848	28% 1,505	69% 3,691	31% 1,662	92% 4,898	8% 455			
		toxic	19% 186	81% 813	8% 79	92% 920	27% 269	73% 730			
		Gemma3-IT		LLaMA3.1-IT		LLaMA3.1-Magpie					
		non-tox.	78% 4,153	22% 1,200	83% 4,461	17% 892	96% 5,162	4% 191			
		toxic	14% 135	86% 864	29% 287	71% 712	73% 733	27% 266			
		Qwen3-base		Qwen3-IT		SahabatAI-IT					
		non-tox.	66% 3,515	34% 1,838	72% 3,848	28% 1,505	76% 4,077	24% 1,276			
		toxic	17% 173	83% 826	12% 118	88% 881	10% 97	90% 902			
		SeaLLMs-v3		non-tox.		toxic		non-tox.		toxic	
		non-tox.	82% 4,364	18% 989							
		toxic	27% 268	73% 731							
		non-tox.		toxic							
predicted											

Figure 6: Per-model confusion matrices, NEUTRAL condition, on the evaluation subset with perfect-disagreement ties excluded ($n = 6,352$; gold toxic rate 15.7%)—the same basis as Table 8. Cells are shaded by row-normalised proportion and labelled with that proportion and the raw count. Diagonal entries are per-class recall. The upper-left cell is non-toxic recall, so lower values indicate over-flagging.

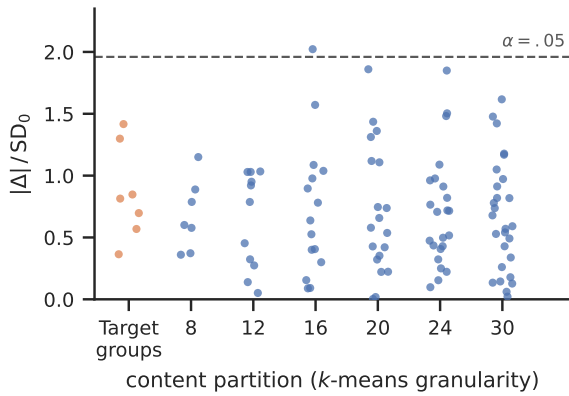


Figure 7: Standardized agreement gaps, $|\Delta|/SD_0$, for all 109 tested content subsets: the seven target groups (left) and the 102 semantic clusters across k -means granularities, where SD_0 is the annotator-level permutation-null standard deviation. The dashed line marks the two-sided 5% threshold; a single cluster at $k = 16$ exceeds it, consistent with chance under multiple testing.

H By-Post Random-Slope Sensitivity

To test whether the pooled RQ2 result depends on assuming a constant gender effect across posts, we refit Eq. 1 with a correlated by-post random intercept and gender slope, $(1 + \text{male} \mid \text{post})$, while retaining the annotator random intercept. The model converged without warnings and was not singular. The male fixed effect remained small and non-significant ($\beta = -0.06$, $\text{SE} = 0.10$, $\text{OR} = 0.94$, 95% CI [0.78, 1.14], $p = .55$), compared with $\beta = -0.09$, $\text{OR} = 0.92$, 95% CI [0.74, 1.14], $p = .43$ in the primary random-intercept model. The by-post model estimated $\text{SD} = 2.50$ for the post intercept, $\text{SD} = 1.46$ for the gender slope, and an intercept-slope correlation of -0.31 . Because 4,789 of 7,524 posts (64%) contain exactly two distinct annotators, we do not interpret the slope variance as substantive evidence of post-level gender heterogeneity; this analysis is used only to assess the stability of the pooled fixed gender effect.

I Topic-Specific Gender Slopes

Finally, we examined whether the demographic effect varies systematically across semantic clusters in the mixed-effects model. Table 15 compares two extensions of Eq. 1 at each granularity $k \in \{8, 12, 16, 20, 24, 30\}$; the $k = 30$ row corresponds to the main semantic partition. Allowing the gender effect to vary across clusters produces a significant random slope at every granularity when annotator-by-cluster variation is omitted. Once annotator-specific topic variation is included, the slope variance lies on the zero boundary for five of the six granularities (fits flagged as singular). At $k = 20$, where the naive slope variance is largest, the controlled fit is non-singular and the remaining slope is non-significant ($\chi^2 = 0.21$, $p = .65$). The fixed gender effect is null across all twelve fits ($\text{OR} = 0.91\text{--}0.94$, all $p > .48$). The apparent topic-specific demographic effect is therefore explained by individual annotators responding differently across topics rather than by a stable interaction between topic and annotator gender.

J Agreement Metrics and Aggregation Robustness

Section 4.3 reports Cohen’s κ under the primary aggregation rule, where annotations are aggregated separately within each gender using majority vote and ties are resolved as TOXIC. To verify that the conclusions do not depend on these choices, we

k	Topic slope only			+ annotator $\times$ cluster		
	$\text{SD}_{\text{male} c}$	χ^2	p	$\text{SD}_{\text{male} c}$	χ^2	p
8	.066	15.05	< .001	.000	0.00	1.00
12	.080	25.61	< .001	.000	0.00	.997
16	.128	102.78	< .001	.000	0.00	.998
20	.161	165.67	< .001	.070	0.21	.650
24	.149	122.97	< .001	.000	0.00	1.00
30	.138	97.95	< .001	.000	0.00	.997

Table 15: Topic-varying gender slopes across clustering granularities. Both specifications extend Eq. 1; likelihood-ratio tests compare each with its counterpart without the cluster-level gender slope ($\text{df} = 2$ and 1). Boundary tests on variance components are conservative. The topic slope is significant at every k but not once annotator-by-cluster variation is modeled.

repeat the analysis across five agreement statistics and four aggregation conventions. Every combination is evaluated against the same annotator-level permutation null.

The agreement statistics are raw agreement, Cohen’s κ , Gwet’s AC1, the log odds ratio ($\ln\text{OR}$), and a discordance statistic computed only on posts where the female and male panels disagree. The aggregation conventions are: (i) ties resolved as TOXIC, (ii) ties resolved as NON-TOXIC, (iii) posts containing tied panels removed, and (iv) ties resolved as TOXIC after restricting to posts where the two panels have equal size.

Across all 20 pooled specifications, the observed gap remains within the central 95% of the corresponding permutation null (Table 16). None of the 200 underlying model-specific tests reaches $p < .05$ (smallest uncorrected $p = .065$). Changing the aggregation rule has a larger effect than changing the agreement statistic: prevalence-robust statistics yield somewhat larger standardized gaps under the published aggregation rule, but these remain non-significant, while changing the tie rule substantially changes the observed gap. This supports the interpretation that the apparent group-level differences arise primarily from calibration and label aggregation rather than stable demographic agreement.

K Permutation Detection Thresholds

Several of our findings are nulls, so for each estimand we report an approximate two-sided 95% detection threshold under the annotator-level permutation null, computed as 1.96SD_0 (Table 17). This quantity is a significance-based detection threshold

Stat.	Convention	Obs.	SD ₀	$\frac{ \text{Obs.} }{\text{SD}_0}$	p
Raw	ties→tox	−.014	.016	0.90	.35
	ties→non	−.007	.018	0.38	.69
	ties dropped	−.010	.017	0.57	.57
	size-matched	−.017	.022	0.77	.54
κ	ties→tox	−.021	.057	0.36	.72
	ties→non	−.020	.066	0.30	.76
	ties dropped	−.028	.069	0.40	.70
	size-matched	−.035	.095	0.37	.76
AC1	ties→tox	−.032	.025	1.29	.16
	ties→non	−.009	.025	0.36	.65
	ties dropped	−.014	.025	0.56	.51
	size-matched	−.032	.033	0.97	.35
lnOR	ties→tox	−.263	.198	1.33	.15
	ties→non	−.094	.250	0.38	.65
	ties dropped	−.162	.255	0.64	.47
	size-matched	−.328	.289	1.13	.28
Disc.	ties→tox	−.081	.063	1.29	.16
	ties→non	−.024	.067	0.36	.68
	ties dropped	−.047	.075	0.63	.48
	size-matched	−.094	.092	1.02	.36

Table 16: Robustness to agreement statistic and aggregation rule. “Obs.” is the pooled female–male agreement gap across the ten models, SD₀ is the standard deviation of the annotator-level permutation null, and p is two-sided. Every pooled specification and every underlying model-specific test is non-significant.

rather than a minimum detectable effect from a statistical power calculation. The effective resolution is determined primarily by the 29 annotators rather than the much larger number of post-annotator observations.

Every observed effect falls below the detection threshold of the analysis that produced it. Resolution varies by more than an order of magnitude across estimands: the persona-matching analysis has a detection threshold of about 0.5 percentage points, whereas topic-level comparisons on small subsets have thresholds above ten percentage points. Together with the permutation analyses, this indicates that the absence of significant demographic agreement is consistent with the statistical resolution of the available annotator pool rather than evidence that demographic effects cannot exist.

Estimand	SD ₀	Thresh.	Obs.
$\Delta\kappa$, pooled	.057	.111	.021
$\Delta\kappa$, per model	.029–.085	.057–.167	.039
Group-label gap, pooled	1.58	3.09	1.42
Group-label gap, per model	1.40–4.99	2.74–9.78	2.64
Annotation gap, pooled cond.	1.94–2.00	3.80–3.92	0.34
Annotation gap, model×cond.	1.66–3.99	3.26–7.83	1.26
Annotation gap, by topic [†]	3.08	6.03	7.80
Δ_{match} , pooled	0.24	0.47	0.11
Δ_{match} , per model	0.15–0.85	0.30–1.67	0.53

Table 17: Permutation-null detection thresholds; all effects except $\Delta\kappa$ are in percentage points. Threshold = 1.96 SD₀; ranges span prompting conditions or models as indicated, and Obs. is the largest absolute estimate in that row. [†] The by-topic row is the annotation-level gap with models pooled; SD₀ is the median across the 136 prompting-condition–content-subset tests. Its 7.80-point maximum is Neutral Cluster 15, whose own threshold is 10.35 points (Table 11).