Robust Persona-Aware Toxicity Detection with Prompt Optimization and a Learned Ensemble of Prompts

Anonymous ACL submission

Abstract

Judgment of toxic language is highly subjective, as well as dependent on contextual factors such as the identity of the target, the audience, or the sociocultural environment of use. Prior work shows that toxicity perception can differ for judges from different demographic groups, such as race or gender. We study whether LLMs can better reflect these differences through persona-based prompting. We investigate a range of prompting strategies for toxicity judgements from several LLMs, using demographically distinct personas. We find that prompts that instruct the LLM to adopt a specific *persona* often improve performance, but not consistently across the board. No single prompting strategy achieves reliable performance across conditions. We therefore investigate ensembles of multiple prompts for the same LLM. We find that a learned SVM achieves the best performance, surpassing majority voting. Error analysis of 16 prompt-response configurations provides insights into the specific benefits of the learned ensemble.

1 Introduction

The identification of toxic language is intrinsically subjective, influenced by human raters' perspectives, identities, beliefs, social priors, and the social groups targeted in the text (Mostafazadeh Davani et al., 2024; Sorensen et al., 2025a; Fleisig et al., 2023). As a consequence, the same text can be judged differently by different raters. In fact, previous work suggests that toxicity should be annotated by members of the group that the toxic language targets (Mostafazadeh Davani et al., 2024; Fleisig et al., 2023). Such findings have motivated *persona-aware* or *perspective-conditioned* toxicity detection, where systems are evaluated (and potentially deployed) with explicit conditioning variables to approximate the judgments of different groups (Mostafazadeh Davani et al., 2024; Sorensen et al., 2025a). To replace expensive human raters, there has been increasing use of LLMs as judges (Zheng et al., 2023; Gu et al., 2024). In light of the subjective nature of toxicity, this paper explores the reliability of **persona-conditioned LLM judgments** for toxicity detection.

In this work, we study the robustness of persona-conditioned LLM judgments of toxicity across prompt designs and LLMs. We conduct a systematic comparison of four prompt variants using personas defined by race, gender and ethnicity: (i) a definition-based toxicity prompt, (ii) perspective-taking persona prompts (Xu et al., 2024), (iii) value-profile conditioning (Sorensen et al., 2025a), which provides a textual description of human values/beliefs based on example toxicity judgments by human, and (iv) an automatically optimized persona prompt obtained via TextGrad-style textual "differentiation" (Yuksekgonul et al., 2024). Across multiple open-source LLMs, **no single prompting strategy consistently performs best,** including reasoning enhanced models: performance varies substantially by persona and by model.

To characterize this consistency challenge, we further analyze **prompt variability** by treating the four prompt-dependent predictions as an ensemble. We show that disagreements in responses from different prompting methods is common and persona-dependent. Unsurprisingly, unanimous agreement tends to be relatively correct, but errors rise sharply for cases with any disagreement, with majority voting for 2-2 ties and singleton disagreements being incorrect nearly equally often, given random tie-breaking. We speculated that majority voting might obscure informative patterns in the ties and singletons that could be learned.

Motivated by the broader success of ensembling for improving reliability when individual systems have heterogeneous error patterns (Jiang et al., 2023; Yang et al., 2023; Ai et al., 2025), we evaluate multiple ensembling techniques to combine

the four prompting approaches. While ensembles improve performance, we observed patterns where a single prompting method remains preferable for a particular persona or base model. Thus we develop a lightweight **prompt-ensembler** that uses a support vector machine (SVM) (Cortes and Vapnik, 1995) to learn how to map a 4-bit vector of prompt predictions to a final label. While the input space is small, the prompt-ensembler provides a data-driven learned classifier that yields the best performance.

Contributions. (i) We provide a systematic comparison of persona-aware toxicity detection via different prompting methods across multiple open-source LLMs, showing substantial prompt sensitivity and persona-dependent performance variation. (ii) We introduce the use of a prompt-optimization method for persona prompting (TextGrad (Yuksekgonul et al., 2024)), and show it is competitive with value-profile conditioning (Sorensen et al., 2025a) while exhibiting complementary errors. (iii) We conduct an error analysis across prompting methods to analyze their complementary weaknesses. (iv) We develop an SVM prompt-ensembler that improves robustness over individual prompts and majority voting. In sum, we show that the subjectivity of toxicity judgments can be modeled through a learned ensemble of persona-based prompts.

2 Related Work

In this section, we review offensive language detection and biases of LLMs, subjectivity in toxicity judgments, and LLMs as judges. We conclude with a discussion of work on ensembling LLMs.

Toxicity and offensive language detection. A large body of work studies automatic detection of toxic, hateful, abusive, or offensive language in online text. Common benchmarks include Twitter- and Wikipedia-based corpora (Davidson et al., 2017; Founta et al., 2018; Wulczyn et al., 2017), and shared tasks such as OffensEval (Zampieri et al., 2019). Later datasets include explanations (Mathew et al., 2021), implicit hate (ElSherief et al., 2021), and spurious correlations involving minority groups (Hartvigsen et al., 2022). These datasets assign a single label and therefore do not capture differences in toxicity perception across raters, or sociodemographic factors as conditioning variables.

Bias and unintended correlations in toxicity models. Toxicity detectors are known to rely on surface patterns (e.g., profanity) and to correlate with specific identity mentions, which can harm minority groups through over-flagging (Dixon et al., 2018; Borkan et al., 2019). Dialect variation further amplifies this issue: demographic-aligned language features (e.g., African-American English) lead to non-normative judgments (Blodgett et al., 2016). Further, insensitivity to dialect has been shown to induce racial bias in hate-speech annotation and downstream classifiers (Sap et al., 2019). These findings motivate studying toxicity detection under explicit perspective conditioning rather than treating toxicity as purely lexical.

Subjectivity, perspective variation, and disagreement in toxicity. Annotation research increasingly argues that disagreement can be systematic rather than mere noise (Aroyo and Welty, 2015; Fleisig et al., 2024). In abusive-language research, dataset construction, annotation guidelines, and annotator pools vary substantially, with consequences for both fairness and performance (Vidgen and Derczynski, 2020). Toxicity judgments have been shown to vary with annotator identities and beliefs (Sap et al., 2022). Recent work extends this direction through cross-cultural offensiveness labels (Mostafazadeh Davani et al., 2024) and value profiles that represent human variation beyond demographics (Sorensen et al., 2025a). Our work is aligned with this motivation, but focuses on persona-conditioned labels and robustness across prompts and LLMs, rather than modeling full within-persona disagreement distributions.

LLMs as toxicity classifiers and judges under prompting. LLMs have been used as zero-/few-shot toxicity classifiers (Atil et al., 2026) and for scalable annotation of subjective NLP tasks (Zheng et al., 2023; Gu et al., 2024). In toxicity detection, they can align reasonably well with crowd labels, but remain sensitive to prompt wording, task definitions, and output format (Li et al., 2023; Mishra and Chatterjee, 2023; Roy et al., 2023; Koh et al., 2024). In particular, Koh et al. (2024) show that prompt design choices strongly affect LLM toxicity judgments, motivating our focus on prompt variability in persona-conditioned settings.

Aggregation for reliability. Aggregation is a common way to improve LLM reliability under prompt sensitivity and nondeterminism. Self-consistency aggregates multiple sampled outputs (Wang et al., 2022), while other LLM ensembling frameworks balance complementary strengths

(Jiang et al., 2023; Ai et al., 2025). Multi-LLM methods such as MUSE aggregate across distinct models for uncertainty calibration (Kruse et al., 2025). Our setting is distinct: we aggregate multiple prompting strategies *within the same base LLM* to reduce prompt- and persona-induced variability and we learn a lightweight prompt-ensembler rather than relying only on voting.

3 Methods

Our goal is to leverage persona-aware toxicity classification, thus for each LLM, we compare multiple prompting strategies that condition on a sociodemographic persona (or unconditioned), and then study whether their predictions can be combined to improve performance. This setup allows us to analyze two related questions: (i) how sensitive persona-conditioned toxicity predictions are to prompt design, and (ii) whether disagreement across prompts can be exploited through ensembling methods.

3.1 Prompting Methods

We evaluate four prompting variants for each model-persona pair: no persona condition, a simple persona condition, adding value profiles (Sorensen et al., 2025a) for each persona, and a prompt-optimizer applied to persona-conditioned prompts. All prompts are shown in Table 7 in the Appendix.

Default prompt. As a non-persona baseline, we use a simple toxicity-classification prompt that provides a definition of offensiveness, with no persona.

Persona prompt. To test explicit perspective conditioning, we use a perspective-taking prompt (Xu et al., 2024) that asks the model to judge offensiveness from the perspective of a given persona.

Value-profile prompt. We also experiment with *Value Profiles* (Sorensen et al., 2025b), which describe in natural language the preferences and sensitivities associated with a social group. Instead of inserting raw in-context examples, this method summarizes group-specific tendencies into a compact profile. Following (Sorensen et al., 2025b), we generate the profiles with Gemini 1.5 Pro. For each persona, we select 7 offensive and 7 non-offensive examples from posts with annotator disagreement and use them to construct the profile.

Prompt Optimization To strengthen persona conditioning, we optimize the persona prompt separately for each model-persona pair using TextGrad (Yuksekgonul et al., 2025; Yüksekgönül et al., 2024), which iteratively revises prompts using natural-language feedback from an optimizer LLM. This searches for prompt formulations that elicit more accurate persona-conditioned judgments. For each model-persona pair, we use 100 training and 100 validation examples, optimize for classification accuracy, and use GPT-4o as the optimizer. This produces one optimized persona prompt per model-persona pair. To our knowledge, prior work has not explored automated prompt optimization for persona-aware toxicity classification.

3.2 Aggregation Setup

We also study whether predictions from different prompts can be combined. For a fixed LLM and persona, we combine the binary predictions produced by the four prompting variants above. This differs from prior work that ensembles multiple LLMs; here, the aim is to reduce *prompt-induced variability* for persona-aware classification.

For each persona, we split the data into 20% train, 10% validation, and 70% test. The training split is used to fit aggregation methods, and the validation split for SVM hyperparameter selection. All reported results, including individual prompting methods, are evaluated on the same test split.

3.3 Aggregation Baselines

We compare against three aggregation baselines.

Accuracy-based Weighted Majority Voting. Our first baseline follows (Yang et al., 2023) and assigns each prompt variant a weight based on its training accuracy.

Theoretical Optimal Weighted Majority Voting. Our second baseline uses the weighting scheme proposed by Ai et al. (2025), where the weight for method m is

$$\log \frac{acc_m}{1 - acc_m}.$$

As above, weights are estimated from the training split and then used to combine prompt predictions.

Best Unweighted Majority Voting. As an upper bound for unweighted voting, we evaluate all subsets of the four prompting methods and report the best-performing subset under simple majority voting. Because this method selects the best subset using test performance, it should be interpreted as an oracle comparison rather than a deployable baseline. In case of ties, we predict *offensive*.

Persona	# of Examples	Offensiveness %
Asian Woman	2829	51.11
Asian Man	4468	36.59
Black Woman	4414	54.05
White Woman	32125	54.09
White Man	31964	56.65
Hispanic Woman	4109	65.40
Hispanic Man	3007	50.05
Native Man	471	40.51

Table 1: Dataset statistics with number of examples and percentage of offensiveness labels.

Proposed Learned Ensembler For each example, we form a 4-dimensional vector of binary predictions from the default, persona, value-profile, and optimized-persona prompts. This vector is the input to a learned classifier.

We use an SVM with an RBF kernel (denoted SVM in the paper), and compare it to a linear SVM. We chose the RBF kernel because the signal appears to lie in specific patterns of agreement and disagreement across prompts rather than in prompt accuracies alone. A nonlinear classifier can capture these interactions. This is especially useful in our setting, where the patterns of disagreement per method matters. The SVM is trained on the same training split as the other aggregation methods, with hyperparameters selected on the validation split. We use the default scikit-learn settings.

4 Dataset and Models

4.1 Dataset

We experiment with the Social Bias Frames dataset (Sap et al., 2020), which has 44k unique social media posts with annotations for offensiveness, target group etc. Also, for each post, they have multiple annotations with demographic information of the annotator. In this work, we focus on the combination of gender and race (we refer to each gender and race combination as a "persona."). We choose the personas that have a reasonable (> 400) number of examples. Table 1 gives the statistics about the dataset for each persona.

4.2 Models

We experiment with Llama3.1 70b/8b (Dubey et al., 2024), and Qwen2.5 14b/32b (Qwen et al., 2025) models. Additionally, we analyze the effect of reasoning-enhanced models. For this, we use R1-distilled models (e.g. R1 Distilled Qwen 14b/32b and R1 Distilled Llama 8b/70b) (Guo et al., 2025).

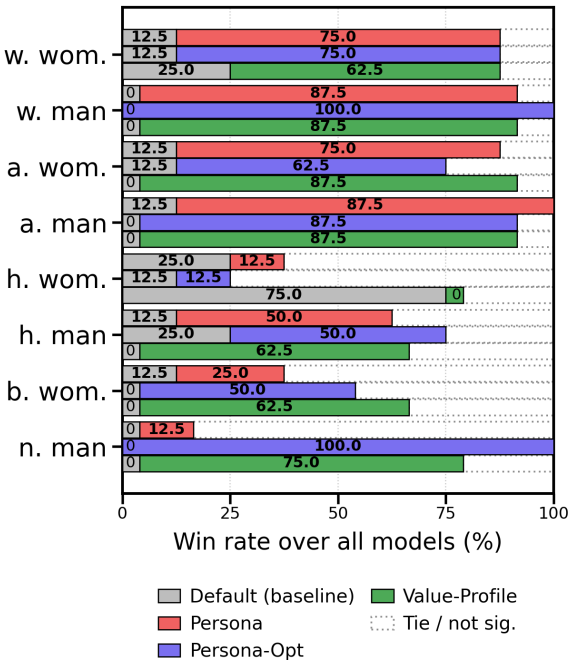

Figure 1: Prompting comparisons with **default prompting** as the baseline.

5 Results

We compare methods using McNemar's test (McNemar, 1947), a paired nonparametric test for two classifiers evaluated on the same examples. For each model–persona setting, we apply McNemar to a pair's correctness indicators and count a *significant win* when the null of equal error rates is rejected (at α=0.05). Table 2 summarizes these wins across all 64 settings (8 models $\times$ 8 personas). Table 5 reports bootstrap confidence intervals for aggregated accuracy by disagreement of prompts. We analyze (i) persona-based prompting effects (Fig. 1), (ii) ensembling improvements over the strongest voting baseline (Fig. 2), and (iii) an SVM ablation study (Fig. 3). Detailed per-model/per-persona results and comparisons to the baselines are provided in the Appendix.

5.1 Reasoning Models

We compare each base model against its reasoning-enhanced counterpart under the default prompt. Overall, the effect of explicit reasoning is mixed and depends on the underlying LLM (see Appendix Table 16). This supports the broader observation that improvements from reasoning-oriented training do not reliably transfer to socially grounded judgments such as toxicity detection, motivating perspective-aware prompting and robustness-oriented aggregation.

5.2 Persona-based Prompt Effects

Figure 1 compares three persona-conditioning strategies against the default prompt, aggregated over models; full model-by-persona results are in the Appendix (Tables 8, 9, and 11). Overall, persona conditioning improves over the default prompt, but the gains vary substantially across personas and LLMs. This variability suggests that different prompting strategies capture complementary aspects of persona-conditioned toxicity judgments.

Persona Prompt. The persona prompt adds demographic context explicitly (e.g., "you are a *X* evaluating offensiveness"). It improves performance in many model-persona settings relative to the default prompt (Figure 1), but inconsistently. For some personas it helps little, and for *hispanic woman* it can hurt performance. Manual analysis shows a recurring type of failure, where persona conditioning raises the rate of *non-offensive*.

Persona optimized Prompt The persona-optimized prompt adapts the persona instruction to each model-persona pair. Compared to the base persona prompt, it usually yields additional gains (Appendix Tables 9, 11), showing that small changes in phrasing can affect persona conditioning. However, it does not remove all failure cases: *hispanic woman* remains challenging,

Value Profile Prompt The value-profile prompt conditions the model through a structured description of values and norms rather than demographic identity alone. Overall, it is one of the strongest single-prompt methods and is often comparable to persona-optimized prompting (Figure 1; Appendix Tables 8, 10). This suggests that persona effects can be elicited through explicit normative framing as well as demographic labels. However, similar to the optimized persona prompt, it can backfire for *hispanic woman*.

Overall, the methods rank as **Optimized Prompt = Value Profile > Persona Prompt > default Prompt**

5.3 Ensembling Methods

Figure 2 compares ensembling methods against the strongest voting baseline, Theoretical Optimal Weighted Majority Voting. Table 2 reports significant win counts across all 64 model–persona settings. We further analyze when ensembling helps using disagreement-bucket confidence intervals (Table 5) and persona-specific disagreement errors (Table 6). Additional comparisons appear in Appendix Figures 6–8.

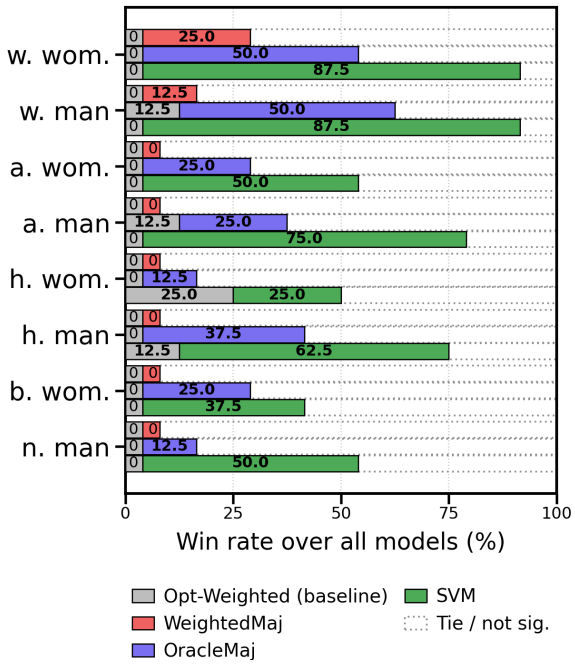

Figure 2: Ensembling comparisons with **Theoretical Optimal Weighted Majority** as the baseline.

Accuracy-based Weighted Majority Voting This method assigns larger weights to prompts that are more accurate on training data. It usually improves over single prompts, but its gains vary across models and personas. This suggests that prompt reliability shifts across settings, and that a fixed weighting scheme is not robust.

Best Unweighted Majority Voting This method is an *upper bound* on unweighted voting because it selects the best prompt subset using test performance. As expected, it is often competitive, but it is not a deployable method. We therefore treat it as a diagnostic ceiling rather than a primary baseline. Additional comparisons are in Appendix Figure 7.

Theoretical Optimal Weighted Majority Voting This is a strong fixed-rule baseline with theoretically motivated weights. It performs similarly to accuracy-based weighted voting and improves over many single-prompting methods. But, similar to other fixed voting rules, it cannot model nonlinear disagreement patterns.

SVM Our SVM prompt-ensembler takes the 4-bit vector of prompt predictions as input and learns a deterministic combiner. Across the 64 model-persona settings, it achieves the most consistent

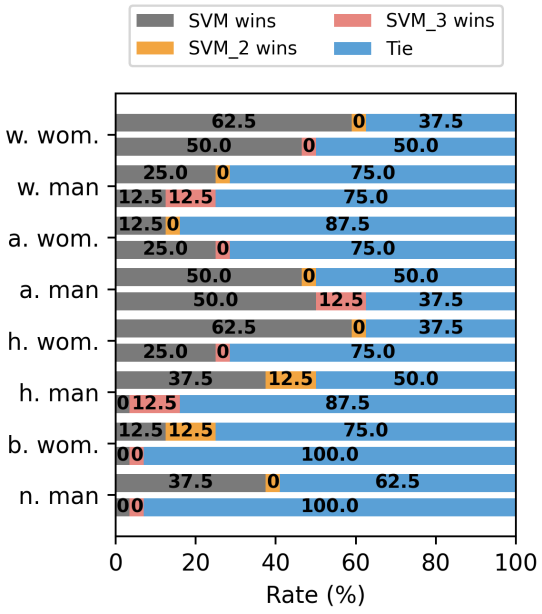

Figure 3: Comparison of two ablations with the full SVM. SVM is the original model, SVM_2 uses two inputs and SVM_3 uses three different inputs.

improvements and the largest number of significant wins (Table 2). Section 5.5 further shows that its gains are largest when any prompt disagrees, rather than on easy cases where all prompts already agree.

Table 2 compares the significant differences among each pair of methods and Table 16 in Appendix compares each method in terms of F1 scores. They show that ensembling outperforms single prompting methods, and SVM performs best overall. Best Unweighted Majority Voting is sometimes competitive, but is an oracle based on the test set results. Across methods, the overall ordering is **SVM > Best Unweighted Majority Voting > Theoretical Optimal Weighted Majority Voting = Accuracy-based Weighted Majority Voting**. Among all methods, SVM is the only one that is consistently strong across models and personas.

5.4 Error Analysis of Disagreements

We conduct an error analysis centered on the nature of different types of *prompt disagreement*. For each model-persona pair, we partition test examples into three buckets based on the four prompting strategies' binary predictions: (i) **Unanimous** (4-0), where all four promptings predict the same label; (ii) **Weak agreement** (3-1), where exactly one prompting disagrees with the other three; and (iii) **Tie** (2-2), where predictions are evenly split. For each bucket, we report its prevalence and the corresponding accuracy of methods. This analysis quantifies (a) how often prompt disagreement occurs and (b) how strongly disagreement relates to failures. It thus highlights where the SVM prompt-ensembler is most beneficial.

Disagreement occurs 25% of the time. Across all models and personas, the four promptings are unanimous on 74.7% of examples, while 18.4% fall into the 3-1 bucket and 6.9% fall into the 2-2 tie bucket. Thus, prompt disagreement appears in 25.3% of the data, with 6.9% representing the highest disagreement (2-2 ties).

Disagreements are model-dependent. The percentage of examples with disagreement (3-1 or 2-2) varies substantially across base models: `llama8b` (31.92%), `dqwen32b` (29.90%), `dqwen14b` (28.77%), `dllama8b` (28.51%), `llama70b` (25.61%), `dllama70b` (22.19%), `qwen32b` (18.47%), and `qwen14b` (17.11%). This confirms that prompt sensitivity is not uniform and depends strongly on the underlying model.

Disagreements are persona-dependent. Aggregating across models, the fraction of examples with at least some disagreement also varies by persona: black woman (29.77%), asian woman (29.04%), asian man (27.73%), native man (27.33%), white man (26.09%), hispanic woman (24.96%), hispanic man (24.90%), and white woman (23.31%). Hence, disagreement is systematic and persona-dependent rather than random.

5.5 Distinctive Input Patterns and Cases Where Ensembling Helps

We can enumerate all 16 configurations of the 4-bit vector (BPVO for base/persona/value-profile/persona-optimized) and measure their probability of occurrence in ground-truth data.

Disagreement predicts failures. Accuracy decreases sharply with disagreement. When all prompts agree (4-0), the average accuracy is approximately 0.78. With any disagreement, accuracy ranges from 0.43-0.63. Notably, the cases of ties versus singletons have overlapping accuracy ranges, suggesting that both cases represent comparably challenging examples.

Distinctiveness of the ensembling patterns. On the test data, we calculate the distinctiveness of the ensembling patterns relative to the ground truth

	Pers. Prompt	Pers. opt	Value Prof.	Weigh. Maj.	Best Maj.	Weigh. Maj. theo.	SVM
Default	7 (34)	5 (43)	8 (42)	0 (50)	0 (57)	0 (50)	0 (58)
Pers. Prompt		12 (23)	13 (25)	4 (37)	0 (40)	4 (36)	1 (51)
Value Prof			18 (19)	10 (25)	0 (24)	10 (23)	0 (47)
Pers. opt				4 (23)	0 (28)	6 (23)	1 (44)
Weigh. Maj.					2 (20)	3 (0)	3 (40)
Best Maj.						21 (1)	2 (29)
Weigh. Maj. theo.							3 (41)

Table 2: For 64 combinations (8 models x 8 personas), number of times the method in the row is better than the method in the column (the method in the column is better than the method in the row).

Pattern (r)	Type	$p(y{=}1 \mid r)$	$\lvert p-0.5\rvert$	SVM Conf.
`B0P0V0O0`	4–0	0.109	0.391	-1.00
`B0P1V1O0`	2–2	0.214	0.286	-0.60
`B0P0V0O1`	3–1	0.217	0.283	-0.63
`B1P1V1O1`	4–0	0.747	0.247	1.00
`B0P1V0O1`	2–2	0.254	0.246	-0.72

Table 3: Top-5 most distinctive BPVO patterns aggregated over models (all personas pooled), ranked by $|p(y{=}1 \mid r) - 0.5|$. BPVO represents base (B), persona (P), value-profile (V), persona-optimized (O).

Persona	Pattern (r)	Type	$p(y{=}1 \mid r)$	$\lvert p-0.5\rvert$	SVM Conf.
a. m	`B0P0V0O0`	4-0	0.050	0.450	-1.0
	`B0P1V0O0`	3-1	0.108	0.392	-0.97
a. w	`B0P0V0O0`	4-0	0.109	0.391	-1.0
	`B0P1V0O0`	3-1	0.238	0.262	-0.96
b. w	`B0P0V0O0`	4-0	0.155	0.345	-1.0
	`B0P1V0O0`	3-1	0.269	0.231	-0.83
h. m	`B0P0V0O0`	4-0	0.085	0.415	-1.0
	`B0P1V0O0`	3-1	0.189	0.311	-0.89
h. w	`B0P0V0O0`	4-0	0.183	0.317	-1.0
	`B1P1V1O1`	4-0	0.814	0.314	1.0
n. m	`B0P0V0O0`	4-0	0.035	0.465	-1.0
	`B1P0V0O0`	3-1	0.071	0.429	-1.0
w. m	`B0P0V0O0`	4-0	0.114	0.386	-1.0
	`B1P1V1O1`	4-0	0.752	0.252	1.0
w. w	`B0P0V0O0`	4-0	0.102	0.398	-1.0
	`B0P0V0O1`	3-1	0.213	0.287	-1.0

Table 4: Top-2 most distinctive BPVO patterns per persona aggregated over models, ranked by $|p(y{=}1 \mid r) - 0.5|$. The persona column uses an abbreviation for the race/ethnicity and gender.

Bucket	Method	Accuracy	95% CI
overall	`SVM`	0.750	[0.749, 0.751]
	`SVM linear`	0.743	[0.742, 0.744]
	`Best Maj.`	0.739	[0.738, 0.740]
	`Acc. W. Maj.`	0.734	[0.733, 0.735]
	`Theo. W. Maj.`	0.732	[0.731, 0.733]
	`Val. Profile`	0.726	[0.725, 0.727]
	`Persona`	0.724	[0.723, 0.725]
	`Persona Opt.`	0.716	[0.715, 0.717]
	`Default`	0.703	[0.702, 0.704]
3–1	`SVM`	0.667	[0.664, 0.670]
	`SVM linear`	0.644	[0.641, 0.647]
	`Best Maj.`	0.629	[0.626, 0.632]
	`Acc. W. Maj.`	0.599	[0.596, 0.602]
	`Theo. W. Maj.`	0.589	[0.586, 0.593]
	`Val. Profile`	0.576	[0.573, 0.580]
	`Persona`	0.574	[0.571, 0.577]
	`Persona Opt.`	0.535	[0.532, 0.538]
	`Default`	0.493	[0.490, 0.497]
2–2	`SVM`	0.650	[0.644, 0.655]
	`SVM linear`	0.609	[0.604, 0.615]
	`Theo. W. Maj.`	0.595	[0.589, 0.600]
	`Acc. W. Maj.`	0.595	[0.589, 0.600]
	`Best Maj.`	0.589	[0.584, 0.595]
	`Val. Profile`	0.541	[0.535, 0.546]
	`Persona`	0.521	[0.516, 0.526]
	`Persona Opt.`	0.508	[0.503, 0.514]
	`Default`	0.430	[0.424, 0.435]

Table 5: Bootstrap 95% confidence intervals for accuracy, aggregated over models and personas (pooled examples), grouped by disagreement bucket.

labels by assigning a probability to each of the 16 possible 4-bit prompt-output configurations of the four prompt predictions. We fix the pattern order as BPVO: base/default prompt (B), persona prompt (P), value-profile prompt (V), and persona-optimized prompt (O); and compute for each example the corresponding bit string (e.g., `B0 P1 V0 O0`). Aggregating over models (by concatenating predictions), we estimate for each pattern *r* its frequency and toxicity skew $p(y = 1 \mid r)$. Table 3 reports the top-5 globally (all personas pooled) patterns, while Table 4 reports the top-2 per persona. The last column of Table 3-4 shows a distinctiveness rank assigned to patterns by $\mid p(y = 1 \mid r) - 0.5 \mid$.

Table 3 shows that a few patterns are strongly associated with non-offensive label. For example, `B0P0V0O0` corresponds to unanimous non-toxic predictions and is strongly associated with the non-toxic label. Importantly, the top-5 list also includes both disagreement patterns (3-1, 2-2). This shows that disagreement patterns can still be informative, and that disagreement does not always mean ambiguity. Overall, the prompt outputs form structured patterns that a learned ensembler can use.

Table 4 shows that these informative patterns also vary by persona. Some recurring patterns are unanimous, such as all prompts predicting non-

Persona	TOWMV Err (3–1)	SVM Err (3–1)	$\Delta_{3\text{–}1}$	TOWMV Err (2–2)	SVM Err (2–2)	$\Delta_{2\text{–}2}$
asian man	0.379	0.271	0.108	0.354	0.229	0.125
asian woman	0.446	0.339	0.107	0.364	0.330	0.034
black woman	0.420	0.439	-0.020	0.504	0.424	0.080
hispanic man	0.427	0.331	0.096	0.390	0.314	0.077
hispanic woman	0.388	0.560	-0.172	0.434	0.544	-0.109
native man	0.346	0.270	0.076	0.361	0.249	0.112
white man	0.419	0.434	-0.014	0.434	0.411	0.023
white woman	0.403	0.329	0.074	0.369	0.324	0.045

Table 6: Error rates in the disagreement buckets (3-1 and 2-2), aggregated over models for each persona. We compare Theoretical Optimal Weighted Majority Voting (TOWMV) to the SVM prompt-ensembler. Δ denotes absolute error reduction (TOWMV -SVM); positive values indicate improvement from SVM.

toxic, but the size of the $p(y = 1 \mid r)$ and the identity of the next most distinctive pattern differ across personas. In several personas, the second most distinctive pattern is a case with one disagreement (3-1), but the pattern still points strongly to one label. In others, a unanimous all-toxic pattern is among the most distinctive. This shows that the relation between prompt-output patterns and labels is persona-dependent.

How the SVM leverages distinct disagreement patterns. Aggregation helps little when all prompting methods agree, but it becomes important in the cases with disagreement, where voting is less reliable (Table 5). In these cases, the SVM can learn which disagreement patterns are associated with the correct label. Table 6 shows that SVM consistently reduces tie errors relative to theoretical optimal weighted majority voting.

We also calculate the decision scores that correlates with the confidence of SVM for each pattern. Tables 3-4 show that SVM captures informative disagreement patterns that simpler voting rules miss.

5.6 Ablation Study on SVM

All four prompts are complementary. Figure 3 ablates the SVM by varying the number of prompt outputs used as input features. Using all four prompt predictions consistently outperforms any subset of two or three prompts, suggesting that the prompting strategies contain complementary information that cannot be recovered by any smaller set.

RBF kernel helps capture disagreement patterns. Table 5 compares SVM with an RBF versus linear kernel. The nonlinear model performs better, indicating that prompt disagreement patterns are not linearly separable and that a more flexible decision boundary is useful.

6 Conclusion

Toxicity detection is not only a modeling problem but also a perspective problem: demographically distinct personas can legitimately disagree on the toxicity of the same post. Hence, persona-conditioned classification is needed. However, across open-source LLMs, the best prompting strategy varies by both model and persona, and reasoning-enhanced variants do not reliably stabilize these judgments. Treating prompts as complementary estimators, we find that a lightweight ensembler using SVM with an RBF kernl yields the most consistent improvements and the largest number of significant wins, with bootstrap confidence intervals showing that gains occur primarily where the different prompting methods disagree. Moreover, enumerating all 16 prompt-output configurations provides an interpretable account of which disagreement patterns are most distinctive globally and per persona. We recommend that future work should model variation of toxicity judgments for given personas by predicting how judgments are distributed across annotators, and should broaden coverage to additional identities, languages, and closed-source models, to better reflect real-world moderation settings.

Limitations

Our study focuses on limited set of personas. We do not look at queer gender types and some other common races. In addition, we focus on gender and race based personas only. Further, we experiment with models from two family. Our findings might differ in closed-source models. In addition, we focus only on English but there are other languages that the models behave differently. Also, culture is an important factor that affects the perception of toxicity, but taking the culture into account is future

work. Last but not least, performance disparities among different personas remain in all methods.

References

Rui Ai, Yuqi Pan, David Simchi-Levi, Milind Tambe, and Haifeng Xu. 2025. Beyond majority voting: Llm aggregation by leveraging higher-order information. *arXiv preprint arXiv:2510.01499*.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Berk Atil, Namrata Sureddy, and Rebecca J. Passonneau. 2026. Something just like trust : Toxicity recognition of span and target. *Preprint*, arXiv:2506.02326.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. *CoRR*, abs/1903.04561.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11.1, pages 512–515.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. 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.

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. 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 2279–2292, Mexico City, Mexico. Association for Computational Linguistics.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12.1.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *The Innovation*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. *arXiv preprint arXiv:2306.02561*.

Hyukhun Koh, Dohyung Kim, Minwoo Lee, and Kyomin Jung. 2024. Can LLMs recognize toxicity? a structured investigation framework and toxicity metric. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6092–6114, Miami, Florida, USA. Association for Computational Linguistics.

Maya Kruse, Majid Afshar, Saksham Khatwani, Anoop Mayampurath, Guanhua Chen, and Yanjun Gao. 2025. Simple yet effective: An information-theoretic approach to multi-LLM uncertainty quantification. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30493–30504, Suzhou, China. Association for Computational Linguistics.

Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2023. "hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *Preprint*, arXiv:2304.10619.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35.17, pages 14867–14875.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Saurabh Mishra and Anwesha Chatterjee. 2023. Exploring chatgpt for toxicity detection in github. *Preprint*, arXiv:2312.13105.

Aida Mostafazadeh Davani, Mark Diaz, Dylan K 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.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing LLMs for hate speech detection: strengths and vulnerabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6116–6128, Singapore. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5477–5490.

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.

Taylor Sorensen, Pushkar Mishra, Roma Patel, Michael Henry Tessler, Michiel A. Bakker, Georgina Evans, Iason Gabriel, Noah Goodman, and Verena Rieser. 2025a. Value profiles for encoding human variation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2047–2095, Suzhou, China. Association for Computational Linguistics.

Taylor Sorensen, Pushkar Mishra, Roma Patel, Michael Henry Tessler, Michiel A Bakker, Georgina Evans, Iason Gabriel, Noah Goodman, and Verena Rieser. 2025b. Value profiles for encoding human variation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2047–2095.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):e0243300.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. *Preprint*, arXiv:1610.08914.

Rongwu Xu, Zian Zhou, Tianwei Zhang, Zehan Qi, Su Yao, Ke Xu, Wei Xu, and Han Qiu. 2024. Walking in others' shoes: How perspective-taking guides large language models in reducing toxicity and bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8341–8368, Miami, Florida, USA. Association for Computational Linguistics.

Han Yang, Mingchen Li, Huixue Zhou, Yongkang Xiao, Qian Fang, and Rui Zhang. 2023. One llm is not enough: Harnessing the power of ensemble learning for medical question answering. *medRxiv*.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. Textgrad: Automatic "differentiation" via text. *arXiv preprint arXiv:2406.07496*.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Pan Lu, Zhi Huang, Carlos Guestrin, and James Zou. 2025. Optimizing generative ai by backpropagating language model feedback. *Nature*, 639(8055):609–616.

Mert Yüksekgönül, Yuxuan Zhang, and Mohit Bansal. 2024. Textgrad: Automatic "differentiation" via text. *Preprint*, arXiv:2406.07496.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

A Prompts Used in the Experiments

Prompt Type	Prompt
Definition	Offensiveness denotes the overall rudeness, disrespect, or toxicity of a post. whether a post could be considered offensive to anyone.
Persona	You are supposed to decide if the given text toxic or offensive based on the following definition: *{definition}* Imagine you were persona and how you would feel about the given statement if you were *{persona}*.
Value Profile	You are supposed to decide if the given text toxic or offensive based on the following definition: *{definition}* Value Profile: *{Value Profile}*
Default	You are an helpful assistant and expert in linguistic analysis. You will decide if the given statement is offensive. *{definition}*

Table 7: Prompts used in baseline methods and the definition of toxicity used in the prompts.

B Pairwise Comparison Figures

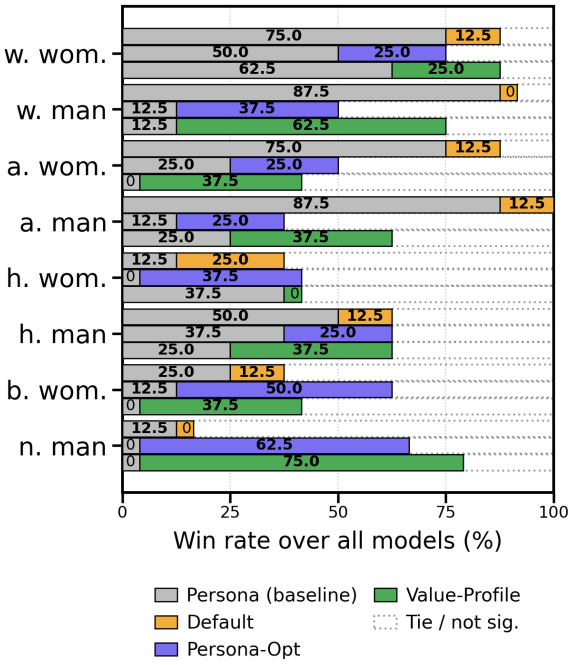

Figure 4: Prompting comparisons with **persona prompt** as the baseline.

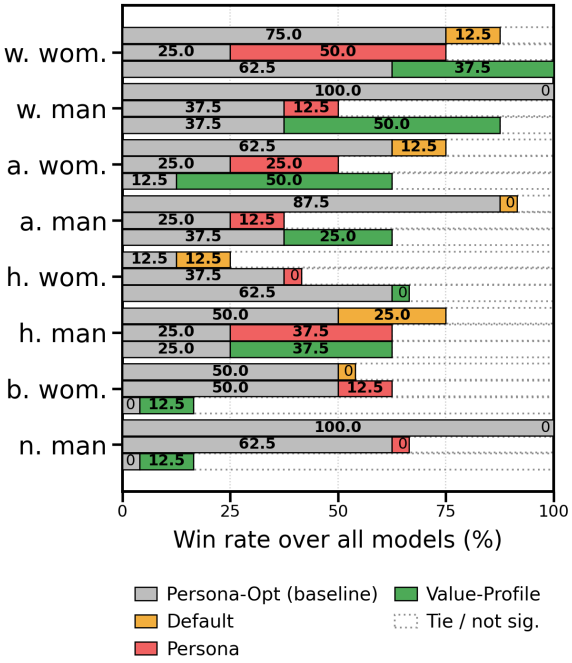

Figure 5: Prompting comparisons with **persona optimized prompt** as the baseline.

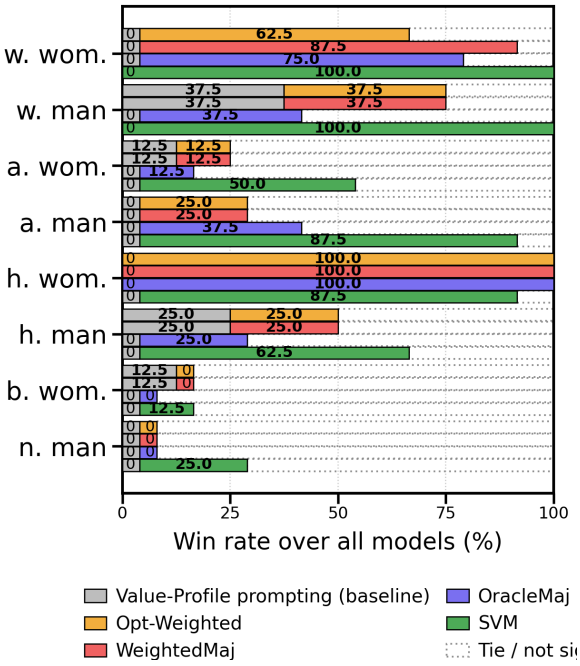

Figure 6: Ensembling comparisons with **Value Profile prompting** as the baseline.

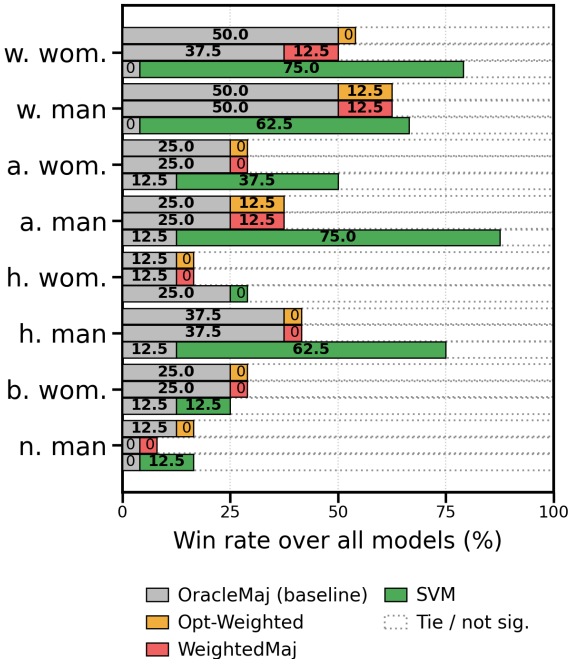

Figure 7: Ensembling comparisons with **Best Unweighted Majority** as the baseline.

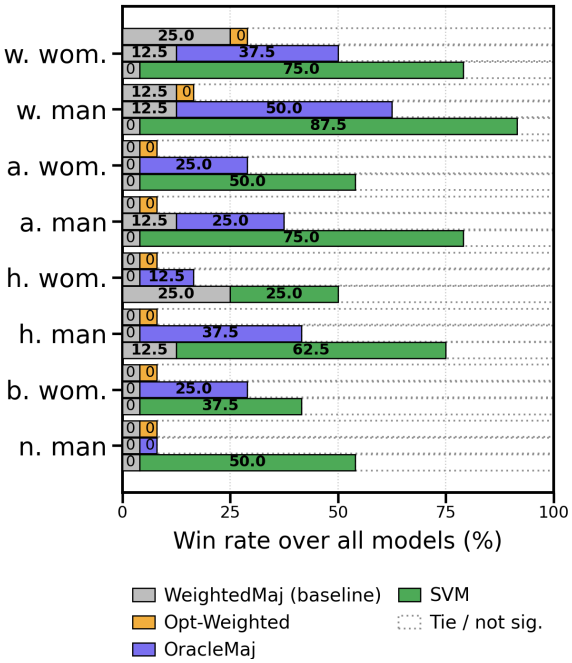

Figure 8: Ensembling comparisons with **Accuracy-based Weighted Majority** as the baseline.

C Pairwise Comparison Tables

Model/Persona	Pers. Prompt	Pers. opt	Value Prof.	Weigh. Maj.	Best Maj.	Weigh. Maj. theo.	SVM
llama8b	7 (0)	8 (0)	7 (0)	8 (0)	8 (0)	8 (0)	8 (0)
dllama8b	0 (5)	4 (1)	4 (1)	3 (0)	5 (0)	3 (0)	8 (0)
llama70b	4 (1)	5 (1)	3 (1)	5 (0)	8 (0)	5 (0)	6 (0)
dllama70b	4 (0)	4 (1)	6 (2)	7 (0)	7 (0)	7 (0)	7 (0)
qwen14b	5 (0)	6 (1)	6 (1)	6(0)	7 (0)	7 (0)	8 (0)
dqwen14b	2 (0)	4 (1)	2 (2)	7 (0)	7 (0)	7 (0)	7 (0)
qwen32b	7 (1)	6 (0)	7 (0)	7 (0)	7 (0)	7 (0)	7 (0)
dqwen32b	5 (0)	6 (0)	6 (0)	7 (0)	7 (0)	7 (0)	7 (0)
Total	34 (7)	43 (5)	42 (8)	50 (0)	57 (0)	50 (0)	58 (0)
white woman	6 (1)	6 (1)	5 (2)	7 (0)	7 (0)	7 (0)	8 (0)
white man	7 (0)	8 (0)	7 (0)	8 (0)	8 (0)	8 (0)	8 (0)
asian woman	6 (1)	5 (1)	7 (0)	7 (0)	8 (0)	8 (0)	8 (0)
asian man	7 (1)	7 (0)	7 (0)	8 (0)	8 (0)	8 (0)	8 (0)
hispanic woman	1 (2)	1 (1)	0 (6)	3 (0)	3 (0)	5 (0)	3 (0)
hispanic man	4 (1)	4 (2)	5 (0)	7 (0)	7 (0)	7 (0)	8 (0)
black woman	2 (1)	4 (0)	5 (0)	4 (0)	4 (0)	6 (0)	8 (0)
native man	1 (0)	8 (0)	6 (0)	6 (0)	6 (0)	8 (0)	7 (0)

Table 8: Out of 8 models or personas, how many times is the approach significantly better than default prompt (how many times it is worse) Pers. Prompt refers to persona prompt, Pers. opt refers to Persona optimized prompt, Value Prof. refers to Value profile prompting, Weight. Maj. refers to Accuracy-based Weighted Majority Voting, Best Maj. refers to Best Unweighted Majority Voting, Weigh. Maj. theo. refers to Theoretical Optimal Weighted Majority Voting

Model/Persona	Pers. opt	Value Prof.	Weigh. Maj.	Best Maj.	Weigh. Maj. theo.	SVM
llama8b	7 (0)	6 (1)	8 (0)	8 (0)	8 (0)	7 (0)
dllama8b	5 (1)	7 (1)	7 (0)	7 (0)	7 (0)	7 (0)
llama70b	2 (2)	4 (1)	2 (0)	3 (0)	2 (0)	3 (0)
dllama70b	2 (4)	4 (1)	5 (0)	5 (0)	5 (0)	7 (0)
qwen14b	0 (0)	3 (2)	2(2)	3 (0)	2 (2)	8 (0)
dqwen14b	4 (1)	2 (3)	6 (0)	8 (0)	6 (0)	7 (0)
qwen32b	1 (2)	1 (3)	1 (2)	2 (0)	1 (2)	4 (1)
dqwen32b	2 (2)	2 (0)	6 (0)	4 (0)	5 (0)	8 (0)
Total	23 (12)	25 (13)	37 (4)	40 (0)	36 (4)	51 (1)
white woman	2 (4)	2 (5)	4 (2)	4 (0)	3 (2)	6 (0)
white man	3 (1)	5 (1)	6 (0)	7 (0)	6 (0)	8 (0)
asian woman	2 (2)	3 (0)	4 (0)	4 (0)	4 (0)	6 (0)
asian man	2 (1)	3 (2)	2 (0)	4 (0)	2 (0)	7 (0)
hispanic woman	3 (0)	0 (3)	6 (0)	5 (0)	6 (0)	5 (0)
hispanic man	2 (3)	3 (2)	4 (1)	5 (0)	4 (1)	7 (1)
black woman	4 (1)	3 (0)	5 (1)	5 (0)	5 (1)	6 (0)
native man	5 (0)	6 (0)	6 (0)	6 (0)	6 (0)	6 (0)

Table 9: Out of 8 models or personas, how many times is the approach significantly better than persona prompt (how many times it is worse)

Model/Persona	Pers. opt	Weigh. Maj.	Best Maj.	Weigh. Maj. theo.	SVM
llama8b	4 (0)	3 (1)	4 (0)	2 (1)	4 (0)
dllama8b	5 (1)	1 (2)	1 (0)	1 (2)	7 (0)
llama70b	5 (1)	5 (1)	4 (0)	5 (1)	5 (0)
dllama70b	1 (5)	2 (1)	2 (0)	2 (1)	5 (0)
qwen14b	1 (2)	3 (2)	3 (0)	3 (2)	7 (0)
dqwen14b	4 (1)	5 (0)	5 (0)	7 (0)	7 (0)
qwen32b	3 (2)	4 (1)	4 (0)	4 (1)	5 (0)
dqwen32b	0 (3)	2 (2)	1 (0)	1 (2)	7 (0)
Total	19 (18)	25 (10)	24 (0)	23 (10)	47 (0)
white woman	5 (3)	7 (0)	6 (0)	5 (0)	8 (0)
white man	3 (4)	3 (3)	3 (0)	3 (3)	7 (0)
asian woman	1 (4)	1 (3)	1 (0)	1 (3)	5 (0)
asian man	3 (2)	3 (1)	3 (0)	3 (1)	8 (0)
hispanic woman	5 (0)	8 (0)	8 (0)	8 (0)	8 (0)
hispanic man	2 (3)	3 (2)	3 (0)	3 (2)	6 (0)
black woman	0 (1)	0 (1)	0 (0)	0 (1)	2 (0)
native man	0 (1)	0 (0)	0 (0)	0 (0)	3 (0)

Table 10: Out of 8 models or personas, how many times is the approach significantly better than value profile prompt (how many times it is worse)

D F1 Score Comparison Tables

Model/Persona	Weigh. Maj.	Best Maj.	Weigh. Maj. theo.	SVM
llama8b	0 (2)	2 (0)	0 (4)	3 (0)
dllama8b	2 (0)	2 (0)	2 (0)	7 (0)
llama70b	3 (0)	5 (0)	3 (0)	4 (0)
dllama70b	5 (1)	5 (0)	4 (0)	6 (0)
qwen14b	4 (1)	5 (0)	4 (1)	8 (0)
dqwen14b	4 (0)	5 (0)	4 (0)	6 (0)
qwen32b	1 (0)	2 (0)	1 (0)	5 (1)
dqwen32b	4 (0)	3 (0)	4 (0)	5 (0)
Total	23 (4)	28 (0)	23 (6)	44 (1)
white woman	4 (2)	4 (0)	4 (3)	8 (0)
white man	5 (0)	7 (0)	5 (1)	8 (0)
asian woman	3 (0)	4 (0)	3 (0)	6 (0)
asian man	2 (1)	3 (0)	2 (1)	6 (0)
hispanic woman	2 (1)	3 (0)	2 (1)	2 (0)
hispanic man	3 (0)	3 (0)	3 (0)	5 (1)
black woman	3 (0)	3 (0)	3 (0)	5 (0)
native man	1 (0)	2 (0)	1 (0)	4 (0)

Table 11: Out of 8 models or personas, how many times is the approach significantly better than persona optimized prompt (how many times it is worse)

Model/Persona	Best Maj.	Weigh. Maj. theo.	SVM
llama8b	4 (0)	0 (2)	4 (0)
dllama8b	2 (0)	0 (0)	7 (0)
llama70b	3 (0)	0 (0)	2 (1)
dllama70b	2 (0)	0 (0)	5 (0)
qwen14b	4 (0)	0 (0)	8 (0)
dqwen14b	1 (1)	0 (0)	3 (1)
qwen32b	2 (0)	0 (0)	4 (1)
dqwen32b	2 (1)	0 (1)	7 (0)
Total	20 (2)	0 (3)	40 (3)
white woman	3 (1)	0 (2)	6 (1)
white man	5 (1)	0 (1)	7 (0)
asian woman	3 (0)	0 (0)	5 (0)
asian man	3 (0)	0 (0)	7 (0)
hispanic woman	1 (0)	0 (0)	3 (1)
hispanic man	3 (0)	0 (0)	5 (1)
black woman	2 (0)	0 (0)	3 (0)
native man	0 (0)	0 (0)	4 (0)

Table 12: Out of 8 models or personas, how many times is the approach significantly better than weighted majority voting (how many times it is worse). This compares ensembling methods

Model/Persona	Weigh. Maj. theo.	SVM
llama8b	0 (5)	1 (0)
dllama8b	0 (2)	6 (0)
llama70b	0 (3)	0 (0)
dllama70b	0 (2)	4 (0)
qwen14b	0 (4)	6 (0)
dqwen14b	1 (1)	4 (1)
qwen32b	0 (2)	3 (1)
dqwen32b	0 (2)	5 (0)
Total	1 (21)	29 (2)
white woman	0 (4)	6 (0)
white man	1 (5)	4 (0)
asian woman	0 (3)	4 (0)
asian man	0 (3)	6 (0)
hispanic woman	0 (1)	0 (1)
hispanic man	0 (3)	5 (1)
black woman	0 (2)	2 (0)
native man	0 (0)	2 (0)

Table 13: Out of 8 models or personas, how many times is the approach significantly better than best majority voting (how many times it is worse). This compares ensembling methods

Model/Persona	SVM
llama8b	5 (0)
dllama8b	7 (0)
llama70b	2 (1)
dllama70b	5 (0)
qwen14b	8 (0)
dqwen14b	3 (1)
qwen32b	4 (1)
dqwen32b	7 (0)
Total	41 (3)
white woman	7 (1)
white man	7 (0)
asian woman	5 (0)
asian man	7 (0)
hispanic woman	3 (1)
hispanic man	5 (1)
black woman	3 (0)
native man	4 (0)

Table 14: Out of 8 models or personas, how many times is the approach significantly better than theoretical optimal weighted majority voting (how many times it is worse). This compares ensembling methods

Model-config	W Wom.	W Man	A. Wom.	A. Man	H. Wom.	H. Man	B. Wom.	N. Man
Reasoning Effect								
llama8b	0.63	0.66	0.58	**0.65**	0.66	**0.68**	0.62	**0.74**
dllama8b	**0.65**	**0.68**	**0.59**	0.61	**0.73**	0.67	**0.63**	0.66
llama70b	**0.75**	**0.76**	**0.68**	**0.74**	0.77	**0.80**	**0.70**	**0.82**
dllama70b	0.73	0.75	0.65	0.68	0.77	0.76	0.68	0.77
qwen14b	0.70	0.71	0.62	0.63	**0.77**	0.71	0.67	0.75
dqwen14b	**0.71**	**0.73**	**0.64**	**0.68**	0.76	**0.73**	**0.68**	**0.76**
qwen32b	**0.71**	**0.74**	**0.63**	**0.66**	**0.77**	**0.74**	**0.67**	**0.75**
dqwen32b	0.70	0.72	0.62	0.64	0.76	0.72	0.66	0.73
Persona Effect								
llama8b	0.63	0.66	0.58	0.65	0.66	0.68	0.62	0.74
llama8b-persona	**0.68**	**0.69**	0.63	0.71	**0.70**	**0.71**	**0.65**	**0.77**
llama8b-val_prof	**0.69**	**0.69**	0.68	0.75	**0.67**	**0.73**	**0.65**	**0.81**
dllama8b	**0.65**	0.68	0.59	0.61	**0.73**	0.67	0.63	0.66
dllama8b-persona	0.64	0.68	0.55	0.58	**0.73**	0.65	0.60	0.63
dllama8b-val_prof	**0.65**	**0.70**	**0.61**	**0.64**	0.71	**0.68**	**0.67**	**0.68**
llama70b	0.75	0.76	0.68	0.74	**0.77**	0.80	0.70	0.82
llama70b-persona	**0.78**	**0.78**	0.73	**0.82**	0.74	**0.81**	**0.71**	0.84
llama70b-val_prof	0.77	0.76	**0.74**	0.81	0.75	**0.81**	**0.71**	**0.85**
dllama70b	0.73	0.75	0.65	0.68	**0.77**	0.76	0.68	0.77
dllama70b-persona	**0.75**	**0.77**	0.69	0.78	0.76	0.76	0.69	0.79
dllama70b-val_prof	0.72	**0.77**	**0.70**	**0.79**	0.76	**0.81**	**0.70**	**0.87**
qwen14b	0.70	0.71	0.62	0.63	**0.77**	0.71	0.67	0.75
qwen14b-persona	**0.72**	0.74	0.65	0.70	**0.77**	**0.76**	**0.68**	0.75
qwen14b-val_prof	0.70	**0.76**	**0.66**	**0.71**	0.75	0.73	0.67	**0.83**
dqwen14b	**0.71**	0.73	0.64	0.68	**0.76**	0.73	0.68	0.76
dqwen14b-persona	**0.71**	**0.74**	0.63	**0.70**	**0.76**	**0.74**	0.66	0.76
dqwen14b-val_prof	0.70	**0.74**	**0.65**	0.69	0.75	**0.74**	**0.69**	**0.86**
qwen32b	0.71	0.74	0.63	0.66	**0.77**	0.74	0.67	0.75
qwen32b-persona	**0.76**	0.77	**0.69**	**0.77**	0.74	**0.81**	**0.70**	0.81
qwen32b-val_prof	0.74	**0.78**	**0.69**	0.73	0.75	0.79	**0.70**	**0.83**
dqwen32b	0.70	0.72	0.62	0.64	**0.76**	0.72	0.66	0.73
dqwen32b-persona	**0.71**	0.73	0.65	0.71	0.75	0.77	0.67	0.75
dqwen32b-val_prof	**0.71**	**0.76**	**0.66**	**0.72**	0.75	**0.78**	**0.68**	**0.82**
Automated Prompt Engineering								
lama8b-persona	0.68	0.69	0.63	0.71	0.70	0.71	0.65	0.77
llama8b-persona_opt	**0.70**	**0.71**	**0.64**	**0.77**	**0.75**	**0.73**	**0.66**	**0.82**
dllama8b-persona	**0.64**	0.68	0.55	0.58	0.73	0.65	**0.60**	0.63
dllama8b-persona_opt	0.63	**0.71**	**0.60**	**0.63**	0.73	**0.67**	0.56	**0.71**
llama70b-persona	**0.78**	0.78	**0.73**	0.82	0.74	0.81	**0.71**	0.84
llama70b-persona_opt	0.77	0.78	0.65	0.82	**0.76**	**0.82**	0.70	**0.86**
dllama70b-persona	0.75	**0.77**	**0.69**	**0.78**	0.76	**0.76**	**0.69**	0.79
dllama70b-persona_opt	**0.76**	0.75	0.65	0.76	**0.77**	0.74	0.68	**0.87**
qwen14b-persona	0.72	0.74	**0.65**	**0.70**	**0.77**	**0.76**	**0.68**	0.75
qwen14b-persona_opt	**0.73**	0.74	0.64	0.69	0.76	0.74	0.67	**0.79**
dqwen14b-persona	0.71	0.74	0.63	**0.70**	0.76	**0.74**	0.66	0.76
dqwen14b-persona_opt	**0.72**	0.74	**0.69**	0.69	**0.77**	0.71	**0.68**	0.76
qwen32b-persona	**0.76**	0.77	0.69	0.77	0.74	**0.81**	**0.70**	0.81
qwen32b-persona_opt	0.71	0.77	0.69	0.77	**0.77**	0.80	0.69	**0.83**
dqwen32b-persona	**0.71**	0.73	0.65	0.71	0.75	**0.77**	0.67	0.75
dqwen32b-persona_opt	0.69	**0.74**	0.65	**0.72**	**0.76**	0.71	**0.69**	**0.83**

Table 15: F1 scores of each configuration for each persona

Model-config	W Wom.	W Man	A. Wom.	A. Man	H. Wom.	H. Man	B. Wom.	N. Man
llama8b	0.63	0.66	0.58	0.65	0.66	0.68	0.62	0.74
llama8b-persona	0.68	0.69	0.63	0.71	0.70	0.71	0.65	0.77
llama8b-val_prof	0.69	0.69	**0.68**	0.75	0.67	0.73	0.65	0.81
llama8b-persona_opt	0.70	0.71	0.64	**0.77**	**0.75**	0.73	0.66	0.82
llama8b-best_majority	0.70	**0.72**	**0.68**	**0.77**	**0.75**	**0.74**	0.66	**0.83**
llama8b-weighted_maj	**0.71**	0.71	0.66	0.75	0.72	**0.74**	0.66	0.82
llama8b-weighted_maj_theo	**0.71**	0.71	0.66	0.75	0.72	**0.74**	0.66	0.82
llama8b-SVM	**0.71**	**0.72**	0.67	0.76	**0.75**	0.73	**0.67**	0.80
dllama8b	0.65	0.68	0.59	0.61	0.73	0.67	0.63	0.66
dllama8b-persona	0.64	0.68	0.55	0.58	0.73	0.65	0.60	0.63
dllama8b-val_prof	0.65	0.70	0.61	0.64	0.71	0.68	**0.67**	0.68
dllama8b-persona_opt	0.63	0.71	0.60	0.63	0.73	0.67	0.56	0.71
dllama8b-best_majority	0.65	0.71	0.61	0.64	**0.74**	0.68	**0.67**	0.71
dllama8b-weighted_maj	0.65	0.71	0.59	0.63	**0.74**	0.68	0.64	0.69
dllama8b-weighted_maj_theo	0.65	0.71	0.59	0.63	**0.74**	0.68	0.64	0.69
dllama8b-SVM	**0.69**	**0.73**	**0.65**	**0.71**	**0.74**	**0.72**	0.66	**0.76**
llama70b	0.75	0.76	0.68	0.74	0.77	0.80	0.70	0.82
llama70b-persona	**0.78**	0.78	0.73	**0.82**	0.74	0.81	0.71	0.84
llama70b-val_prof	0.77	0.76	**0.74**	0.81	0.75	0.81	0.71	0.85
llama70b-persona_opt	0.77	0.78	0.65	**0.82**	0.76	0.82	0.70	0.86
llama70b-best_majority	**0.78**	**0.79**	**0.74**	**0.82**	**0.78**	**0.83**	**0.72**	**0.87**
llama70b-weighted_maj	**0.78**	0.78	0.73	**0.82**	0.77	0.82	0.70	0.85
llama70b-weighted_maj_theo	**0.78**	0.78	0.73	**0.82**	0.77	0.82	0.70	0.85
llama70b-SVM	**0.78**	0.78	**0.74**	**0.82**	**0.78**	0.82	0.70	0.85
dllama70b	0.73	0.75	0.65	0.68	**0.77**	0.76	0.68	0.77
dllama70b-persona	0.75	0.77	0.69	0.78	0.76	0.76	0.69	0.79
dllama70b-val_prof	0.72	0.77	0.70	0.79	0.76	0.81	**0.70**	0.87
dllama70b-persona_opt	0.76	0.75	0.65	0.76	**0.77**	0.74	0.68	0.87
dllama70b-best_majority	0.76	0.77	0.70	0.79	**0.77**	0.81	**0.70**	0.87
dllama70b-weighted_maj	0.75	0.77	0.69	0.78	**0.77**	0.78	**0.70**	0.87
dllama70b-weighted_maj_theo	0.75	0.77	0.69	0.78	**0.77**	0.78	**0.70**	0.87
dllama70b-SVM	**0.77**	**0.78**	**0.71**	**0.80**	**0.77**	**0.82**	**0.70**	**0.88**
qwen14b	0.70	0.71	0.62	0.63	0.77	0.71	0.67	0.75
qwen14b-persona	0.72	0.74	0.65	0.70	0.77	0.76	0.68	0.75
qwen14b-val_prof	0.70	0.76	0.66	0.71	0.75	0.73	0.67	0.83
qwen14b-persona_opt	0.73	0.74	0.64	0.69	0.76	0.74	0.67	0.79
qwen14b-best_majority	0.73	0.76	0.66	0.71	**0.78**	0.76	0.68	0.83
qwen14b-weighted_maj	0.72	0.74	0.65	0.70	**0.78**	0.75	0.68	0.80
qwen14b-weighted_maj_theo	0.72	0.74	0.65	0.70	**0.78**	0.75	0.68	0.80
qwen14b-SVM	**0.75**	**0.77**	**0.69**	**0.75**	**0.78**	**0.79**	**0.69**	**0.85**
dqwen14b	0.71	0.73	0.64	0.68	0.76	0.73	0.68	0.76
dqwen14b-persona	0.71	0.74	0.63	0.70	0.76	0.74	0.66	0.76
dqwen14b-val_prof	0.70	0.74	0.65	0.69	0.75	0.74	0.69	**0.86**
dqwen14b-persona_opt	0.72	0.74	0.69	0.69	0.77	0.71	0.68	0.76
dqwen14b-best_majority	0.72	**0.76**	**0.69**	0.73	**0.78**	0.75	0.69	**0.86**
dqwen14b-weighted_maj	0.72	**0.76**	0.68	0.70	**0.78**	0.74	0.69	0.85
dqwen14b-weighted_maj_theo	0.72	**0.76**	0.68	0.70	**0.78**	0.74	0.69	0.85
dqwen14b-SVM	**0.75**	**0.76**	0.68	**0.76**	0.77	**0.78**	**0.70**	**0.86**
qwen32b	0.71	0.74	0.63	0.66	**0.77**	0.74	0.67	0.75
qwen32b-persona	**0.76**	0.77	**0.69**	0.77	0.74	**0.81**	**0.70**	0.81
qwen32b-val_prof	0.74	**0.78**	**0.69**	0.73	0.75	0.79	**0.70**	0.83
qwen32b-persona_opt	0.71	0.77	**0.69**	0.77	**0.77**	0.80	0.69	0.83
qwen32b-best_majority	**0.76**	**0.78**	**0.69**	0.77	**0.77**	**0.81**	**0.70**	0.83
qwen32b-weighted_maj	0.74	0.77	**0.69**	0.77	**0.77**	**0.81**	0.69	0.82
qwen32b-weighted_maj_theo	0.74	0.77	**0.69**	0.77	**0.77**	**0.81**	0.69	0.82
qwen32b-SVM	**0.76**	**0.78**	**0.69**	**0.78**	**0.77**	0.79	**0.70**	**0.86**
dqwen32b	0.70	0.72	0.62	0.64	0.76	0.72	0.66	0.73
dqwen32b-persona	0.71	0.73	0.65	0.71	0.75	0.77	0.67	0.75
dqwen32b-val_prof	0.71	0.76	0.66	0.72	0.75	0.78	0.68	0.82
dqwen32b-persona_opt	0.69	0.74	0.65	0.72	0.76	0.71	**0.69**	0.83
dqwen32b-best_majority	0.71	0.76	0.66	0.72	**0.77**	0.78	**0.69**	0.83
dqwen32b-weighted_maj	**0.73**	0.75	0.66	0.72	**0.77**	0.76	**0.69**	0.81
dqwen32b-weighted_maj_theo	**0.73**	0.75	0.66	0.72	**0.77**	0.76	**0.69**	0.81
dqwen32b-SVM	**0.73**	**0.77**	**0.68**	**0.78**	0.76	**0.80**	**0.69**	**0.86**

Table 16: F1 scores of each configuration for each persona. This table compares all methods.