The Subjectivity of Respect in Police Traffic Stops: Modeling Community Perspectives in Body-Worn Camera Footage

Preni Golazizian[1], **Elnaz Rahmati**[1], **Jackson Trager**[2], **Zhivar Sourati**[1], **Nona Ghazizadeh**[1,2], **Georgios Chochlakis**[1], **Jose Alcocer**[6*], **Kerby Bennett**[4,5*], **Aarya Vijay Devnani**[1*], **Parsa Hejabi**[1*], **Harry G. Muttram**[3*], **Akshay Kiran Padte**[1*], **Mehrshad Saadatinia**[1*], **Chenhao Wu**[1*], **Alireza S. Ziabari**[1*], **Michael Sierra-Arévalo**[7†], **Nick Weller**[3†], **Shrikanth Narayanan**[1,2†], **Benjamin A. T. Graham**[5†] and **Morteza Dehghani**[1,2†]

[1]Department of Computer Science, University of Southern California,
[2]Department of Psychology, University of Southern California,
[3]Department of Political Science, University of California Riverside,
[4]Department of Anthropology, University of California Los Angeles,
[5]Department of Political Science and International Relations, University of Southern California,
[6]Harvard Law School, Harvard University, [7]Department of Sociology, The University of Texas at Austin
golazizi@usc.edu

Abstract

Traffic stops are among the most frequent police–civilian interactions, and body-worn cameras (BWCs) provide a unique record of how these encounters unfold. Respect is a central dimension of these interactions, shaping public trust and perceived legitimacy, yet its interpretation is inherently subjective and shaped by lived experience, rendering community-specific perspectives a critical consideration. Leveraging unprecedented access to Los Angeles Police Department BWC footage, we introduce the first large-scale traffic-stop dataset annotated with *respect ratings* and free-text *rationales* from multiple perspectives. By sampling annotators from police-affiliated, justice-system-impacted, and non-affiliated Los Angeles residents, we enable the systematic study of perceptual differences across diverse communities. To this end, *i)* we develop a domain-specific evaluation rubric grounded in procedural justice theory, LAPD training materials, and extensive fieldwork; *ii)* we introduce a rubric-driven preference data construction framework for perspective-consistent alignment, and *iii)* we propose a perspective-aware modeling framework that predicts personalized respect ratings and generates annotator-specific rationales for both *officers* and *civilian drivers* from traffic-stop transcripts. Across all three annotator groups, our approach improves both rating prediction performance and rationale alignment. Our perspective-aware framework enables law enforcement to better understand diverse community expectations, providing a vital tool for building public trust and procedural legitimacy.

*These authors contributed equally to this work.
†Senior authors.

1 Introduction

Police traffic stops are common, but fraught interactions between the public and their government. When officer-driver communication is ineffective, it undermines public trust and may lead to violent outcomes (Tyler and Huo, 2002). At present, the vast majority of stops in major police jurisdictions in the U.S. are captured on bodyworn cameras (BWCs), a practice now becoming common worldwide (e.g., Laming, 2019), offering high-quality data to understand and improve these interactions. The sheer volume of data, however, precludes the possibility of manual review.

The interpretation of complex social interactions, such as police-civilian encounters, is inherently *subjective*. Perspectives on good communication vary across communities, influenced by lived experiences and systemic differences rooted in past encounters with the justice system (Sierra-Arévalo et al., 2025). For example, previous research shows that Black people experience more frequent and more negative police interactions (Pierson et al., 2020; Xu et al., 2024), affecting perceptions of police legitimacy and potentially contributing to fear, anxiety, and distrust within communities (Pickett et al., 2022).

Prior research has established *respect* as a measurable and consequential dimension of police-civilian interactions (e.g., Voigt et al., 2017; Camp and Voigt, 2024). What remains less explored is how differences in lived experience shape interpretation and how to incorporate such variation into computational models. This gap is particularly critical because when the subjectivity of core

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 33939–33961

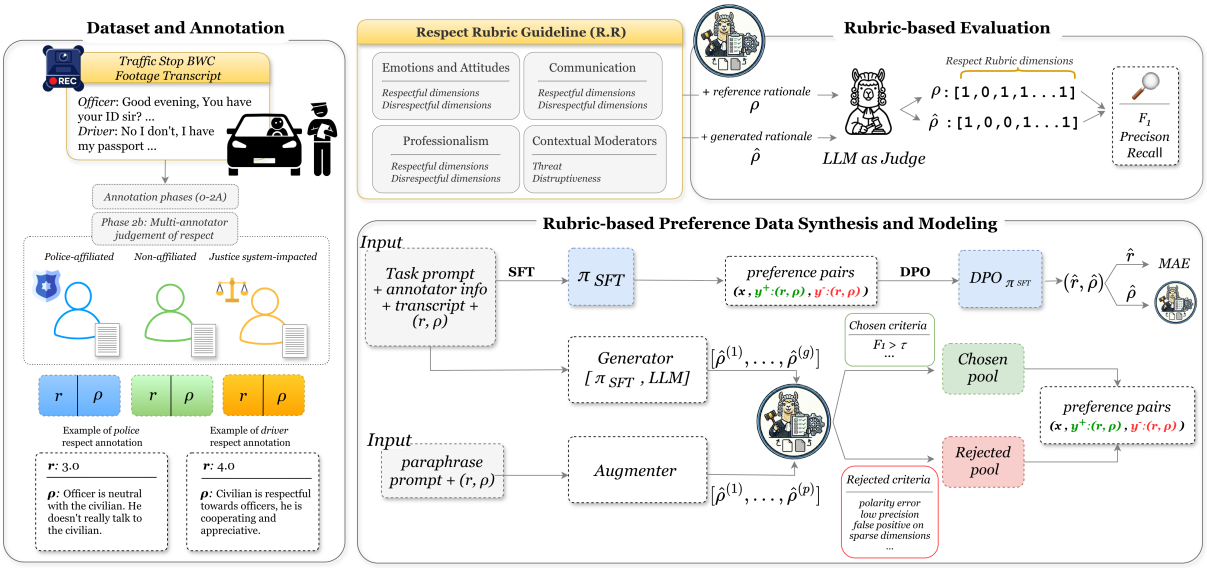

Figure 1: Overview of the proposed framework, illustrating multi-perspective respect annotation from BWC transcripts, and domain-specific rubric development. The framework integrates rubric-guided preference data synthesis with supervised fine-tuning and alignment to produce perspective-aware respect ratings and rationales.

concepts, like respect, is ignored in ML, efforts to establish a single ground truth have historically privileged the perspectives of the dominant institutional and majority-group perspectives (Turner Lee, 2018; Barabas et al., 2020). This focus on a single perspective overlooks how respect is interpreted differently across communities and limits the ability of models to reflect these differences.

In this work, we focus on *respect*, as expressed by both officers and drivers. Respect is a central dimension of police-civilian interactions, influencing public trust and the perceived legitimacy of the state's authority (Tyler and Huo, 2002; Voigt et al., 2017), and consistently emerging as a key factor in how individuals interpret encounter outcomes. This focus is built upon established procedural justice literature, extensive survey, and qualitative interviews in Los Angeles (Sierra-Arévalo et al., 2025).

Recognizing the subjective nature of respect in police-driver interactions, it is essential to employ subjective computational modeling techniques that accommodate multiple plausible perspectives and interpretations. Therefore, we undertook an 18-month effort to annotate Los Angeles Police Department (LAPD) BWC data from a demographically and experientially diverse pool of annotators: *police-affiliated* individuals, *justice-system-impacted* individuals, and other Los Angeles residents. This multi-perspective approach captures both shared and differing views of respect.

In this work: *(1)* we introduce a large-scale dataset of annotations of LAPD traffic-stop BWC footage, consisting of multi-annotator, fine-grained respect ratings and rationales collected from a demographically and experientially diverse pool of annotators. The dataset reflects the subjective and heterogeneous ways community members interpret officer and civilian behavior. *(2)* we develop a domain-specific, theory- and data-grounded *respect rubric* that synthesizes community perspectives, LAPD training materials, and prior research to characterize respectful and disrespectful communicative behaviors in traffic stops. The rubric is operationalized through an LLM-as-a-judge framework to evaluate human and model-generated rationales, enabling interpretable comparison and the construction of a rubric-guided preference dataset. *(3)* we model the subjective, group- and individual-dependent nature of perceived respect with a single annotator-aware and group-aware model, which conditions on annotator background to personalize predictions across community members.

Our framework, as illustrated in Figure 1, predicts respect ratings and annotator-specific rationales that reflect group's interpretive lens, with rubric-guided preference data improving model alignment and rationale quality across groups.

2 The LAPD Respect Dataset

Our dataset comprises BWC videos and their annotations from a stratified random sample of about 1,000 LAPD traffic stops recorded between

09/2021- 09/2022. These videos and their annotations were collected through a large-scale, multi-year collaboration with LAPD, facilitated by the Office of the Inspector General. Footage is recorded from officers' BWCs, which are typically activated at the start of a stop, though coverage can vary due to activation timing and technical factors.

We collect fine-grained annotations from a diverse pool of trained annotators (§2.1) across multiple phases (§2.2). This work focuses on the phase involving *respect* ratings and free-text *rationales* for both officers and civilians, using conversation transcripts as the primary input modality.

2.1 Annotator Selection and Setting

The core objective of this project is to systematically document and model the heterogeneity of community perceptions regarding police-civilian encounters. To achieve this, we have conducted extensive fieldwork, including officer and community focus groups and interviews, a representative survey of over 2,000 LA residents, an analysis of LAPD training manuals, and ride-alongs with LAPD officers. Crucially, these diverse engagements reveal that individuals hold different, and often conflicting, preferences for the same communicative behaviors; this finding confirms previous research (e.g., Weaver and Lerman, 2010) that lived experiences with the justice system strongly shape the interpretation of respect, validating the central premise of our subjective modeling approach. Based on our fieldwork, annotators are recruited from three groups: *(1) Police-affiliated* (G_{PA}): individuals with prior law enforcement experience; *(2) Justice-system-impacted* (G_{JI}): individuals with a history of incarceration or arrest; and *(3) Non-affiliated* (G_{NA}): community members without prior law enforcement experience or history of incarceration/arrest. This design ensures that the subjectivity motivating our study is reflected in the collected respect annotations, rather than collapsing judgments into a single perspective.

2.2 Annotation Phases and Guideline Development

The annotation framework was developed through iterative refinement informed by fieldwork to ensure the schema reflects community expectations and institutional constraints (Trager et al., 2024). To operationalize this process, we built a domain-specific, secure annotation platform (Appendix A). The annotation proceeded in the following stages:

	Statistic	G_{PA}	G_{NA}	G_{JI}	All
	#Annotators	5	20	2	27
	#Annotations	300	872	190	1362
Officer	Mean(ratings)	3.69	3.61	4.07	3.69
	Std(ratings)	1.00	0.78	0.63	0.83
	Rationale length (tokens)	51.6	42.7	39.1	44.2
Driver	Mean(ratings)	3.97	3.65	4.03	3.77
	Std(ratings)	0.88	0.83	0.77	0.85
	Rationale length (tokens)	43.9	42.0	39.8	42.1

Table 1: Phase 2B respect annotation statistics by annotator group and entity.

Preliminary Phases. In *Phase 0*, annotators corrected noisy WhisperX (Bain et al., 2023) transcripts. Subsequent stages involved tagging individuals and objects (*Phase 1A*), labeling perceived demographics (*Phase 1B*), and tracking emotions alongside objective stop outcomes (*Phase 2A*).
Phase 2B: Multi-Annotator Judgments of Respect. As the central component of this work, annotators produced fine-grained subjective judgments of respect ratings and rationales for both officers and civilians for a subset of traffic stops. Each sampled video is labeled by one to four annotators with varied police-related lived experiences, enabling the analysis of perceptual differences. Find summary statistics of this phase annotations in Table 1.

We empirically validate a core assumption underlying this dataset: that annotators' free-text rationales reflect their corresponding respect ratings for both officers and civilians (see Appendix F).

3 Modeling Subjective Respect Annotations

We describe our task and the modeling components to predict and explain annotator-specific perceived respect in police–civilian interactions. Our approach relies on a domain-specific rubric for evaluating rationale quality, and uses this rubric to construct preference data for personalized alignment. We begin by defining the task, then present the development and motivation for the rubric, followed by rubric-guided preference data construction.

3.1 Task Definition

Given a set of conversations $C = \{c_i\}_{i=1}^{N}$ from traffic-stop interactions between officers and civilians, and a set of annotators $A = \{a_j\}_{j=1}^{M}$, we represent the respect annotation provided by annotator a_j for conversation c_i as $y_{ij} = (r_{ij}, \rho_{ij})$, where $r_{ij} \in [1, 5] \cap \mathbb{Z}$ is the respect rating (from

very disrespectful to *very respectful*) and ρ_{ij} is a free-text rationale explaining the annotator's rating.

Each annotator a_j is associated with a group label $g(a_j)$ reflecting their prior exposure to the criminal justice system: *Police-affiliated* (G_{PA}), *Justice-system-impacted* (G_{JI}), or *Non-affiliated* (G_{NA}), as described in §2.1. Each annotator also has a set of demographic attributes $\mathrm{demo}(a_j)$. These annotator-specific attributes are used in our modeling to signal differences in their perspectives.

We denote the input prompt for conversation c_i, conditioned on annotator a_j, as x_{ij}. The resulting dataset is defined as: $D = \{d_{ij} : x_{ij} \mapsto y_{ij}\}$ where each data point d_{ij} maps a prompt to its corresponding rating and rationale.

Our goal is to: *(1)* predict the overall level of respect $\hat{r}_{ij}$ displayed by the officer and civilian in conversation c_i, from the perspective of annotator a_j; *(2)* generate a rationale $\hat{\rho}_{ij}$ that explains the reason behind the rating provided by annotator a_j, and *(3)* assess how well the model captures the distinct perspectives of each group G_{PA}, G_{JI}, and G_{NA} by evaluating performance at the group level.

We formulate this task as a conditional language generation problem. The prompting format used for model training is provided in Appendix C.

3.2 Respect Rubric Development

Evaluating the subjective nature of perceived respect in traffic stops requires more than predicting a numeric score; it requires understanding the reasoning behind each judgment. For this reason, our annotation process includes a question asking for a free-text rationale accompanying every respect score. These rationales allow us to capture the nuanced cues and diverse perceptions of what constitutes respectful versus disrespectful behavior across different community members.

A central challenge in this task is determining what counts as a *good* rationale. Standard text-similarity metrics (e.g., ROUGE, BLEU) are ill-suited for evaluating respect-related reasoning, as they cannot capture the nuanced, domain-specific cues that annotators rely on. More broadly, evaluating rationales in this setting requires a method that compares generated explanations to reference rationales in a structured and interpretable manner. Such an evaluation should be anchored in the specific dimensions of communication relevant to traffic stops, making explicit which elements of respect or disrespect are present, absent, or mischaracterized in a given rationale. Overall, our evaluation framework needs to satisfy two criteria: (1) *interpretability and reasoning*: It must reflect the key elements and reasons behind a respect judgment; (2) *policing-domain grounding*: It must be explicitly grounded in policing domain expertise and synthesize the full range of respectful and disrespectful communicative behaviors.

Respect Rubric Guideline. To meet these requirements, we develop a survey- and theory-informed guideline that identifies the concrete aspects of officer-civilian communication that annotators commonly reference when explaining their provided respect ratings.

The rubric emerges from the following complementary sources: *(i)* the annotation manual, which already links objective and subjective variables to prior research; *(ii)* extensive fieldwork, including the citywide survey, qualitative interviews, researcher ride-alongs, and insights from LAPD training materials; and *(iii)* academic literature on procedural justice, transparency, and de-escalation (Sunshine and Tyler, 2003a; Mazerolle et al., 2013). Drawing on these sources, the rubric organizes respect into three core (overlapping) categories: *Emotions*, *Professionalism*, and *Communication*. A separate category of *Contextual Moderators* is included to capture situational conditions that can influence the interpretation of the core categories.

The resulting rubric, shown in Appendix B, specifies both respectful and disrespectful elements within each category. For example, emotional respect may involve warmth, empathy, or calm body language, while disrespect may manifest as offensiveness or unnecessary anger. Professionalism includes greetings, introductions, and composure under stress, while lapses include abrupt commands or mocking language. Communication covers explanations, comprehension checks, and signaling when civilians are free to leave. Contextual moderators, such as threats or environmental disruptions, shape how these elements are expressed and how annotators justify their ratings. By anchoring rationale evaluation in these descriptors, the rubric allows for systematic, transparent comparison of annotator-written and model-generated reasoning.

LLM-as-a-judge. We implement a rubric-aware LLM-as-a-judge, validated by human annotators, that maps both ground-truth rationales (ρ) and model-generated rationales ($\hat{\rho}$) to structured representations grounded in the respect rubric. Using `LLaMA-3-70B` (Grattafiori et al., 2024), the

judge identifies which respect-related elements are present in a given rationale. This rubric-grounded representation enables systematic comparison of human and model reasoning and serves as a shared evaluation primitive throughout our framework. Judge prompts are provided in Appendix B.

Metric Formulation. Let the respect rubric consist of K dimensions. Given a rationale ρ, the judge produces a binary activation vector $z(\rho) \in \{0,1\}^K$, indicating the presence or absence of each rubric element.

We evaluate a generated rationale $\hat{\rho}$ by computing macro precision $P(\hat{\rho})$, recall $R(\hat{\rho})$, and F_1 between its activation vector and the reference $z(\rho)$. This rubric-based metric provides an interpretable measure of rationale quality and is used both for evaluation and for constructing preference data in the alignment stage.

3.3 Rubric-Grounded Preference Data Synthesis

Rationales in this domain are nuanced, structurally varied, and context-dependent. To align models toward such reasoning, we introduce a rubric-grounded preference dataset synthesis framework composed of the following modules (see Figure 1).

GENERATOR MODULE. Given an input prompt x_{ij} and a target model π_t, the Generator Module produces a set of candidate rationales, $\{\hat{\rho}_{ij}^{(g)}\}_{g=1}^{G} \sim \pi_t(\cdot \mid x_{ij})$. These candidates may originate from different target models, including a supervised fine-tuned model or a larger base model.

AUGMENTER MODULE. To augment high-quality rationales, we apply a paraphrasing model π_{phr} to the reference rationale ρ_{ij}, $\{\hat{\rho}_{ij}^{(p)}\}_{p=1}^{P} \sim \pi_{\text{phr}}(\cdot \mid \rho_{ij})$, producing paraphrased variants that preserve the underlying reasoning. These paraphrases are evaluated in the same manner as generated candidates and may be included in the chosen pool if they satisfy the rubric-alignment criteria.

JUDGE MODULE. We reuse the rubric-aware judge to evaluate both generated candidates and reference rationales within the preference synthesis pipeline. For each candidate, $\hat{\rho}_{ij}^{(g)}$ or $\hat{\rho}_{ij}^{(p)}$, and the corresponding reference rationale ρ_{ij}, the judge assigns binary rubric activations $z(\hat{\rho}_{ij})$ and $z(\rho_{ij})$.

Candidate rationales are assigned to *chosen* or *rejected* pools based on rubric-based alignment with the reference rationale. For candidates derived from generator module $\hat{\rho}_{ij}^{(g)}$, we apply empirically determined thresholds, τ_{ch} and τ_{rej}, to ensure clear separation between high- and low-quality reasoning. A candidate is assigned to the *chosen* pool if $F_1(\hat{\rho}) \geq \tau_{ch}$. A candidate is assigned to the *rejected* pool if it satisfies any rejection condition, including *(i)* low overall alignment ($F_1(\hat{\rho}) < \tau_{rej}$), *(ii)* low precision or recall ($P(\hat{\rho}) < \tau_{rej}$ or $R(\hat{\rho}) < \tau_{rej}$), *(iii)* systematic false positives on commonly over-generated rubric dimensions (e.g., *warmth*), or *(iv)* false negatives on sparse but critical rubric dimensions. Candidates that satisfy neither condition are discarded.

For $\hat{\rho}^{(p)}$, we apply only the *chosen* criteria: paraphrases that meet the alignment threshold are included in the chosen pool, and others are discarded.

Preference Pair Construction. For each instance d_{ij}, we construct up to k preference pairs by pairing elements from the chosen—which also includes ground truths—and rejected pools: $(x_{ij}, y_{ij}^{+}, y_{ij}^{-})$. The resulting preference pairs dataset is used for downstream preference optimization of the target model to make the model better aligned with the rubric-grounded respect values embedded in the ground truth respect annotations.

4 Experiments

We first present a mixed-effects analysis to quantify how lived experience drives divergent interpretations of respect. We then describe our experiments, which proceed in three stages: *(i)* prompt-level, *(ii)* model-level, and *(iii)* alignment-level.

4.1 Mixed-Effects Analysis of Respect Annotations.

A core motivation for our modeling framework is the substantial subjectivity with which annotators perceive respect in police-civilian encounters. As views vary with lived experience, we first examine how annotator background including group identity $g(a_j)$ and demographic attributes $\text{demo}(a_j)$, explain variation in the ratings r_{ij} they assign. To quantify this variation, we fit mixed-effects regression models (details in Appendix D), treating annotators as random effects and incorporating annotator background as predictor variables.

Across both officer-respect and civilian-respect models, we find annotator group identity $g(a_j)$ as the only annotator-level attribute that significantly

predicted perceived respect. Specifically, annotator group identity has a significant effect on respect annotations of officers ($\beta = 0.541, p < .001$) as well as on respect annotations of drivers ($\beta = 0.501, p = .046$). In contrast, other annotator demographic variables, including age, gender, and race, do not exhibit significant main effects on respect judgments. Taken together, these results suggest that lived experience with the criminal justice system is the dominant lens through which annotators interpret respect. Additional model specifications and coefficients are included in the Appendix D.

4.2 Prompt-level Experiments

This stage serves two goals: establishing baseline performance, and assessing whether explicitly conditioning on annotator background information improves subjective predictions. To this end, we augment the task prompt with different subsets of the annotator background variables and evaluate the zero-shot performance of the base model on both rating and rationale generation.

This allows us to test whether making the annotator's perspective explicit via prompting enables the model to approximate annotator a_j's subjective interpretation *without* any parameter updates.

As confirmed by our findings in Section 4.1, $g(a_j)$ is the strongest predictor of subjective rating variation. Hence, we begin with conditioning the model on $g(a_j)$ and incrementally incorporate additional background variables. The prompt configurations (examples in Appendix C) are: *(i)* $Base$: No annotator information included; *(ii)* $Base^{(g)}$: Group-Personalized Baseline; the prompt is conditioned only on the annotator's group identity $g(a_j)$; *(iii)* $Base^{(g+demo)}$: Annotator-Personalized Baseline; the prompt includes group identity and annotator demographic attributes, including age, race, and gender; and *(iv)* $Base^{(g+demo+ent)}$: Target Entity-Personalized Baseline; the prompt includes annotator group, annotator demographics, and perceived demographic attributes of the officer or civilian involved in the conversation.

4.3 Model-level Experiments

To adapt the model to the domain and learn personalized generation patterns, we apply parameter-efficient fine-tuning via LoRA (Hu et al., 2022). We condition the model on the $(g + demo)$, since this configuration yields better result in zero-shot setting. The prompt template is shown in Appendix C.

During training, we apply the standard autoregressive language modeling objective, computing the negative log-likelihood loss only over the tokens corresponding to the rating r_{ij} and the rationale ρ_{ij}. This produces the supervised model π_{sft} that captures annotator-specific patterns directly from data, enabling it to more accurately predict subjective ratings and produce rationales aligned with annotator a_j's interpretive perspective.

4.4 Alignment-level Experiments

To evaluate the impact of alignment, we compare two types of preference datasets: *(i)* pairs derived from the original human annotations (Ground-truth-based) and *(ii)* augmented rubric-grounded preference data as explained in Section 3.3.

Ground-truth-based Preference Data. For this dataset, we construct preference pairs directly from human annotations. For each annotated instance d_{ij}, we define the *chosen output* as the annotator's rating and rationale, $y_{ij}^{+} = (r_{ij}, \rho_{ij})$. The *rejected outputs* are the ratings and rationales assigned to the same instance d_{ij} by other annotators, denoted $y_{i\ell}^{-} = (r_{i\ell}, \rho_{i\ell})$ for all $\ell \neq j$. This process yields preference tuples of the form $(x_{ij}, y_{ij}^{+}, y_{i\ell}^{-})$, capturing natural disagreement across annotators and ensuring that each annotator's perspective is preferred in the alignment objective.

Rubric-Grounded Preference Data. Rubric-driven preference pairs, introduced in Section 3.3, are constructed by contrasting *chosen* high-quality rationales with *rejected* rationales, according to the rubric-based criteria (see Appendix A.2 for details). Candidate rationales are generated either by the target supervised fine-tuned model π_{sft} or by a larger language model in a zero-shot setting, specifically `Qwen3-30B-A3B-Instruct-2507-FP8` (Yang et al., 2025). Training the aligned model on this dataset encourages it to generate more specific reasoning, better capture sparse but domain-critical disrespect dimensions, and correct systematic errors in its explanations, thereby moving beyond generic or overly positive statements.

We apply Direct Preference Optimization (DPO; Rafailov et al., 2023) using both preference datasets and evaluate four alignment configurations: *(i)* DPO_{gt}: DPO with original data; DPO applied directly to the base model using the original annotation-derived preference pairs. *(ii)* DPO_{rub}: DPO with rubric-augmented data; DPO applied to

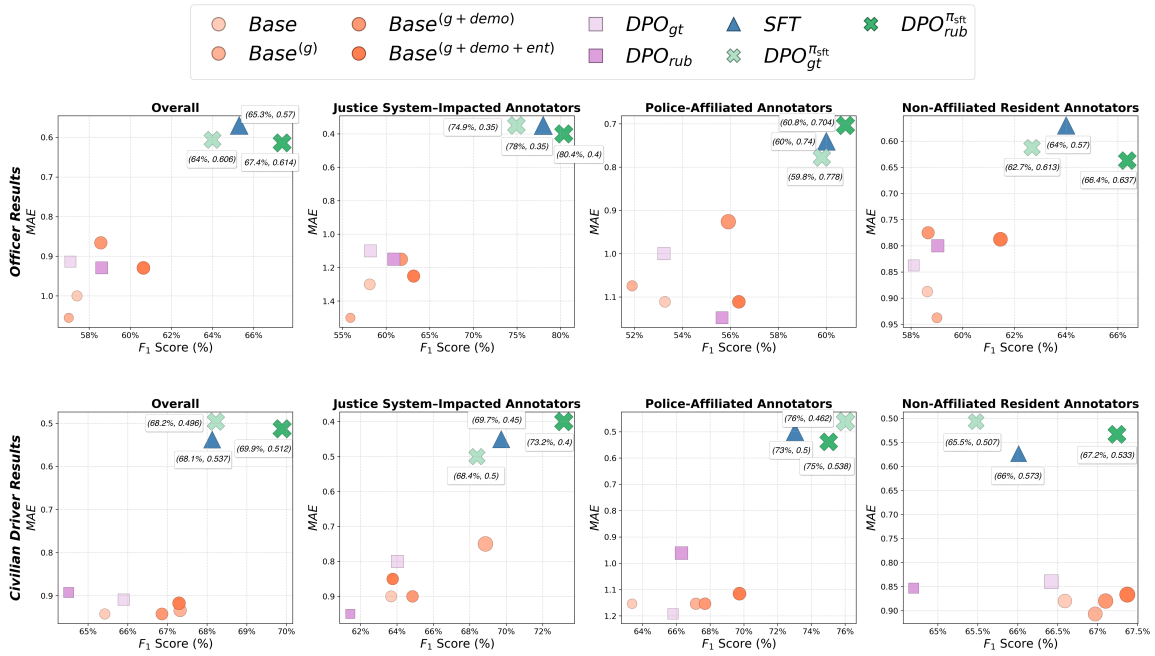

Figure 2: Rating MAE (lower is better) and $F_1(\hat{\rho})$ (higher is better) for officers (top) and drivers (bottom).

the base model using the rubric-constructed preference pairs described above. *(iii)* $DPO_{gt}^{\pi_{sft}}$: DPO with original data post-SFT; DPO applied to the supervised model π_{sft} using the original annotation-derived preference pairs. *(iv)* $DPO_{rub}^{\pi_{sft}}$: DPO with rubric-augmented data post-SFT; DPO applied to π_{sft} using rubric grounded preference data.

This setup enables a controlled and systematic assessment of the impact of each modeling component, allowing us to isolate and attribute the performance gains, particularly those driven by our rubric-based alignment. The implementation details can be found in Appendix E.

5 Results

We evaluate our models for the two primary interactants in traffic stops: the *officer* and the *driver*. For rating prediction, we report mean absolute error (MAE; lower is better). For rationale generation, we report rubric-based $F_1(\hat{\rho}_\pi)$; higher is better. Results are shown in Figure 2, aggregated over the test set and per annotator group.

5.1 Overall Performance

Across both officer and driver evaluations, $DPO_{rub}^{\pi_{sft}}$ achieves the highest overall $F_1(\hat{\rho}_\pi)$. Zero-shot models perform poorly along both dimensions, exhibiting higher MAE and lower $F_1(\hat{\rho}_\pi)$, and applying DPO without prior SFT does not yield consistent improvements. In contrast, SFT yields substantial gains, which we further build upon by applying DPO with our constructed datasets ($DPO_{rub}^{\pi_{sft}}$ and $DPO_{gt}^{\pi_{sft}}$) consistently improving $F_1(\hat{\rho}_\pi)$over SFT alone, while maintaining rating MAE on par with the SFT model. Specifically for officer rationales, the $F_1(\hat{\rho}_\pi)$ improves from 65.3% to 67.4% ($\uparrow 2.1\%$), and for driver rationales from 68.1% to 69.9% ($\uparrow 1.8\%$).

5.2 Group-Specific Performance

We analyze annotator group-specific performance to assess how our alignment method differentially benefit annotators with distinct perspectives. For each group, we report performance on officer and driver respect evaluations, focusing on both rationale quality and rating error.

Justice-System-Impacted Annotators. Alignment over SFT improves the performance most strongly for *Justice-system-impacted* annotators. Both SFT and $DPO_{gt}^{\pi_{sft}}$ achieve the lowest rating error, with MAE of 0.35. Additionally, $DPO_{rub}^{\pi_{sft}}$ improves rationale quality over SFT, increasing $F_1(\hat{\rho}_\pi)$ from 78.0% to 80.4% ($+2.4\%$). For driver, $DPO_{rub}^{\pi_{sft}}$ achieves the best overall performance, attaining an MAE of 0.40 and improving rationale quality to 73.2%, a gain of $+3.5\%$ over SFT.

Police-Affiliated Annotators. For this group, preference-based alignment yields consistent improvements in rationale quality for both officer and driver evaluations. For officer respect, SFT

achieves an $F_1(\hat{\rho}_\pi)$ of 60.0% with a corresponding MAE of 0.74. Applying $DPO_{rub}^{\pi_{sft}}$ improves $F_1(\hat{\rho}_\pi)$ to 60.8% while reducing rating error to 0.704. For driver respect, gains from preference alignment are more pronounced. SFT attains an $F_1(\hat{\rho}_\pi)$ of 73% with $\text{MAE}_{\text{rating}} = 0.50$, while $DPO_{gt}^{\pi_{sft}}$ improves rationale quality to 76% and reduces rating error to 0.462.

Non-Affiliated Resident Annotators. For officer respect, SFT attains an $F_1(\hat{\rho}_\pi)$ of 64.0% and MAE of 0.57. Applying $DPO_{rub}^{\pi_{sft}}$ further improves the $F_1(\hat{\rho}_\pi)$ to 66.4% but results in higher rating error (MAE = 0.637). For driver respect, $DPO_{rub}^{\pi_{sft}}$ performs better relative to SFT alone for both rating and rationale. However, the baseline models perform on par with $DPO_{rub}^{\pi_{sft}}$ on rationale generation, indicating limited benefit from alignment for this group's rationale generation. Nevertheless, $DPO_{rub}^{\pi_{sft}}$, achieves a lower rating error than the best-performing baseline ($\downarrow$ 0.3).

Taken together, these results indicate that while SFT provides a strong foundation—particularly for rating prediction—rubric-grounded preference alignment is crucial for improving perspective-specific rationale generation. The largest relative gains from post-SFT alignment arise for justice-system-impacted annotators, followed by police-affiliated annotators, highlighting the value of aligning rationales to distinct community perspectives.

6 Related Work

Respect and Communication in Policing. Research shows that officer respect during traffic stops significantly influences public trust and perceived legitimacy (Tyler, 1988, 2017; Worden and McLean, 2018; Sunshine and Tyler, 2003b; Nagin and Telep, 2017; Jackson et al., 2012). These interactions shape broader relations with the state. Black drivers systematically receive less respect in routine stops (Voigt et al., 2017), a reality that likely shapes community expectations and demands for respectful treatment (Camp et al., 2024). Recent research uses BWC transcripts to evaluate officer training on respectful traffic stops (Camp et al., 2024), as well as how civilian demeanor affects officer behavior (Sunde et al., 2023).

Personalization in LLMs. Personalization is a major focus in LLM research, with surveys detailing its evolution. Input-level methods (Liu et al., 2025a) encode identity via persona-guided prompts (Ryan et al., 2025), embeddings (Li and Liang, 2021; Doddapaneni et al., 2024; Huber et al., 2025; Liu et al., 2025b), soft adapters (Hebert et al., 2024), or distilled prompts (Ramos et al., 2024), while conversational systems generate tailored narratives (Sayana et al., 2025). Model-level approaches utilize parameter-efficient adaptation, such as personalized LoRA variants for various tasks (Zhang et al., 2024; Tan et al., 2024; Zhu et al., 2024; Kong et al., 2024; Long et al., 2024) and personality customization via mixture of experts (Dan et al., 2025). Alignment-based strategies modify objectives through group optimization (Zhao et al., 2023), personalized reinforcement learning (Li et al., 2024), or parameter merging (Jang et al., 2023), alongside inference-time steering for retraining-free control (He et al., 2024; Cao et al., 2024; Bo et al., 2025; Zhang et al., 2025a). Recent advances include memory-based systems (Zhang et al., 2025b), latent difference modeling (Qiu et al., 2025a,b), causal alignment (Zhao et al., 2025), and lifelong adaptation (Wang et al., 2024).

Reinforcement Learning with AI Feedback. Aligning models for personalization relies on human feedback (Ouyang et al., 2022), yet scaling diverse alignment data is challenging. Reinforcement Learning with AI Feedback (RLAIF) addresses this by using LLMs to generate and rank preference pairs (Bai et al., 2022; Lee et al., 2024), though it often suffers from low response diversity. Strategies to enhance diversity include using ensembles of different model families (Cui et al., 2024), iterative prompt and response refinement (Dong et al.), and policy gradient-based prompt updates (Zhou et al., 2025). While RLAIF typically ranks candidates via specific criteria (Yuan et al., 2024) or generation probabilities (Lee et al., 2024), standard Likert-style scoring often yields inconsistent judgments (Lee et al., 2025). To improve reliability, recent work proposes rubric-based evaluations where models answer targeted binary questions across multiple dimensions (Lee et al., 2025; Wei et al., 2025; Cook et al., 2024; Ruan et al., 2025).

7 Conclusion

This work introduces large-scale LAPD BWC annotations that capture the heterogeneous ways diverse community members interpret officer and civilian behavior. By modeling respect as an inherently subjective task, we move beyond a singular ground-truth to incorporate the divergent perspectives of justice-system-impacted, police-affiliated,

and other community annotators. We develop a domain-specific rubric grounded in community insights and a framework combining SFT with DPO. By constructing preference pairs via rubric alignment, our approach mitigates reasoning failures and improves rating accuracy and rationale quality.

Our results show that incorporating group-specific context is essential for effective alignment. Gains are most significant for justice-system-impacted annotators, highlighting the necessity of modeling divergent perspectives. Overall, this work demonstrates a framework for aligning language models to nuanced, explanation-dependent judgments, offering a pathway toward perspective-aware AI systems in high-stakes social domains.

8 Limitations

We acknowledge that while our dataset includes annotators from multiple stakeholder groups with diverse lived experiences, annotator diversity can always be expanded. In particular, additional demographic, geographic, and experiential perspectives, both within and beyond the groups considered here, may further enrich the range of interpretations captured and improve the robustness of personalized modeling. Second, our analysis focuses exclusively on the textual modality derived from body-worn camera transcripts. Although language is a central channel through which respect is communicated, nonverbal cues such as tone, prosody, facial expressions, and physical positioning also play a critical role in shaping perceptions of respect in police–civilian interactions. Incorporating audio and visual modalities remains an important avenue for our future work. Finally, our modeling and evaluation framework is grounded in a domain-specific respect rubric developed for traffic-stop encounters. While this rubric enables structured and interpretable reasoning in this context, its dimensions and criteria may not directly transfer to other forms of police–community interactions or to different institutional settings without adaptation. Extending this approach to additional domains will require careful reconsideration of rubric design and stakeholder perspectives.

9 Ethical Statement

Bias Amplification and Reification of Group Categories. Conditioning on annotator group identity enables modeling of subjective variation but risks reifying coarse group labels or reinforcing stereotypes if treated as essential. We emphasize that these groups serve as analytic proxies for lived experience, not fixed representations, and that substantial within-group heterogeneity remains.

Fairness and Differential Stakeholder Impact. By engaging diverse stakeholders and reflecting their perspectives in the AI tools we develop, we are working explicitly to increase fairness and stakeholder representation in this field.

Privacy and Data Sensitivity. This work relies on sensitive police body-worn camera data. All annotations were conducted within secure, on-premises infrastructure, and we obtain informed consent from annotators to study their annotations. We paid annotators wages ranging from $17.28 per hour for undergraduate students to $30 per hour for more experienced professionals. Wages were based on experience level. Consistent with hiring practices at [university anonymized], annotators were recruited through a range of measures, including the posting of job listings on a range of public platforms. The study was approved by the IRB Board at [university anonymized].

Data sharing is strictly constrained by contractual agreements with LAPD. We do not release raw video, audio, or unredacted transcripts, and any disseminated artifacts are anonymized and access-controlled. Despite these safeguards, modeling real-world interactions involving vulnerable individuals carries inherent privacy risks.

Safeguards and Mitigation Strategies. We take several steps to mitigate these risks: (1) framing respect as inherently subjective and rejecting a single ground truth; (2) evaluating models using rubric-based, interpretable criteria rather than opaque scores; (3) emphasizing human-in-the-loop analysis rather than automated decision-making; and (4) limiting claims about deployment. Future work could explore gated release mechanisms and participatory governance with community stakeholders.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. In *Interspeech 2023*, pages 4489–4493.

Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. 2020. Studying up: reorienting the study of algorithmic fairness around issues of power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 167–176.

Jessica Y Bo, Tianyu Xu, Ishan Chatterjee, Katrina Passarella-Ward, Achin Kulshrestha, and D Shin. 2025. Steerable chatbots: Personalizing llms with preference-based activation steering. *arXiv preprint arXiv:2505.04260*.

Nicholas P Camp and Rob Voigt. 2024. Body camera footage as data: Using natural language processing to monitor policing at scale & in depth. *Behavioral Science & Policy*, 10(2):16–25.

Nicholas P Camp, Rob Voigt, MarYam G Hamedani, Dan Jurafsky, and Jennifer L Eberhardt. 2024. Leveraging body-worn camera footage to assess the effects of training on officer communication during traffic stops. *PNAS nexus*, 3(9):pgae359.

Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. 2024. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization. *Advances in Neural Information Processing Systems*, 37:49519–49551.

Jonathan Cook, Tim Rocktäschel, Jakob Nicolaus Foerster, Dennis Aumiller, and Alex Wang. 2024. Ticking all the boxes: Generated checklists improve llm evaluation and generation. In *Language Gamification-NeurIPS 2024 Workshop*.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, and 1 others. 2024. Ultrafeedback: Boosting language models with scaled ai feedback. In *ICML*.

Yuhao Dan, Jie Zhou, Qin Chen, Junfeng Tian, and Liang He. 2025. P-react: Synthesizing topic-adaptive reactions of personality traits via mixture of specialized lora experts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6342–6362.

Sumanth Doddapaneni, Krishna Sayana, Ambarish Jash, Sukhdeep Sodhi, and Dima Kuzmin. 2024. User embedding model for personalized language prompting. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 124–131, St. Julians, Malta. Association for Computational Linguistics.

Qingxiu Dong, Li Dong, Xingxing Zhang, Zhifang Sui, and Furu Wei. Self-boosting large language models with synthetic preference data. In *The Thirteenth International Conference on Learning Representations*.

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

Jerry Zhi-Yang He, Sashrika Pandey, Mariah L Schrum, and Anca Dragan. 2024. Context steering: Controllable personalization at inference time. *arXiv preprint arXiv:2405.01768*.

Liam Hebert, Krishna Sayana, Ambarish Jash, Alexandros Karatzoglou, Sukhdeep S Sodhi, Sumanth Doddapaneni, Yanli Cai, and Dima Kuzmin. 2024. Persoma: Personalized soft prompt adapter architecture for personalized language prompting. *CoRR*.

Parsa Hejabi, Akshay Kiran Padte, Preni Golazizian, Rajat Hebbar, Jackson Trager, Georgios Chochlakis, Aditya Kommineni, Ellie Graeden, Shrikanth Narayanan, Benjamin AT Graham, and 1 others. 2024. Cvat-bwv: A web-based video annotation platform for police body-worn video. In *International Joint Conferences on Artificial Intelligence Organization*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Bernd Huber, Ghazal Fazelnia, Andreas Damianou, Sebastian Peleato, Max Lefarov, Praveen Ravichandran, Marco De Nadai, Mounia Lalmas-Roellke, and Paul N Bennett. 2025. Embedding-to-prefix: Parameter-efficient personalization for pretrained large language models. *arXiv preprint arXiv:2505.17051*.

Jonathan Jackson, Ben Bradford, Mike Hough, Andy Myhill, Paul Quinton, and Tom R Tyler. 2012. Why do people comply with the law? legitimacy and the influence of legal institutions. *British journal of criminology*, 52(6):1051–1071.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*.

Xiaoyu Kong, Jiancan Wu, An Zhang, Leheng Sheng, Hui Lin, Xiang Wang, and Xiangnan He. 2024. Customizing language models with instance-wise lora for sequential recommendation. *Advances in Neural Information Processing Systems*, 37:113072–113095.

Erick Laming. 2019. Police use of body worn cameras. *Police practice and research*, 20(2):201–216.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and 1 others. 2024. Rlaif vs. rlhf: scaling reinforcement learning from human feedback with ai feedback. In

Proceedings of the 41st International Conference on Machine Learning, pages 26874–26901.

Yukyung Lee, JoongHoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. 2025. CheckEval: A reliable LLM-as-a-judge framework for evaluating text generation using checklists. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15782–15809, Suzhou, China. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Xinyu Li, Ruiyang Zhou, Zachary C Lipton, and Liu Leqi. 2024. Personalized language modeling from personalized human feedback. *arXiv preprint arXiv:2402.05133*.

Jiahong Liu, Zexuan Qiu, Zhongyang Li, Quanyu Dai, Wenhao Yu, Jieming Zhu, Minda Hu, Menglin Yang, Tat-Seng Chua, and Irwin King. 2025a. A survey of personalized large language models: Progress and future directions. *arXiv preprint arXiv:2502.11528*.

Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. 2025b. LLMs + persona-plug = personalized LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9373–9385, Vienna, Austria. Association for Computational Linguistics.

Guodong Long, Tao Shen, Jing Jiang, Michael Blumenstein, and 1 others. 2024. Dual-personalizing adapter for federated foundation models. *Advances in Neural Information Processing Systems*, 37:39409–39433.

Lorraine Mazerolle, Sarah Bennett, Jacqueline Davis, Elise Sargeant, and Matthew Manning. 2013. Procedural justice and police legitimacy: a systematic review of the research evidence. *Journal of Experimental Criminology*, 9(3):245–274.

Daniel S Nagin and Cody W Telep. 2017. Procedural justice and legal compliance. *Annual review of law and social science*, 13(1):5–28.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Justin T Pickett, Amanda Graham, and Francis T Cullen. 2022. The american racial divide in fear of the police. *Criminology*, 60(2):291–320.

Emma Pierson, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, Phoebe Barghouty, Cheryl Phillips, Ravi Shroff, and 1 others. 2020. A large-scale analysis of racial disparities in police stops across the united states. *Nature human behaviour*, 4(7):736–745.

Yilun Qiu, Tianhao Shi, Xiaoyan Zhao, Fengbin Zhu, Yang Zhang, and Fuli Feng. 2025a. Latent inter-user difference modeling for LLM personalization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10610–10628, Suzhou, China. Association for Computational Linguistics.

Yilun Qiu, Xiaoyan Zhao, Yang Zhang, Yimeng Bai, Wenjie Wang, Hong Cheng, Fuli Feng, and Tat-Seng Chua. 2025b. Measuring what makes you unique: Difference-aware user modeling for enhancing llm personalization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21258–21277.

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.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Jerome Ramos, Bin Wu, and Aldo Lipani. 2024. Peapod: Personalized prompt distillation for generative recommendation. *arXiv preprint arXiv:2407.05033*.

Jie Ruan, Inderjeet Nair, Shuyang Cao, Amy Liu, Sheza Munir, Micah Pollens-Dempsey, Tiffany Chiang, Lucy Kates, Nicholas David, Sihan Chen, and 1 others. 2025. Expertlongbench: Benchmarking language models on expert-level long-form generation tasks with structured checklists. *arXiv preprint arXiv:2506.01241*.

Michael J. Ryan, Omar Shaikh, Aditri Bhagirath, Daniel Frees, William Held, and Diyi Yang. 2025. SynthesizeMe! inducing persona-guided prompts for personalized reward models in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8045–8078, Vienna, Austria. Association for Computational Linguistics.

Krishna Sayana, Raghavendra Vasudeva, Yuri Vasilevski, Kun Su, Liam Hebert, James Pine, Hubert Pham, Ambarish Jash, and Sukhdeep Sodhi. 2025. Beyond retrieval: Generating narratives in conversational recommender systems. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2411–2420.

Michael Sierra-Arévalo, Jose Alcocer, Lauren Brown, Raquel Delerme, Benjamin AT Graham, Harry Muttram, Jackson Trager, and Nicholas Weller. 2025. Police as policymakers: How experiences with policy implementation shape policy representation.

Hans Myhre Sunde, Don Weenink, and Marie Rosenkrantz Lindegaard. 2023. Revisiting the demeanour effect: a video-observational analysis of encounters between law enforcement officers and citizens in amsterdam. *Policing and society*, 33(8):953–969.

Jason Sunshine and Tom R. Tyler. 2003a. The Role of Procedural Justice and Legitimacy in Shaping Public Support for Policing. *Law & Society Review*, 37(3):513–547.

Jason Sunshine and Tom R Tyler. 2003b. The role of procedural justice and legitimacy in shaping public support for policing. *Law & society review*, 37(3):513–547.

Zhaoxuan Tan, Zheyuan Liu, and Meng Jiang. 2024. Personalized pieces: Efficient personalized large language models through collaborative efforts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6459–6475, Miami, Florida, USA. Association for Computational Linguistics.

Jackson Trager, Kerby Bennett, Lauren Brown, Georgios Chochlakis, Morteza Dehghani, Aarya Devnani, Raquel Delerme, Brittany Friedman, Ellie Graeden, Benjamin A. T. Graham, Preni Golazizian, Rajat Hebbar, Parsa Hejabi, Aditya Komminen, Lateefah Mirza, Oliver Chavez, Akshay Kiran Padte, Michael Sierra-Arévalo, Nicholas Weller, and Shrikanth Narayanan. 2024. Everyday respect project phase 0 manual: Improving automated transcripts. Technical report, Open Science Framework. Annotation manual for the Everyday Respect Project.

Nicol Turner Lee. 2018. Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society*, 16(3):252–260.

Tom Tyler. 2017. Procedural justice and policing: A rush to judgment? *Annual review of law and social science*, 13(1):29–53.

Tom R Tyler. 1988. What is procedural justice?: Criteria used by citizens to assess the fairness of legal procedures. *Law & society review*, 22(1):103–135.

Tom R Tyler and Yuen J Huo. 2002. *Trust in the law: Encouraging public cooperation with the police and courts*. Russell Sage Foundation.

Rob Voigt, Nicholas P Camp, Vinodkumar Prabhakaran, William L Hamilton, Rebecca C Hetey, Camilla M Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L Eberhardt. 2017. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, 114(25):6521–6526.

Tiannan Wang, Meiling Tao, Ruoyu Fang, Huilin Wang, Shuai Wang, Yuchen Eleanor Jiang, and Wangchunshu Zhou. 2024. Ai persona: Towards life-long personalization of llms. *arXiv preprint arXiv:2412.13103*.

Vesla M Weaver and Amy E Lerman. 2010. Political consequences of the carceral state. *American Political Science Review*, 104(4):817–833.

Tianjun Wei, Wei Wen, Ruizhi Qiao, Xing Sun, and Jianghong Ma. 2025. Rocketeval: Efficient automated LLM evaluation via grading checklist. In *The Thirteenth International Conference on Learning Representations*.

Robert E Worden and Sarah J McLean. 2018. Measuring, managing, and enhancing procedural justice in policing: Promise and pitfalls. *Criminal justice policy review*, 29(2):149–171.

Wenfei Xu, Michael Smart, Nebiyou Tilahun, Sajad Askari, Zachary Dennis, Houpu Li, and David Levinson. 2024. The racial composition of road users, traffic citations, and police stops. *Proceedings of the National Academy of Sciences*, 121(24):e2402547121.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. 2024. Self-rewarding language models. In *Forty-first International Conference on Machine Learning*.

Jinghao Zhang, Yuting Liu, Wenjie Wang, Qiang Liu, Shu Wu, Liang Wang, and Tat-Seng Chua. 2025a. Personalized text generation with contrastive activation steering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7128–7141, Vienna, Austria. Association for Computational Linguistics.

Xinliang Frederick Zhang, Nick Beauchamp, and Lu Wang. 2025b. Prime: large language model personalization with cognitive memory and thought processes. *arXiv e-prints*, pages arXiv–2507.

You Zhang, Jin Wang, Liang-Chih Yu, Dan Xu, and Xuejie Zhang. 2024. Personalized lora for human-centered text understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19588–19596.

Siyan Zhao, John Dang, and Aditya Grover. 2023. Group preference optimization: Few-shot alignment of large language models. *arXiv preprint arXiv:2310.11523*.

Xiaoyan Zhao, Juntao You, Yang Zhang, Wenjie Wang, Hong Cheng, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. 2025. Nextquill: Causal preference modeling for enhancing llm personalization. *arXiv preprint arXiv:2506.02368*.

Yiyang Zhou, Zhaoyang Wang, Tianle Wang, Shangyu Xing, Peng Xia, Bo Li, Kaiyuan Zheng, Zijian Zhang, Zhaorun Chen, Wenhao Zheng, and 1 others. 2025. Anyprefer: An agentic framework for preference data synthesis. In *ICLR*.

Jiachen Zhu, Jianghao Lin, Xinyi Dai, Bo Chen, Rong Shan, Jieming Zhu, Ruiming Tang, Yong Yu, and Weinan Zhang. 2024. Lifelong personalized low-rank adaptation of large language models for recommendation. *arXiv preprint arXiv:2408.03533*.

A Dataset details

The sampled stops statistics is shown in Table 2. To operationalize the annotation process, we design a domain-specific, secure, on-premises annotation software and pipeline. This custom solution is necessitated both by the complex task demands and by strict security constraints that prohibited the raw BWV data from leaving LAPD headquarters, where all annotations are completed (Hejabi et al., 2024).

A.1 Annotator Training and Data Quality Control

Given the socially sensitive and high-stakes nature of police–civilian interactions, we place strong emphasis on annotator training and data quality.

All annotators underwent structured, multi-phase training using detailed, domain-specific annotation guidelines developed in collaboration with subject-matter experts. The guideline itself was iteratively refined through extensive fieldwork, including interviews, surveys, and engagement with LAPD training materials.

Each annotation phase (Phases 0, 1, 2A, 2B) consisted of a 1.5–2 hour training session followed by a practice task in which annotators independently annotated a sample video. Annotators then received feedback on their annotations before proceeding to annotate study data. This training-and-feedback process was repeated for each phase to ensure consistency across tasks.

Annotators were provided with comprehensive training manuals for each phase, including detailed examples and a glossary of relevant terms, which remained available as reference materials during annotation.

To maintain annotation quality over time, annotations were periodically reviewed in a blinded manner. These reviews were conducted approximately every 4–6 weeks in collaboration with LAPD to ensure that annotators continued to produce consistent and high-quality annotations after training.

Data Sharing and Reproducibility. Our contractual agreement with LAPD permits the release of anonymized transcripts and associated annotations. These resources, along with preprocessing scripts, prompts, and evaluation code, will be publicly available at: `https://github.com/preniJee/ER-ACL`. Due to legal and privacy constraints, raw body-worn camera video and audio cannot be released.

A.2 Preference Pairs

We construct preference pairs using *generator* and *augmenter* modules as introduced in the main paper, which leverages ground-truth rationales, and their paraphrased variants as chosen samples. For paraphrasing, we employ `Qwen/Qwen3-Next-80B-A3B-Instruct` (Yang et al., 2025) with temperature 0.7, top-$p = 1.0$, and a maximum generation length of 2000 tokens to generate three semantically equivalent paraphrases for each ground-truth rationale while preserving the original respect rating and rubric-dimensional structure. The paraphrasing prompt is shown in Figure 7. Chosen samples consist of the original ground-truth rationales and their paraphrased versions, ensuring that all chosen samples maintain perfect rating accuracy. Rejected samples are selected based on the rubric-alignment quality metrics described in the Section 3.3. Specifically, we reject model-generated rationales that exhibit any of the following failure modes: (i) binary precision below 0.5, (ii) binary recall below 0.5, or (iii) false-positive activations on target emotional-respect dimensions, namely `emotional_respect.respectful.empathy` or `emotional_respect.respectful.warmth`. For each chosen sample (ground truth or paraphrase), we construct up to 5 preference pairs by randomly sampling from the pool of valid rejected candidates, enforcing a minimum rubric F_1 score gap of 0.2 between chosen and rejected outputs. These thresholds and pairing constraints correspond to the final configuration selected after empirically evaluating multiple alternative settings, and together result in a dataset of $10{,}612$ rubric-grounded preference

pairs. This approach results in a dataset where chosen samples demonstrate high alignment with ground truth annotations (near-perfect F_1 for paraphrases), while rejected samples demonstrate substantially lower alignment, with average precision of 0.41 and recall of 0.51, creating clear preference signals for alignment training.

Stops Statistic	**Count (%)**
Number of stops	1008 (–)
Timeframe of stops	–
Total duration (hours)	696.7 (–)
Mean duration (minutes)	14.9 (SD = 13.7)
Reason	
Traffic Violation	990 (82.8%)
Reasonable suspicion of crime	173 (14.5%)
Parole/Probation	16 (1.3%)
Arrest Warrant/Wanted Person	14 (1.2%)
Consensual encounter w/search	2 (0.2%)
Possible Danger to Self&Others/5150	1 (0.1%)
Action taken	
Search of person conducted	520 (43.5%)
None	452 (37.8%)
Handcuffed/flex cuffed	415 (34.7%)
Search of property conducted	393 (32.9%)
Ordered from vehicle	361 (30.2%)
Curbside detention	265 (22.2%)
Req. consent to search property	146 (12.2%)
Patrol car detention	131 (11.0%)
Req. consent to search person	79 (6.6%)
Vehicle impounded	67 (5.6%)
Firearm pointed at person	40 (3.3%)
Field sobriety test	32 (2.7%)
Property was seized	31 (2.6%)
Person photographed	9 (0.8%)
Physically removed from vehicle	3 (0.3%)
Result	
Warning	369 (30.9%)
Citation for infraction	348 (29.1%)
FI card completed	236 (19.7%)
No action	207 (17.3%)
Arrest w/o warrant	164 (13.7%)
Warrant arrest	24 (2.0%)
In-field cite and release	8 (0.7%)
Psychiatric hold	3 (0.3%)

Table 2: Summary statistics of the LAPD traffic stop videos.

B Respect Rubric

The prompts used for the LLM-as-a-judge to evaluate the rationales based on the respect rubric for officer and driver are shown in Figures 4 and 5 respectively.

C Task prompts

Example of the prompt used for training and evaluation of our models is shown in Figure 6

D Subjectivity in Respect Annotations

To investigate the subjectivity inherent in respect ratings during police stops, we employed mixed-effect models that incorporate both fixed and random effects. These models assess how annotator-related variables contribute to variation in respect ratings. Our primary focus is on the identity and experience of the annotators, particularly whether they have previously been incarcerated or affiliated with law enforcement, as well as their demographics (race, gender, age). We also include perceived demographic and socioeconomic attributes of the drivers and officers, such as race, gender, age, height, presence of a foreign accent, clothing, and car type, as control variables to account for potential confounding influences and to ensure that the estimated effects of annotator group identity are not driven by differences in perceived characteristics of the individuals involved.

Police officers. Annotator group identity was the single significant predictor of perceived officer respect among annotator background attributes. Justice system-impacted annotators rated officers as more respectful than non-affiliated annotators ($\beta = 0.553, p = .002, CI = [0.206, 0.900]$), and police-affiliated annotators also provided higher officer-respect ratings than non-affiliated annotators ($\beta = 0.526, p = .007, CI = [0.141, 0.911]$). No other annotator-level demographic variables (e.g., race, gender, age) significantly influenced judgments, underscoring that lived experience with the criminal justice system was the primary factor shaping evaluations of police behavior.

Controlling for perceived officer and driver characteristics did not change the observed effect of annotator group on respect annotations. Overall, officers were judged as less respectful when they were perceived to have a foreign accent ($\beta = -1.358, p < .001, CI = [-2.031, -0.685]$). Several driver attributes also shaped how respectful the officer's behavior appeared: officers were rated as more respectful when the driver was female ($\beta = 0.282, p = .002, CI = [0.107, 0.458]$), when the driver was short ($\beta = 0.397, p = .007, CI = [0.110, 0.684]$), and when the driver was perceived to be driving a middle-class car, compared to a working-class car ($\beta = 0.265, p = .020, CI = [0.042, 0.487]$). Officers were judged as less respectful when the driver spoke with a foreign accent ($\beta = -0.344, p = .004, CI = [-0.581, -0.108]$).

Drivers. Similarly, annotator group identity emerged as the single significant predictor of perceived driver respect among annotator background attributes. Police-affiliated annotators rated drivers as significantly more respectful than non-affiliated annotators ($\beta = 0.837, p = .010, CI = [0.203, 1.470]$), whereas we found no evidence of a difference between formerly incarcerated annotators and the non-affiliated group ($\beta = 0.174, p = .588, CI = [-0.457, 0.805]$). Individual annotator demographics (race, gender, age) were not predictive of the annotators' respect judgment, indicating that group identity remained the primary factor shaping perceptions of driver behavior.

Controlling for perceived officer and driver characteristics did not change the observed effect of annotator group on respect annotations. Overall, a small number of stop-level characteristics also influenced perceived driver respect. Drivers perceived as Hispanic were rated as more respectful than those perceived as White ($\beta = 0.320, p = .012, CI = [0.071, 0.569]$). Several officer-related attributes also shaped judgments: when the officer was perceived as Asian, drivers were rated as less respectful ($\beta = -0.728, p = .018, CI = [-1.331, -0.124]$), and when the officer was perceived to have a foreign accent, drivers were judged as substantially less respectful ($\beta = -1.513, p < .001, CI = [-2.137, -0.889]$).

Discussion. Across both officer and driver evaluations, annotator group membership, particularly having been incarcerated or affiliated with law enforcement, stood out as the sole annotator-level factor associated with systematic differences in respect annotations. Interestingly, we observed an inverse pattern of expectations: former officers rated drivers as more respectful, while formerly incarcerated individuals rated officers as more respectful, relative to the non-affiliated group. This counters the intuition that shared background or adversarial experience should yield negative evaluations and instead underscores the complex, subjective nature of respect judgments.

In contrast, demographic characteristics of annotators (e.g., race, gender, age) did not significantly predict respect ratings. Incorporating perceived demographic and socioeconomic attributes of officers and drivers as control variables revealed additional context-specific patterns, but did not alter the observed effect of annotator group identity on respect annotations. Together, these findings indicate that experiential background related to the criminal justice system plays a central role in shaping perceptions of respect in police–civilian encounters, beyond what can be explained by demographic characteristics alone.

E Implementation Details

We use Qwen2.5-7B-Instruct (Qwen et al., 2025) as the base model in our experiments. We fine-tune Qwen2.5-7B-Instruct using LoRA, and then, using Direct Preference Optimization (DPO) with parameter-efficient LoRA adapters. Our implementation builds on the HuggingFace `transformers` and `trl` libraries, with additional infrastructure for custom checkpoint selection, group-aware inference, and training-time evaluation.

Data are split at the conversation level into 80/10/10 train/validation/test partitions to avoid transcript leakage. Fine-tuning is performed with the TRL `DPOTrainer` using the standard sigmoid DPO loss, optimized with AdamW and a learning rate of 1×10^{-7}. The effective batch size is 32 (batch size 4 with gradient accumulation of 8), and training proceeds for up to 5 epochs with evaluation every 10 steps. All evaluation checkpoints are saved for post-hoc scoring, and the best model is selected based on validation-set rating MAE.

During inference, we use sampling config parameters `temperature` = 0.5, `top_p` = 0.85, and `top_k` = 30. All evaluation metrics are computed both overall and per annotator group.

All experiments are conducted on NVIDIA H200 GPUs. We fix the random seed to 42 for reproducibility, and we store hyperparameters, configuration files, and training logs with each checkpoint. Training traces and experiment metadata are logged to Weights & Biases.

F Rationales' Predictiveness of Ratings

We empirically validate that the rationales provided by the annotators are actually reflective of their ratings they provide for respect. To do so, we condition LLMs on the rationale, ρ_{ij}, and predict the corresponding rating, r_{ij}.

Because some of the rationales explicitly mention the ratings, e.g., "I rate this behavior as *respectful*", or "I rated the officer's respect a 4.", we preprocess the rationales to remove any mention of the label names or their Likert value, and replace these mentions with `[MUTE]`. In this manner, we focus on the reasons provided for the rating, and

Category	Hyperparameter	Value
Base Model	Model	Qwen2.5–7B
	Precision	bf16 / f16
Training	Learning rate	1×10^{-7}
	Optimizer	AdamW
	Batch size	4
	Grad. accum.	8
	Eff. batch size	32
	Epochs	5
	Eval freq.	Every 10 steps
LoRA	Rank	32
	Target modules	Linear layers
	α	64
DPO	Loss type	Sigmoid
	Beta (β)	0.1
	Ref. model	Frozen base
	Truncation	keep_end
Data	Split	80/10/10
	Split level	Conversation
	Augment.	Synthetic pairs

Table 3: Hyperparameters for all experiments.

therefore use the LLMs as models for the reasoning of the annotators, rather than merely tool to retrieve the rating from the rationale.

We experiment with 5 models, GPT-OSS-20B[1], Llama-3.1-8B 8B-Instruct[2], Llama-3.3-70B-Instruct[3], and Qwen3-4B-Instruct[4] and Qwen3-30B Instruct[5]. We set the reasoning effort of the GPT-OSS models to *low*, with a computing budget of 100 tokens. All models are prompted with 5-shot prompts, and each experiment is repeated 5 times, randomly resampling the demonstrations in the prompts from a small held-out training set. Example 1-shot prompts can be seen in Figure 8.

Table 4 shows the Mean Absolute Error (MAE) with 95% confidence intervals after converting the predictions of the LLM back to the original 1-5 Likert scale. Note that models are about 60% accurate in predicting the level of respect assigned by the annotator. By looking at their MAE, we see that when they err, they usually predict neighboring values. For reference, we note that without redaction of explicit mentions of the ratings from the rationales, Llama-3.1-8B-Instruct achieved 66.1% accuracy and an MAE of 0.366, significantly better than with the redacted version. We conclude that rationales capture the valence of the respect effectively.

[1] https://huggingface.co/openai/gpt-oss-20b

[2] https://huggingface.co/meta-llama/llama-3.1-8b-instruct

[3] https://huggingface.co/meta-llama/llama-3.3-70b-instruct

[4] https://huggingface.co/qwen/qwen3-4b-instruct-2507

[5] https://huggingface.co/qwen/qwen3-30b-a3b-instruct-2507

	MAE	
Model	**Officer**	**Civilian**
GPT-OSS-20B	$0.547_{\pm 0.016}$	$0.529_{\pm 0.011}$
Llama-3.1-8B-Instruct	$0.455_{\pm 0.011}$	$0.462_{\pm 0.007}$
Llama-3.3-70B-Instruct	$0.404_{\pm 0.012}$	$0.424_{\pm 0.005}$
Qwen3-4B-Instrcut	$0.459_{\pm 0.009}$	$0.449_{\pm 0.008}$
Qwen3-30B-Instruct	$0.420_{\pm 0.006}$	$0.439_{\pm 0.006}$

Table 4: Mean Absolute Error (*MAE*) when predicting the respect rating of each annotator based on their provided rationale.

G Additional Analysis

G.1 Error Case Analysis

To better understand the limitations of our best-performing model ($DPO_{rub}^{\pi_{sft}}$), we conduct a structured error analysis along two axes: (i) rating prediction errors and (ii) rationale generation errors grounded in the respect rubric.

G.1.1 Rating Prediction Errors

We analyze model performance across different rating levels for both officers and drivers using mean absolute error (MAE). Table 5 summarizes the worst-performing categories.

Entity	Worst Rating	MAE
Officer	1 (Very Disrespectful)	3
Driver	3 (Neutral)	0.952

Table 5: Rating prediction error analysis for worst-performing categories.

These results indicate that the model struggles most with (i) extreme negative officer behaviors and (ii) ambiguous driver interactions, suggesting difficulty in capturing both rare edge cases and inherently subjective mid-spectrum judgments.

G.1.2 Rationale Generation Errors

We further analyze errors in generated rationales using the rubric-based evaluation framework (§3.2), focusing on dimensions with the lowest macro F1 scores.

Officer Rationales. For officer rationales, the model most frequently fails to capture *communicative respect (disrespectful)* dimensions, with an aggregated macro F1 of 0.471. The lowest-performing subdimensions are shown in Table 6.

Subdimension	Category	Macro F1
lack_of_explanation	communicative_respect	0.460
lack_of_options_next_steps	communicative_respect	0.464
lack_of_reason_ask	communicative_respect	0.466

Table 6: Lowest-performing rubric dimensions for officer rationale generation.

Driver Rationales. For driver rationales, the model exhibits a similar pattern, with the lowest performance again concentrated in communicative respect (disrespectful) dimensions (macro F1 = 0.496). The lowest-performing subdimensions are shown in Table 7.

Subdimension	Category	Macro F1
interrupts	communicative_respect	0.494
aggressive_opening	professional_respect	0.496
non_responsive	communicative_respect	0.498

Table 7: Lowest-performing rubric dimensions for driver rationale generation.

Across both entities, the model consistently under-detects or underrepresents *disrespectful communicative behaviors*, particularly those involving missing explanations, lack of dialogue, and conversational breakdown (e.g., interruptions or non-responsiveness).

G.2 Validation of LLM-as-a-Judge

Given the high-stakes nature of evaluating respect in police–civilian interactions, we assess the reliability of our LLM-as-a-judge framework through human validation.

To evaluate the robustness of the judge, we conduct human validation on a random 10% subset of the dataset. We evaluate officer and driver rationales separately. For each rationale, both the LLM judge and human annotators assign binary labels over the full set of rubric dimensions (see Appendix B). We compute agreement at the dimension level by comparing the binary activations produced by the LLM judge and human annotators. Agreement is then averaged across all rubric dimensions to obtain a macro-average agreement score for each entity. Table 8 reports the agreement between human annotations and the LLM-as-a-judge.

Entity	Macro-Average Agreement
Officer	96.1%
Driver	96.7%

Table 8: Agreement between human annotators and the LLM-as-a-judge across rubric dimensions.

These results indicate high consistency between the LLM judge and human annotators across both entities, suggesting that the judge reliably captures rubric-grounded reasoning.

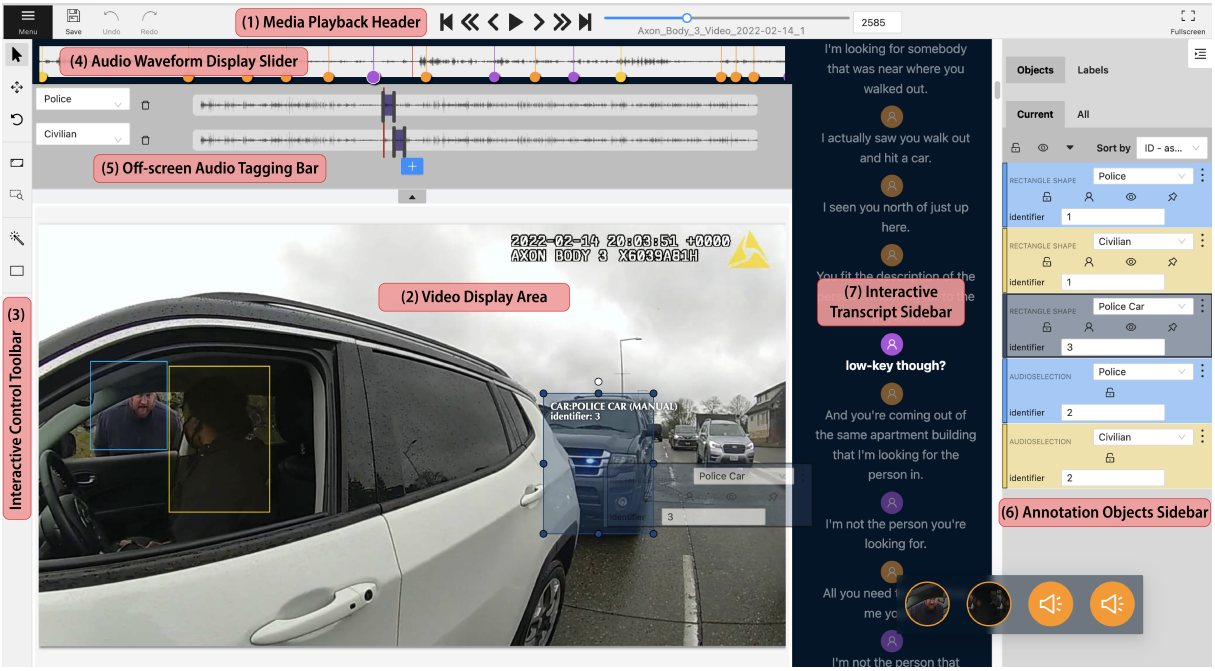

Figure 3: Our annotation platform developed for annotating BWV data.

Officer respect rubric prompt

You are tasked with evaluating rationales explaining ratings of officer respect in police traffic stops. Use the rubric below to guide your evaluation. The rubric describes four major elements of respect: Emotional Respect, Professional Respect, Communicative Respect, and Contextual Moderators of Respect. These elements are not mutually exclusive or hierarchical. Instead, they capture overlapping aspects of officer behavior that annotators may draw on when explaining their ratings.
Your role is to judge whether each of the respect dimensions are referenced in the provded rationale.

1. Emotional Respect
Emotional respect refers to how the officer's language, tone, and demeanor convey care, warmth, or hostility. It highlights the affective dimension of respectful or disrespectful communication.

Respectful Emotional Descriptives
Warmth :The officer conveys friendliness or kindness.
Example: "The officer smiled and used a polite tone."
Empathy: The officer acknowledges or validates the civilian's perspective or feelings.
Example: "The officer said they understood why the driver was nervous."
Apology/Thanks: The officer apologizes for inconvenience or expresses gratitude.
Example: "They apologized for the delay."

Disrespectful Emotional Descriptives
Lack of Warmth: The officer is cold, distant, or unfriendly.
Lack of Empathy: The officer doesn't show any understanding or concern.
Example: "The officer was not empathetic"
Lack of Apology/Thanks: The officer doesn't apologize or thank.
Offensiveness/Bias: The officer uses insulting, hateful, or biased language.
Example: "They implied bias by questioning whether the driver 'belonged here.'"
Unnecessary Escalation: The officer amplifies tension rather than calming the situation.
Example: "They raised their voice aggressively even though the driver was compliant."
Disrespect for Time: The officer prolongs the stop without clear justification.
Example: "The driver was left waiting on the curb for an extended period without explanation."

2. Professional Respect
Professional respect reflects formality, neutrality, and composure. It concerns whether the officer behaves in a manner consistent with professional standards.

Respectful Professionalism
Greeting: The officer begins with a polite greeting rather than a command.
Introduction: The officer introduces themselves or their department.
Professional Language: The officer avoids slang, sarcasm, or mocking tones.
Example: "They used clear, formal language."
Composure/Deflection: The officer maintains calm and redirects tension.
Example: "They calmly de-escalated when the driver grew agitated."

Disrespectful Unprofessionalism
Order Opening: The officer starts with an order instead of a greeting.
Example: "They asked for license and registration, without greeting."
Non-Introduction: The officer fails to identify themselves or their department.
Example: "They never told the driver who they were."
Unprofessional Tone or Language: The officer uses sarcasm, slang, or mockery.
Example: "They laughed at the driver's explanation."

3. Communicative Respect
Communicative respect centers on whether the officer provides transparency, voice, and fairness in the interaction.

Respectful Communicative Dialogue
Reason Ask: The officer invites the civilian to share their perspective.
Example: "The officer asked the driver to explain what happened."
Reason Given: The officer allows the civilian to provide their own account.
Example: "The driver explained they were late to pick up their child."
Explanation: The officer clarifies the reason for the stop or decision.
Example: "They explained the traffic law that was violated."
Options/Next Steps: The officer informs the civilian of their choices.
Example: "The officer said you can contest this ticket in court."
Comprehension Check: The officer ensures the civilian understands.
Example: "The officer made sure the driver unserstood why they were pulled over"
Free to Leave: The officer explicitly signals when the encounter is over.
Example: "The officer signaled the driver is free to go, and said drive safely."

Disrespectful Communicative Dialogue
Lack of Reason Ask: The officer doesn't ask for the civilian's perspective.
Lack of Reason Given: The officer doesn't allow explanation.
Lack of Explanation: The officer doesn't explain the reason for the stop.
Lack of Options/Next Steps: The officer doesn't inform about choices.
Lack of Comprehension Check: The officer doesn't verify understanding.
Lack of Free to Leave: The officer doesn't state that the encounter is over.
Interrupts: The officer cuts off or talks over the civilian.
Example: "The driver tried to explain, but the officer repeatedly interrupted."

4. Contextual Moderators of Respect
Contextual moderators are situational factors that shape how respect is expressed or constrained. They are not inherently respectful or disrespectful but help explain why annotators may justify ratings differently.

Threat
Threaten Violence: The civilian makes verbal or physical threats.
Example: "The driver threatened to harm the officer."
Non-Visible Hands: The civilian hides or refuses to show hands.
Example: "The officer repeatedly asked to see the driver's hands."
Movement Resistance: The civilian resists physical or verbal compliance.
Example: "They refused to step out of the vehicle."

Disruptiveness
Yelling: The civilian shouts over the officer.
Example: "The driver kept screaming, preventing any dialogue."
Extreme Interruptions: The civilian persistently interrupts.
Example: "They wouldn't let the officer finish a sentence."
Environmental Distractions: Noise or activity in the setting impedes communication.
Example: "Passing cars and shouting bystanders made it impossible to hear."

(a) Part I.

RATIONALE TO EVALUATE:
{rationale}

INSTRUCTIONS FOR OUTPUT

After reading the rationale, evaluate it against the rubric above.
For each sub-dimension, answer the following question:
Is this aspect being referenced in the rationale, either explicitly or implicitly implied?
Respond with "yes" or "no" for each field.

When answering:

- Treat respectful and disrespectful dimensions as separate and independent.
- Mark "yes" under a respectful field only if the rationale explicitly mentions the presence of that positive behavior.
- Mark "yes" under a disrespectful field only if the rationale explicitly mentions the absence or violation of that behavior.
- Do not mark both as "yes" unless the rationale clearly describes both respectful and disrespectful moments.
- If the rationale does not reference that aspect at all, mark "no" in both.

You must produce only valid JSON that conforms exactly to the schema below.

```
Schema :
{
  "emotional_respect": {
    "respectful": {
      "warmth": "",
      "empathy": "",
      "apology_thanks": ""
    },
    "disrespectful": {
      "lack_of_warmth": "",
      "lack_of_empathy": "",
      "lack_of_apology_thanks": "",
      "offensiveness_bias": "",
      "unnecessary_escalation": "",
      "disrespect_for_time": ""
    }
  },
  "professional_respect": {
    "respectful": {
      "greeting": "",
      "introduction": "",
      "professional_language": "",
      "composure_deflection": ""
    },
    "disrespectful": {
      "order_opening": "",
      "non_introduction": "",
      "unprofessional_tone_language": ""
    }
  },
  "communicative_respect": {
    "respectful": {
      "reason_ask": "",
      "reason_given": "",
      "explanation": "",
      "options_next_steps": "",
      "comprehension_check": "",
      "free_to_leave": ""
    },
    "disrespectful": {
      "lack_of_reason_ask": "",
      "lack_of_reason_given": "",
      "lack_of_explanation": "",
      "lack_of_options_next_steps": "",
      "lack_of_comprehension_check": "",
      "lack_of_free_to_leave": "",
      "interrupts": ""
    }
  },
  "contextual_moderators": {
    "threat": {
      "threaten_violence": "",
      "non_visible_hands": "",
      "movement_resistance": ""
    },
    "disruptiveness": {
      "yelling": "",
      "extreme_interruptions": "",
      "environmental_distractions": ""
    }
  }
}
```

(b) Part II.

Figure 4: Rubric and instructions used by the LLM-as-a-judge to evaluate officer-respect rationales and output structured rubric activations.

Driver respect rubric prompt

You are tasked with evaluating rationales explaining ratings of civilian respect in police traffic stops. Use the rubric below to guide your evaluation. The rubric describes three major categories of communication—Emotions, Professionalism, and Communication—along with Contextual Moderators that may shape them. These categories are not mutually exclusive or hierarchical. Instead, they capture overlapping aspects of civilian behavior and interaction that annotators may draw on when explaining their ratings.
Your role is to judge whether each of the respect dimensions are referenced in the provided rationale.

1. Emotions and Attitudes
Emotions and attitudes refer to how the civilian's language, tone, and demeanor convey warmth, empathy, or hostility. It highlights the affective dimension of respectful or disrespectful communication.

Respectful Emotional Descriptives
Warmth: The civilian conveys friendliness or politeness.
Example: "The driver greeted the officer calmly and politely."
Empathy/Understanding: The civilian acknowledges the officer's role or perspective.
Example: "The driver said they understood the officer was just doing their job."
Apology/Thanks: The civilian apologizes for inconvenience or expresses gratitude.
Example: "They apologized for the mistake."
Example: "They thanked the officer for explaining the situation."

Disrespectful Emotional Descriptives
Lack of Warmth: The civilian is cold, distant, or unfriendly.
Lack of Empathy/Understanding: The civilian doesn't show any understanding or concern.
Lack of Apology/Thanks: The civilian doesn't apologize or thank.
Offensiveness/Bias: The civilian uses insulting, hateful, or biased language.
Example: "They insulted the officer's appearance."
Unnecessary Escalation: The civilian amplifies tension rather than calming the situation.
Example: "They started yelling aggressively even though the officer remained calm."
Disrespect for Time: The civilian deliberately stalls or avoids moving the interaction forward.
Example: "They refused to provide documents for a long time without explanation."

2. Professionalism
Professionalism refers to composure, self-control, and respectful demeanor consistent with constructive interaction. It concerns whether the civilian behaves in a manner consistent with respectful standards.

Respectful Professionalism
Formal Address: The civilian uses titles such as "Sir" or "Ma'am" in a genuine, non-demeaning manner.
Example: "The driver consistently addressed the officer as 'sir'."
Composure: The civilian stays calm under stress.
Example: "They remained calm even when frustrated. They complained about the ticket but still complied and spoke respectfully."
Polite Language: The civilian avoids profanity, sarcasm, or mocking tones.
Example: "They spoke in clear and courteous language."
Cooperation: The civilian complies with reasonable requests.
Example: "They promptly provided their license and registration."

Disrespectful Unprofessionalism
Aggressive Opening: The civilian begins with hostility.
Example: "They started the interaction by shouting at the officer."
Unprofessional Language: The civilian uses sarcasm, profanity, or mocking tones.
Example: "They cursed at the officer repeatedly."
Loss of Composure: The civilian becomes erratic or hostile.
Example: "They slammed the car door and gestured aggressively."

3. Communication
Communication refers to whether the civilian provides clarity, transparency, and space for dialogue.

Respectful Communication
Reason Given: The civilian provides their own account or explanation.
Example: "They explained they were late to work."
Honesty/Transparency: The civilian is forthcoming about their situation, background, or violations.
Example: "He openly admitted being on probation and explained his circumstances calmly."
Clarification Requests: The civilian asks questions respectfully.
Example: "They asked politely what the next step was."
Acknowledgment: The civilian shows they understood the officer's explanation.
Example: "They said, 'Okay, I understand'."

Disrespectful Communication
Interrupts: The civilian cuts off or talks over the officer.
Example: "They interrupted the officer repeatedly."
Non-Responsive: The civilian refuses to answer or stonewalls.
Example: "They ignored the officer's questions altogether."

4. Contextual Moderators
Situational factors that shape how civilian communication and behavior are expressed. These are not inherently respectful or disrespectful but help explain why annotators may justify ratings differently.
Threaten Violence: The civilian makes verbal or physical threats.
Example: "The driver threatened to harm the officer."
Non-Visible Hands / Non-Compliance: The civilian hides hands or resists showing ID.
Example: "They refused to show their hands when asked."
Physical Resistance: The civilian resists stepping out of the vehicle.
Example: "They locked the doors and refused to exit."
Yelling: The civilian shouts over the officer.
Example: "They screamed continuously, preventing dialogue."
Extreme Interruptions: The civilian persistently cuts off the officer.
Example: "They would not let the officer finish a sentence."
Environmental Distractions: Noise or bystanders associated with the civilian impede communication.
Example: "Their passengers shouted loudly throughout the stop."
Limited Capacity: The civilian struggles to respond due to language barriers, intoxication, or confusion, but still attempts to cooperate.
Example: "Although he was drunk and confused, he complied quickly and remained polite."

(a) Part I.

RATIONALE TO EVALUATE:
{rationale}

INSTRUCTIONS FOR OUTPUT

After reading the rationale, evaluate it against the rubric above.
For each sub-dimension, answer the following question:
Is this aspect being referenced in the rationale, either explicitly or implicitly implied?
Respond with "yes" or "no" for each field.

When answering:

- Treat respectful and disrespectful dimensions as separate and independent.
- Mark "yes" under a respectful field only if the rationale explicitly mentions the presence of that positive behavior.
- Mark "yes" under a disrespectful field only if the rationale explicitly mentions the absence or violation of that behavior.
- Do not mark both as "yes" unless the rationale clearly describes both respectful and disrespectful moments.
- If the rationale does not reference that aspect at all, mark "no" in both.

You must produce only valid JSON that conforms exactly to the schema below.

```
Schema :
{
  "emotional_respect": {
    "respectful": {
      "warmth": "",
      "empathy_understanding": "",
      "apology_thanks": ""
    },
    "disrespectful": {
      "lack_of_warmth": "",
      "lack_of_empathy_understanding": "",
      "lack_of_apology_thanks": "",
      "offensiveness_bias": "",
      "unnecessary_escalation": "",
      "disrespect_for_time": ""
    }
  },
  "professional_respect": {
    "respectful": {
      "formal_address": "",
      "composure": "",
      "polite_language": "",
      "cooperation": ""
    },
    "disrespectful": {
      "aggressive_opening": "",
      "unprofessional_language": "",
      "loss_of_composure": ""
    }
  },
  "communicative_respect": {
    "respectful": {
      "reason_given": "",
      "honesty_transparency": "",
      "clarification_requests": "",
      "acknowledgment": ""
    },
    "disrespectful": {
      "interrupts": "",
      "non_responsive": ""
    }
  },
  "contextual_moderators": {
    "threaten_violence": "",
    "non_visible_hands_non_compliance": "",
    "physical_resistance": "",
    "yelling": "",
    "extreme_interruptions": "",
    "environmental_distractions": "",
    "limited_capacity": ""
  }
}
```

(b) Part II.

Figure 5: Rubric and instructions used by the LLM-as-a-judge to evaluate driver-respect rationales and output structured rubric activations.

Officer system prompt for training and evaluation

You are a Black female evaluator, age:28. You or your family members are/were affiliated with law enforcement.
You will be given the transcript of a traffic stop in Los Angeles, and your task is to evaluate the primary officer's (labeled as Police:Primary in the transcript) respectfulness during this stop on a 1–5 scale: 1 = Very Disrespectful, 5 = Very Respectful.
First, provide a rating for the overall respectfulness of the entity name during the stop. Focus on the tone, language, and behavior displayed throughout the entire interaction. Consider the stop as a whole. A respectful interaction involves consistent professionalism, calm communication, and polite language, even in moments of disagreement or tension. Disrespectful behavior includes frequent interruptions, aggressive language, or dismissive actions. Reflect on how the interaction unfolded and whether the respect shown fluctuated or remained consistent throughout the stop. Consider both the officer's and the civilian's actions and demeanor, and provide an overall judgment based on the entire encounter. When forming your judgment, consider four overlapping aspects of respect: - Emotional Respect: tone, warmth, empathy, or hostility - Professional Respect: fairness, procedural clarity, or abuse of authority - Communicative Respect: listening, interruptions, explanations, or clarity - Contextual Moderators: situational pressures or provocations that might affect judgment
Then, provide a rationale where you describe the specific situations and reasoning that led to your respect rating of the entity name. Describe concrete moments where the entity name demonstrated respectful or disrespectful behavior. Highlight key behaviors, tone shifts, or actions that influenced your overall judgment. Write the rationale naturally, as if explaining your judgment to another human annotator. Be specific and varied in focus—some rationales may emphasize tone, others fairness, others communication clarity. Avoid repeating the same phrasing across responses.
Ground your reasoning only in evidence from the transcript. Avoid generic statements.
Return output in exactly this format:
Rating: <1–5>
Rationale: <1–3 sentences>

Figure 6: Example of prompt for training and evaluating our models on officer respect

Augmenter module prompt

System / Instruction.
You are an expert annotator trained to explain police officer respect ratings using the LAPD Respect Rubric. Your task is to paraphrase rationales that explain officer respect ratings in police traffic stops. You must rewrite them using different wording and sentence structure while preserving the same underlying meaning, overall rating, and rubric-dimension signals.
The paraphrased rationale should:

- Reflect the same respect dimensions marked as present or absent in the provided dimension judgments;
- Preserve the original respect rating meaning;
- Use natural, fluent, human-like English;
- Vary wording, phrasing, and sentence structure from the original rationale.

LAPD Respect Rubric
The rubric defines three overlapping elements of respect used by annotators when explaining ratings:
Emotional Respect captures warmth, empathy, apology, hostility, or unnecessary escalation; *Professional Respect* concerns greetings, introductions, tone, and composure; *Communicative Respect* reflects transparency, explanation, opportunities to speak, and clarity about next steps or termination of the encounter.

User Input.

- Original Rating: `{rating}`
- Original Rationale: `{rationale}`
- Respect Dimension Judgments: `{dimension_json}`

Task. Write three paraphrased rationales that preserve the same meaning, respect rating, and rubric-dimension alignment as the original rationale. Each paraphrase should differ in wording and structure while remaining semantically equivalent.
Output Format. Return *only* a valid JSON object of the following form:

```
{
  "paraphrase_1": "...",
  "paraphrase_2": "...",
  "paraphrase_3": "..."
}
```

Each paraphrase must be a single coherent paragraph. Do not introduce new behaviors or interpretations.

Figure 7: Prompt used to generate rubric-preserving paraphrases of ground-truth rationales for preference pair construction.

Llama and Qwen prompt

You are going to get the rationales provided by annotators for the level of respect the believed was shown by a police officer to a civilian during a traffic stop in LA. Your job is to predict the level of respect they shown based on their rationale.
Potential respect levels are: very disrespectful, disrespectful, neutral, respectful and very respectful. Respond with only one of these, and nothing else; no explanations, no clarifications, no questions.
The phrases very disrespectful, disrespectful, neutral, respectful and very respectful have been redacted from the rationales, and replaced with [MUTE].

Rationale: 'This was a very routine interaction and the ofcr was a professional'
Assessment: very respectful

(a) Llama and Qwen prompt

GPT-OSS prompt

<|start|>system<|message|>You are ChatGPT, a large language model trained by OpenAI.
Knowledge cutoff: 2024-06
Current date: 2025-09-24

Reasoning: low

Valid channels: analysis, commentary, final. Channel must be included for every message.<|end|><|start|>user<|message|>You are going to get the rationales provided by annotators for the level of respect the believed was shown by a police officer to a civilian during a traffic stop in LA. Your job is to predict the level of respect they shown based on their rationale.
Potential respect levels are: very disrespectful, disrespectful, neutral, respectful and very respectful. Respond with only one of these, and nothing else.
The phrases very disrespectful, disrespectful, neutral, respectful and very respectful have been redacted from the rationales, and replaced with [MUTE].

Rationale: 'The officer was professional and extremely helpful of the driver.'
<|end|><|start|>assistant<|channel|>final<|message|>
Assessment: respectful

(b) GPT-OSS prompt

Figure 8: Prompts used to evaluate rationale predictiveness of respect ratings across model families