Perceived Demographic Causally Influences Police Officer Statements to VR Characters in a Virtualized Setting

Anonymous ACL submission

Abstract

We analyze Virtual Reality (VR) scenario simulations in which 79 police officers from three U.S. police departments interacted with virtual characters varying by skin color (*Black/White*) and gender (*Man/Woman*) across bus-stop, convenience-store, and house scenes. We isolate the causal effect of perceiving a *Black Man* character (our "treatment") on the deference of officer speech (tone, e.g., politeness and impartiality). Results indicate a non-zero average treatment effect (ATE): in scene-level analyses—where data aggregation matches the original randomized assignment—officers speak less deferentially to *Black Man* characters overall and in bus scenes. We also observe a possible positive effect in store scenes (where the character was more likely a victim). For inverse-propensity–weighted ATEs, we incorporate LLM-featurized text as covariates in mixed-effects deference prediction models; for doubly robust ATEs, we fine-tune an LLM (Llama 3.1 8B) to produce propensity scores and deference outcome predictions used in downstream estimation. While LLMs are not yet proven for multilevel data, their contextual capabilities make them a potentially promising future tool for such analyses, very common in the social sciences, among other fields, and with survey data.

1 Introduction

In this study, we analyze data from experimental police scenario simulations conducted in a virtual reality (VR) lab to research police behavior as seen through their language in tense situations such as those they might experience in real life. In this setting, police interacted with virtual characters, proxies for people police may encounter, in VR scenes that emulated real life. We hope our study facilitates insights into the relationship between police demographics and attitudes, and their statements to perceived minorities (specifically Black men as proxied by *Black man* VR characters in our study) in everyday situations where there are inherent social power imbalances. These situations transcend cultural boundaries and are common globally.

Given the impactful role of police in public safety, the experiences of actual police officers studied in experimental settings such as ours can help make the case for human capital investments in police training in de-escalation tactics. Implicit bias, reflected in police speech, that we causally attribute to the perception of a *Black man* character, could be addressed via skills training to manage these encounters, such as targeted deescalation and communication training.

While the virtual character responses to the police in our study were rules-based and finite when created in 2020-2021, it is now possible to envision experiments and training exercises incorporating large language model-facilitated verbal responses from virtual characters in crime scenarios that are more human-like and dynamic, such that they could respond negatively or positively to an officers' statements, thus responding like a human might.

For this study, given data from 79 actual police officers engaging in VR police scene simulations, we specifically explore the causal effect that the perceived *Black Man* demographic (based on skin tone and apparent gender) of the virtual character had on the deference (tone) of the police officer speech. Deference was judged by 5,009 Prolific crowd workers, based on their evaluation of speech transcriptions of the police scenario simulations after they occurred. Our approach to computing the causal effect (average treatment effect or ATE) relied on traditional statistical models including mixed effects and regular regression models, with a range of ATE calculation methods, but were facilitated through the use of the Llama 3.1 8B large language model (LLM).

We specifically investigate the research questions:

- *RQ1: Do officers talk to Black Man virtual characters differently than they talk to others.* ; and
- *RQ2: If so, how much of this is explained by their social beliefs and views on policing (as indicated by their survey responses)*

The policing simulations were conducted with officers from three police departments in the Southern, Mid-Atlantic, and Midwest regions of the United States. The officers verbally interacted with virtual characters (persons of interest or victims) in different scenes, simulating real-life settings (house, bus stop, convenience store) where the virtual character's skin tone (*Black* or *White*) and gender (*Man* or *Woman*) were randomly assigned for a given officer in a particular scene.

2 Related Work

2.1 Policing Strategies

2.1.1 Social Factors and Training Trends in Countries with high Police Use of Force Incidents

Key societal factors in the U.S. and other countries with similar issues contribute to high rates of police use of force, disproportionately affecting minorities. Our study focuses on bias through speech toward Black American men, proxied by *Black Man* virtual characters in VR settings. Hirschfield (Hirschfield, 2023) shows that the U.S., along with some Latin American countries, has exceptionally high rates of fatal police violence (FPV, defined as all violence inflicted by police resulting in a documented citizen fatality) and fatal police shootings (FPS, fatalities caused by police gunfire, a subset of FPV), with the U.S. recording the highest known FPS rate of 3.1 per one million individuals in 2019. Necessary conditions for high FPV include societal division, conflict, and mistreatment of minority groups, while social grievance, political inequality, and restricted rights are strongly correlated with elevated rates. Countries with greater grievances and higher firearm prevalence also tend to have shorter or decentralized police training programs, as in the U.S., which may be a necessary condition for high FPV (Hirschfield, 2023). Given these challenges, standardized training and reforms to improve police interactions with people of color—particularly through de-escalation—are critical, and promising evidence suggests such practices are trainable and generally welcomed by U.S. officers. §2.1.2

2.1.2 The Evidence for De-Escalation Training

According to The Police Executive Research Forum (PERF) (Forum, 2021), research on the efficacy of de-escalation training in policing is limited due to inconsistent definitions, officer perceptions that they already practice de-escalation, and the lack of standardized implementation across departments. Nonetheless, several randomized controlled trials (RCTs) show promising results: the Integrating Communications, Assessment, and Tactics (ICAT) program led to significant reductions in use of force, officer injuries, and civilian injuries, with officers reporting favorable attitudes toward the training; the Tact, Tactics, and Trust program, evaluated in Tuscon, Arizona, and Fayetteville, North Carolina, emphasized deliberate, repetitive practice and scenario-based training that officers valued; and a Tempe, Arizona study confirmed officer appreciation of de-escalation practices, with measurable increases in the perceived value of tactics related to compromise, safety, and restraint. Taken together, this evidence suggests that while in-person training is beneficial, reinforcement through departmental policy and repeatable scenario-based simulations—including emerging virtual reality (VR) approaches—may be critical for sustaining and institutionalizing effective de-escalation tactics in U.S. policing.

2.2 Language Model Strengths for Predictive Modeling and Growth Areas for Multilevel Data

When utilizing language models for causal effect estimation with text covariates, it is important to note the strengths and limitations of large language models. We find that LLMs can play a critical role in downstream data analysis, including as inputs to statistical models. Specifically, in our study, we utilize the LLM, Llama 3.1 8B, in a couple of different ways. For our first methodological approach (Pipeline I in §3), we use the LLM to featurize the text in the conversational exchanges between the police officer and the character they are speaking to in a scene. (Tang et al., 2025) note that LLM embeddings for downstream tasks may perform better than feature engineering, particularly for high-dimensional regression tasks. We additionally utilize an LLM to finetune prediction models in our study. Specifically, we use classification and regression models to create the propensity score and

deference outcome prediction inputs to our doubly robust average treatment effect calculation method in our second methodological approach (see §3 for Pipeline II details). For finetuning LLMs for classification, (Minaee et al., 2021) argue that deep learning models have surpassed classical machine learning models in text classification.

Nonetheless, there are documented weaknesses of LLMs for applications to multilevel data in particular; that is, data that is inherently hierarchical or clustered such that multiple groupings exist. With this type of data, observations fall into distinct (or sometimes overlapping) clusters or groups, and groups often can hierarchically be classified into other higher-level groups as well. (Dickinson and Basu, 2005) discuss the dangers of not accounting for multilevel data in statistical modeling. From an LLM capabilities perspective, there are a couple of areas in which LLMs could improve. LLMs need to be stronger at multi-step reasoning (Patel et al., 2024); this may be important for LLM processing and generation related to multilevel data, where each level of the data is uniquely characterized and needs to be understood, i.e. properly accounted for. Inherently, LLMs are optimized for next token prediction, not deeper understanding of data relationships. Finally, LLMs are proven to perform well with supervision such as in classification to learn labels. But, more nuanced unlabeled data patterns are present in multilevel data and such features are often latent. Nonetheless, as we see in the study, long context LLMs, despite these challenges, still exhibit results for multilevel modeling that are within the range of mixed effects models. However, the precision required for statistical analysis is an area for growth.

3 Methodology

In this study, to approximate the impact of skin color and gender on the tone of police officer statements to a person of interest or a victim of a crime, we computed the ATE of the demographic (race and gender) of a virtual character on the average deference rating of police officer utterances. The deference of each officer's verbal statement was defined as the mean deference score applied to a police statement (i.e., utterance) to a VR character, where mean deference is the average of the individual manually assigned ratings of respect, formality, politeness, friendliness, and impartiality (ratings were integers ranging from 1 to 10, see Appendix A for an example of the specific guidance provided to annotators).

We estimate the causal effect of a perceived demographic d_i on $f_i = f(u_i)$, where f is deference of the officer speech, a function of the officer's utterance u (Gelman and Hill, 2007),(Austin, 2011). In our causal graph, we refer to deference as U. Our null hypothesis is that the treatment effect is zero, i.e. the deference rating of the officer's language is not impacted by the VR character demographic. We performed two-tailed tests of statistical significance of ATE at the 95% confidence level. Please see the Appendix A for dataset details.

3.1 Our Causal Model

We define key variable groups as follows:

- *B: Officer beliefs (unobserved)*
- *M: Officer demographics*
- *S: Answers to survey questions about beliefs*
- *U: (deference of the) Utterance of the officer to virtual character (outcome variable)*
- *C: Dialogue context (i.e., the dialogue plus setting [e.g. bus stop, etc.])*
- *D: Demographic of the virtual character (treatment variable)*

In computing the "treatment effect" of $D = 1$ for the *Black Man* character demographic, we wanted to understand how much the treatment impacts how deferential the officer is. However, a challenge in measuring the impact of the treatment D on U is the existence of (B,C,M), which are confounders between D and U. Therefore, to account for these confounds, a key estimator we compute is $P[D = 1|S, C, M]$, the propensity score.

Our causal graph, as seen in Figure 1, illustrates the causal relationship between the treatment variables D and U. In our causal diagram, since D is randomly assigned, it is not caused by other factors. B, C, M and D impact what the officer says at any given time. B and M also cause the context, as does D. It is important to acknowledge that there are other unmeasured variables that may influence B and C, such as the distance between the character and officer in a virtual setting and the volume of the audio. We assume that S is a partial measurement or approximation of B, but that it is independent of everything else in the graph.

3.2 Methods for Average Treatment Effect Calculations

Please reference the Appendix A for rationale on scene-level versus utterance-level models and their

relationship to the original experimental setup for the police scenario simulations.

3.2.1 Calculating the Causal Effect of the Perceived *Black Man* Demographic on Police Deference of Speech

We compute the ATE utilizing statistical and computational (LLM-based) methods. We describe and show results from two distinct methodological pipelines to illustrate not only the powerful capabilities of LLMs today and options for mixed statistical and computational methods, but also to shed light on the complexity of social data, demonstrating that different approaches can yield different results from complex hierarchical (multilevel) data such as ours. While we ultimately believe a hybrid approach utilizing both mixed effects models and the power of LLMs for text featurization is most robust for our use case (Pipeline I), the other LLM-facilitated approach (Pipeline II) we utilized is worth highlighting here, as it or variations on it could be useful for any number of applications.

3.3 Two Pipeline Approaches for Preparing Inputs to the ATE Calculations

3.3.1 Pipeline I: Mixed Effects Models with LLM-created Features from Text for Causal Effect Computation

To ensure statistically robust and accurate estimation of the coefficients on key explanatory variables for our deference prediction models, as well as reliable measurements of ATE for our utterance-level observations, we account for clustered (also known as multilevel) data, such as repeated measures and non-independence of utterances per officer (such as those per scene), as well as officers grouped into different police departments, through the use of mixed effects (or multilevel) regression models(Gelman and Hill, 2007).

First, we computed the ATE where our observations (samples) were at the "utterance" level; i.e., each observation contains an officer's set of statements that had a corresponding VR character response. For this level of observations, we performed IPTW as well. This approach was hybrid in that in addition to utilizing a statistical mixed effects model, it also pulled in LLM-created features (numeric representations) of the scenario conversation text. Specifically, after featurizing the conversational exchanges with Llama 3.1 8B, we performed principal components analysis (PCA[1]) to reduce the dimensionality of the text features to 50 principal components. We then input these along with our other covariates into our prediction models. For the IPTW, we first gathered propensity scores, or the probabilities of the treatment for each utterance-level observation, via a mixed effects logistic regression model regressing the binary VR character race and sex variable (D in our causal graph) with the other fixed effects covariates from the M, S, C groups (with C including the featurized text), with random effects of police department, scene type (also a C variable), and officer id. Then, we ran an outcome (deference) mixed effects model with covariates from the groups D, M, S, and C weighted by the propensity scores.

Second, we computed the ATE based on the true average deference scores, aggregating deference scores (averaging) and the whole set of police and VR character exchanges for each officer in a given scene into a single text, consistent with the original randomization approach where a different VR character (skin tone and gender combination) was assigned by officer for each scenario in which they participated; this enabled examination of results where all assumptions of a randomized control trial should hold. With essentially the same model setup as for utterance-level data but aggregated to the scene-level, we perform the deference outcome model regression with and without text features as shown in §4 to assess effects on our covariates for our second research question.

3.3.2 Pipeline II: Large Language Model Fine-Tuned Predictions as Inputs to Causal Effect Computations

For the LLM-based models of our second pipeline approach, we compute a doubly robust estimator that combines outcome modeling and propensity scoring, where the two types of prediction models are developed via LLM-finetuning. The doubly robust method for this pipeline is intended to mitigate any potential biases in our model, such as from un-(or only partially-) measured confounding variables, that might influence the officer deference of speech, our outcome variable (Funk et al., 2011).

For this approach, for our outcome (deference) prediction model, we fine-tuned the Llama 3.1 8B large language model for regression on utterance-level prompts to obtain our predicted deference estimates for our treatment and control groups. We

[1]Implemented PCA with scikit-learn 1.7.2.

additionally fine-tuned Llama 3.1 8B for classification to obtain our propensity scores, the probability of our treatment ($BlackMan$ VR character), given our covariates as seen in Figure 1. This approach to converting all covariates into a text prompt per observation was the primary difference in the data setup (essentially the same variables and formulas were utilized) with this approach versus that of Pipeline I. See Appendix A for fine-tuning details. Doubly robust causal estimation models such as the inverse propensity weighting approach we used (Funk et al., 2011) utilize not only propensity scores but also outcome model predictions, i.e. the expected value of the deference for treatment and control observations (Funk et al., 2011). The use of observations in the form of text prompts enabled our use of large language models to measure treatment effect, capitalizing on the contextual power of large language models for predictive models. Our training, validation and test splits for fine-tuning our LLMs ensured no data associated with a given officer overlapped between splits.

4 Results

Table 1, Table 3, and Table 4 illustrate the average treatment effects of the *Black Man* demographic treatment *D* on the deference of the officer statements U (Figure 1) obtained from two different pipeline approaches for prediction models and for different levels of text data aggregation (at the scene level versus at the utterance, i.e. police statement followed by character response). Table 2 and Table 5 regression results demonstrate the effects from different officer attributes on the deference outcome variable. We primarily reference our mixed effects model results (Pipeline I) to answer our research questions of *whether police officers talk to Black Man virtual characters differently than they talk to others*, and if so, *how much of this is explained by their social beliefs and views on policing (as indicated by their survey responses)*. While the results from LLM-finetuned prediction models (our implementation for Pipeline II) are less proven as inputs to average treatment effect calculation for multilevel data in terms of statistical robustness, (Table 4) they may be an important and useful option to be explored further by the research community, particularly for multilevel social datasets such as ours, which are common in the social sciences, as well as with complex hierarchical survey data. Our mixed effects model results more closely align with earlier results published for a different study on this data set, and the finetuned model results, which were at the utterance level, show larger statistically significant effect sizes and show a reversal of effect direction for the overall effect, but also show a positive effect for the store scene, similar to the utterance level mixed effects model. While language models are powerful in that they augment contextual knowledge of text, they do not explicitly account for and are unproven in their ability to handle multilevel data structures, as mixed effects models are designed to do.

4.1 Results from Pipeline I, Mixed Effects Models with LLM-Featurized Text

4.1.1 The Perceived *Black Man* as Virtual Character appears to Causally Influence Police Officer Deference of Speech

In reference to our first research question, we find that in fact, police officers do speak differently to *Black Man* virtual characters in VR simulations of real-life police scenarios. In Table 1, for the Pipeline I approach, in highlighted blue, we see statistically significant[2] negative treatment effects of the *Black Man* VR character on the deference of officer statements. Specifically, the numbers suggest that overall (for each and all scenario types the officers participated in), the perception of a *Black Man* virtual character, on a scale of 1 - 10 (with 1 being least deferential and 10 being most), causes a decrease of 0.16 (*naïve*) points in deference of police statements. For the bus scene in particular, deference points decrease 0.27 for the *naïve* model. Subgroups shown include by scene (bus stop, convenience store, and house) of the virtual reality simulation as well as by officer ("off") demographic, or a combination of the two. We denote y_1 as the deference for the treatment group (*Black Man*) and y_0 for the control group (*All Others*).

The naïve calculation of the average treatment effect of the *Black Man* virtual character on the deference of the police officer speech (the *basic* entries) is equal to the difference in the mean of the true deference of the officer statements for the treatment (D = 1, *Black Man* virtual character) versus that of the control (D = 0, *All Others* VR characters) ($mean(y_0) - mean(y_1)$). This calculation, while basic, is relevant for this setup since the level of the dataset observations corresponds to the original

[2] Statistical significance indicates a non-zero effect based on confidence intervals obtained from bootstrap resampling with 1,000 iterations)

ATE Method	ATE Subgroup	N	ATE Value
basic	overall	272	-0.16
basic	scene:bus	87	-0.27
basic	scene:house	90	-0.12
basic	scene:store	95	-0.09
basic	white male off	128	-0.16
basic	black male off	88	-0.15

Table 1: **Pipeline I: Scene-Level ATE of *Black Man* VR character demographic on officer deference of speech (with mixed effects models)**. Table reports the average treatment effect (ATE) of $D = 1$ (*Black Man*) on deference (U) of the officer utterances overall and by subgroup (statistical significance highlighted blue). We show the naïve difference in means (*basic*), equal to $mean(y_1) - mean(y_0)$) is the only ATE calculation needed as the randomized treatment was at the scene level.

Fixed Effects	Estimate	Signif
(Intercept)	5.13	***
char_race_sex_binary	-0.04	
dept	0.10	
agecollapsed	-0.04	
alwayspolice_num	0.04	
blackahardwork_num	-0.01	
blackdisc_num	-0.03	
blackint_num	0.02	
fampolice_num	0.24	.
feeltherm_black_num	-0.05	
feeltherm_pol_num	0.06	
offfemale	0.18	
offraceBiracial/Multiracial	0.10	
offraceHispanic/Latino	0.00	
offraceWhite/European	0.20	.
racialiden_num	0.01	
offeducBA	0.53	*
offeducHS	0.35	
offeducMA	0.72	**
offeducSome	0.32	.
gentrust_avg	0.03	
ipip_avg	-0.01	
polidentity_avg	-0.00	
poltrust_avg	0.03	
racialalien_avg	0.04	
rolepolice_avg	0.12	
socdom_avg	0.18	
srs_avg	0.07	

Table 2: **Mixed effects scene-level regression of deference on VR character's race and sex, police officer background, beliefs, and conversation text (LLM-featurized conversation text included as covariates)**. Baseline (intercept) deference rating is for a Black Male police officer. N= 253 observations from three police departments, 72 police officers and three scene types. Fixed effects variables are shown and fall into the groups seen in Figure 1. Other covariates included the 50 principal components resulting from PCA on the LLM-featurized texts; the police department, officer id, and the scene type were random effects.

randomized experiment, where each virtual character was assigned to the police officer by scenario ("scene" for short). Our outcome model independent variables were categorized into the variable groups delineated in §3 and in Figure 1. Observations were at the scene level with text aggregated such that all officer statements with VR character responses were concatenated into one text per observation, before featurization with our LLM.

In the case of Table 3, where we show the results from the IPTW mixed effects deference prediction model at the utterance level, each text observation (before LLM featurization) corresponded to a different exchange between the police officer and the VR character (the text of the observations was not cumulative in this setup). Notably, while the overall (across all scenes and officers) and bus scenes treatment effects are also negative here, they are much smaller effects and not statistically significant in this case. We do see a positive treatment effect for the store scene (increase of 0.16 deference of officer statement rating) and a negative treatment effect for the house scene (-0.19 deference effect). In this setup, the Black male officer (-0.08 points deference effect), the bi-multiracial female offifcer (-0.12 points deference effect), and bi/multiracial male officer (-0.35 points deference effect) all show slightly less deference when speaking to a perceived *Black Man* VR character. Only the white female officer (0.14 deference effect) appears to have a slightly more deferential tone to a *Black Man* character.

4.1.2 Officer Background Explains more Deferential Statements to VR Characters in Police Scenario Simulations

Regarding our second research question, in which we wanted to know how much deference is explained by officer beliefs or attitudes, it turns out, for the original randomized experimental setup (where there is one observation per unique combination of officer, scene type, and virtual character persona), that officer background attributes are most explanatory of the deference ratings of their statements to virtual characters, not their social beliefs or attitudes for our models. We note that we chose *Black Male* as our baseline officer demographic for these models. For our first model that regresses average deference rating U for police statements for a scene on the VR character's race and gender combination D, officer demographics

M, beliefs (partially measured by survey variables *S*), and with the scene type and conversation text as context (C), we observe that the officer's educational background, in particular having at least a Bachelor of Arts (BA) or a Master of Arts (MA) is associated with an increase of 0.53 and 0.72 points in their deference ratings, respectively. Having family who are police is also associated with a 0.24 point increase in deference ratings of their statements. Additionally, we see a slightly positive effect (0.20 points deference increase) from an officer having a White European racial background. Results for a second similar model can be seen in the Appendix Table 5, where we treat only the scene as context *S* (and do not include and thus do not control for the featurized text covariates), but otherwise have similar explanatory variables, the primary differences in the results are that we now see a statistically significant negative effect from Male officers (-0.34 deference rating points) on their deference of speech. Notably, with this model we also see that Latino officers speak less deferentially (-0.77 points deference), but Latino (Male) police officers speak almost a full point more deferentially than do Latina officers. There are different possible explanations for this finding that the survey answers about police attitudes and beliefs do not play a significant role in their deference of tone to a perceived *Black Man character*. It is possible that subjective police self-reporting on their attitudes and beliefs may be unreliable due to social desirability bias or maybe unconscious bias is not as easily measured through these types of survey questions, whereas more objective measures of education and family ties may be more reliable. Also, the virtual nature of the simulations undoubtedly adds some variation to the tone of the police officer statements that may not otherwise exist.

4.2 Results from Pipeline II: Treatment Effects of the *Black Man* Virtual Character Demographic from LLM-Fine-Tuned Prediction Models Are Similar, but Less Proven

In Table 4, results shown are from the test data set used for inference of the LLM models we fine-tuned, with a total of 456 observations/samples ($N = 456$). We show the effects from a doubly robust estimator (*dr*) that combines outcome modeling and inverse propensity weighting (§3). Specifically, we see an overall positive effect from $D = 1$ (officer perceives and interacts with a *Black Man* virtual character) on the officer speech to become more deferential ($ATE = 0.83$). When broken down by scene, we show a negative treatment effect in the bus scene ($drATE = -0.01$). Likewise, we show a positive treatment effect for the store scene ($drATE = 1.79$). We see a positive treatment effect for the doubly robust estimation ($drATE = 0.26$). Our outcome model independent variables were categorized into the variable groups delineated in §3 and in Figure 1. Observations were at the utterance level with text aggregated such that each set of officer statements that corresponded to a VR character response was concatenated into one text per sample, before featurization with our LLM. The text of the observations (before featurization) were not cumulative in this setup; rather, each observation corresponded to a different exchange between the police officer and the VR character.

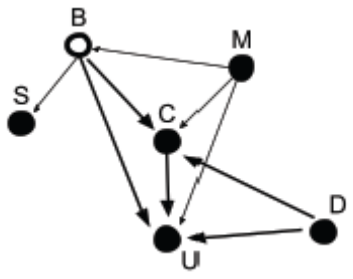

Figure 1: This causal graph illustrates the hypothesized treatment effect of the VR character demographic (D) on the deference rating (U) of police officer utterances. Arrows indicate directions of influence, with several confounding variables also affecting U directly or indirectly. We assume additional unobserved confounders between D and U. Opaque black circles represent observed variable groupings, while the transparent circle (B) denotes unobserved officer beliefs; survey variables (S) serve as a partial proxy for B.

5 Conclusion

Overall, we observe that there are statistically significant treatment effects of the perceived *Black Man* virtual character demographic on the deference of police officer statements; in other words, police do speak differently to the *Black Man* character than other characters. At the scene-level, where the data setup matches the original randomized control trial setup of the original simulations, we see negative effects (*Overall ATE = -0.16* and *bus ATE = -0.27* in Table 1). However, it's worth noting that at the utterance-level of observations, in both the Pipeline I and Pipeline II results, we see some statistically significant results for subgroup

ATE Method	ATE Subgroup	N	ATE Value
iptw_predicted	overall	4483	-0.02
iptw_predicted	scene:bus	1175	-0.04
iptw_predicted	scene:store	1558	0.16
iptw_predicted	scene:house	1750	-0.19
iptw_predicted	white male off	1940	0.04
iptw_predicted	black male off	1506	-0.08
iptw_predicted	black female off	291	-0.06
iptw_predicted	bi/multiracial female off	204	-0.12
iptw_predicted	bi/multiracial male off	92	-0.35
iptw_predicted	latina female off	52	-0.18
iptw_predicted	white female off	349	0.14
iptw_predicted	bus: white male off	497	0.03
iptw_predicted	store: white male off	736	0.15
iptw_predicted	house: white male off	707	-0.08

Table 3: **Pipeline I: Utterance-Level ATE of *Black Man* demographic on officer deference of speech (with mixed effects models)** This table reports the average treatment effect (ATE) of the $D = 1$ (*Black Man*) on the deference (U) of the police officer statements overall and by subgroup, with effect sizes statistically different from zero highlighted blue. Subgroups shown include by scene (bus stop, convenience store, and house) of the VR simulation or by officer ("off") demographic.

ATE Method	ATE Subgroup	N	ATE
dr	overall	456	0.83
dr	scene:bus	79	-0.01
dr	scene:store	182	1.79
dr	scene:house	195	0.26
dr	white male off	334	0.34
dr	black male off	90	0.78
dr	bus: white male off	53	0.22
dr	store: white male off	138	0.45
dr	house: white male off	143	0.22

Table 4: **Pipeline II: Utterance-Level ATE of *Black Man* demographic on officer deference of speech (with LLM-fine tuned prediction models)** This table reports the average treatment effect (ATE) of $D = 1$(*Black Man*) on the deference (U) of the police officer utterances overall and by subgroup, with effect sizes statistically different from zero highlighted blue. Here we show a doubly robust effect (*dr*) that combines outcome modeling and inverse propensity weighting such that the effect aims to correct for any misspecification of either model.

reporting, including some positive effects for the store (this was the scene where the officers should have understood that the character was a victim). The difference between the utterance level results and the scene level results suggests that the text exchanges are potentially too short to be meaningful for standard mixed effects statistical models to have enough signal. On the other hand, we see positive effects overlap between the Pipeline I and Pipeline II models for the house scene. Also, overall, effect sizes are notably larger for the Pipeline II models. This raises the question of if, in spite of the texts being potentially not representative enough for the utterance-level observations, LLM finetuned models, in spite of being unproven for multilevel data, may make up for this to some extent with the powerful context they gain through pretraining.

It is worth noting that given the virtual experimental simulation nature of the data, both positive and negative treatment effects are reasonable. In the case of positive ATE, where the deference of an officer's verbal statements are higher toward a Black Man character versus all other characters, police could be making a concerted effort to speak in a positive, respectful or controlled tone given the experimental nature of the simulation. It is also reasonable that some police may convey implicit or explicit bias via their verbal statements based on preexisting police attitudes and opinions. These attitudes could influence officers to speak less deferentially to a Black Man character, particularly in these simulations which were often tense and with virtual characters that were combative and argumentative. Regardless, we hope the insights from this study can help make a case for local policy changes or human capital investments such as for officer training via police scene simulations, which have proven effective as discussed in §2. Such training could support practice in de-escalation in settings where police interact with persons of interest, or reinforce positive communication habits that we see via our reporting of ATE in some settings.

6 Limitations

Data Limitations. Our sample of 79 police officers and the fact that each officer participated in only a handful of simulations is limited. However, we have attempted to mitigate this by capitalizing on the contextual power of LLMs via the incorporation of LLM-featurized police-character conversational text in the case of IPTW-based ATE cal-

cuations from our mixed effects models (Pipeline I) and by finetuning propensity and outcome prediction models on all the covariates in the form of text prompts, in the case of our doubly robust estimation of ATE in Pipeline II.

Participant Limitations. The set of 79 police officers that participated in original VR simulations from which we analyzed our data were from only a few police departments. It is possible that a larger sampling of police departments could be more representative of the greater U.S. police force.

7 Ethics Statement

We were fortunate to be able to take advantage of data collected under an approved Institutional Review Board application. We do not believe the data analysis that consitutes our study poses meaningful risks to the participants or to society; any risks the direct participants may have had, associated with being involved in the VR scenario simulations, would be addressed under the original IRB for the data collection effort. The data we utilized for this study was fully anonymized before our use. We utilized LLMs for coding assistance.

References

Peter C Austin. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424.

L Miriam Dickinson and Anirban Basu. 2005. Multilevel modeling and practice-based research. *The Annals of Family Medicine*, 3(suppl 1):S52–S60.

Police Executive Research Forum. 2021. What police chiefs and sheriffs need to know about collecting and analyzing use-of-force data. Police Executive Research Forum Washington, DC.

Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. 2011. Doubly robust estimation of causal effects. *American journal of epidemiology*, 173(7):761–767.

Andrew Gelman and Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Paul J Hirschfield. 2023. Exceptionally lethal: American police killings in a comparative perspective. *Annual Review of Criminology*, 6(1):471–498.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, 54(3):1–40.

Nisarg Patel, Mohith Kulkarni, Mihir Parmar, Aashna Budhiraja, Mutsumi Nakamura, Neeraj Varshney, and Chitta Baral. 2024. Multi-logieval: Towards evaluating multi-step logical reasoning ability of large language models. *arXiv preprint arXiv:2406.17169*.

Eric Tang, Bangding Yang, and Xingyou Song. 2025. Understanding LLM embeddings for regression. *Transactions on Machine Learning Research*.

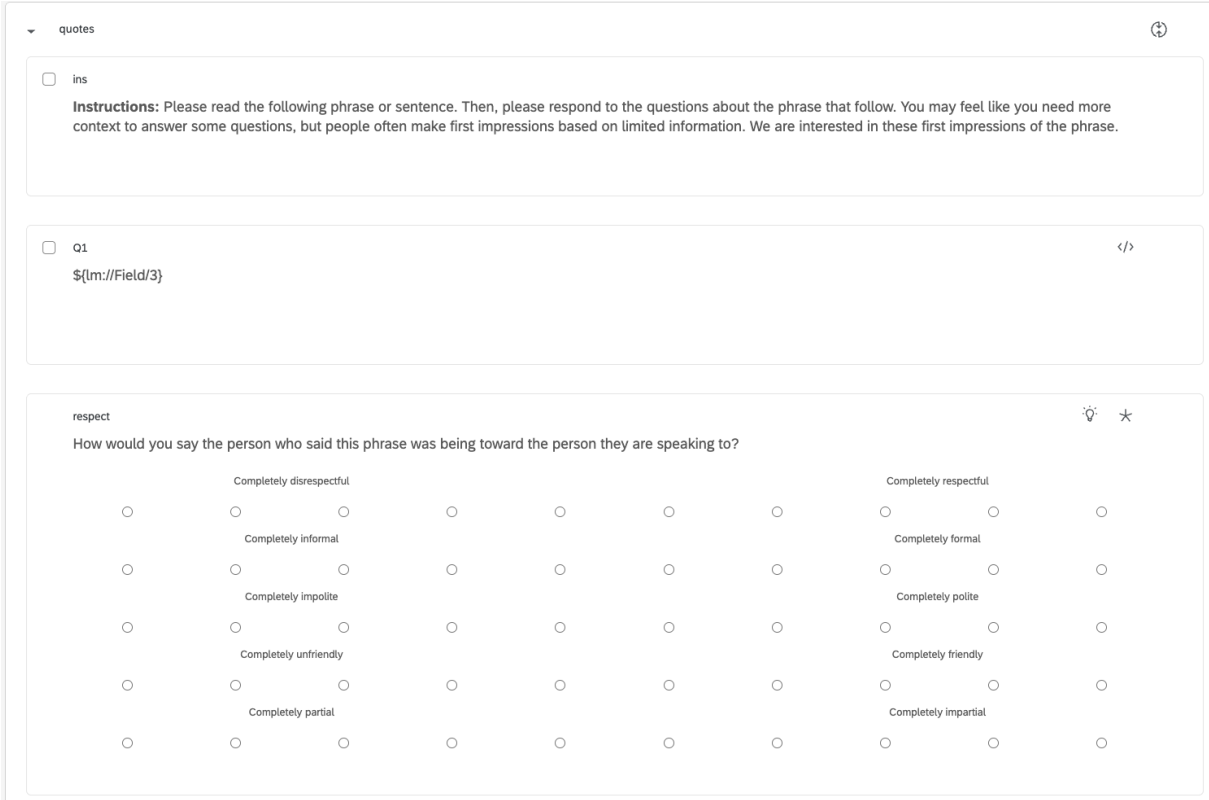

Figure 2: *Prompt for Deference Rating Components*: This was the prompt provided to Prolific workers to rate police officer statements on Likert scales of the following attributes: disrespectful, informal, impolite, unfriendly, partial

A Appendix

A.1 Dataset Details

In terms of the data collection from the simulations, the conversations the officers had with the characters they were assigned was transcribed via automated speech transcription (and subsequently validated/refined manually), and is the basis for the text data that was collected for this study (the numeric form of the transcribed conversational text is a covariate in our models, part of the context C from our causal graph, in addition to the scene type, also part of C). For the demographic variable D, non-*Black Man* characters were grouped into an *All Others* category, since we were most focused on ATE of the perceived *Black Man* demographic in light of police use of force against Black men in the United States. The average deference ratings (U in our causal graph) [3], our outcome variable of interest for our models, are composite scores regarding officer tone of speech assigned at the utterance level by crowd workers.

For our study, we aggregated the average deference scores and separately, the scene conversational exchanges, at two levels of granularity: 1) at the scene level for each each officer for each unique character with which they spoke to in a given scene, and 2) at the utterance level, where each observation contained a single set of police utterances that were followed by a character response.

Additionally, officers were asked to take an online survey including questions about their demographics and educational background (answers to these we consider (M) for demographics) as well as about their beliefs and attitudes toward policing and society (S for survey answers that were a partial measurement of their beliefs). Some officers took this survey before, others took it after their participation in the simulations. We include binary, categorical, and Likert-style variables for all of these variables in our datasets. (see Appendix A for a variable mapping to causal graph groupings.)

[3] See Appendix A for an example of the specific guidance provided to annotators.

Fixed Effects	Estimate	Signif
(Intercept)	5.75	***
char_race_sex_binary	-0.10	
agecollapsed	-0.02	
alwayspolice_num	0.09	
blackahardwork_num	-0.01	
blackdisc_num	-0.05	
blackint_num	0.00	
fampolice_num	0.20	.
feeltherm_black_num	-0.03	
feeltherm_pol_num	0.06	
offgenderMale	-0.34	.
offraceBiracial/Multiracial	-0.09	
offraceHispanic/Latino	-0.77	.
offraceWhite/European	-0.07	
racialiden_num	0.02	
offeducBA	0.55	**
offeducHS	0.41	.
offeducMA	0.82	**
offeducSome	0.34	*
gentrust_avg	-0.01	
ipip_avg	-0.07	
polidentity_avg	-0.00	
poltrust_avg	0.04	
racialalien_avg	0.02	
rolepolice_avg	0.08	
socdom_avg	0.05	
srs_avg	-0.04	
offgenderMale:offraceBiracial/Multiracial	0.13	
offgenderMale:offraceHispanic/Latino	0.99	.
offgenderMale:offraceWhite/European	0.20	

Table 5: **Mixed effects scene-level regression of deference on VR character's race and sex, police officer background, beliefs**. Baseline (intercept) deference rating is for a Black Male police officer (other variable reference values not listed). N= 253 observations from three police departments, 72 police officers and three scene types. Fixed effects variables are shown (see Appendix A for codebook) and fall into the groups seen in Figure 1. Other independent variables that were in the model but not shown here are the 50 principal components (numeric features X0-X49) resulting from Principal Component Analysis on the LLM-featurized texts, as well as the police officer id and the scene description, which we treated as random effects in our model.

A.2 Experimental Setup of Simulations and Implications for our Models

To reiterate, the experimental design of the police scene simulations was such that officers were randomly assigned a VR character of a specific demographic (*Black*, *White*, and *Man* or *Woman*) as part of a specific scene they experienced (bus, store, or house). In terms of our causal analysis framework, the treatment is the *Black Man* VR character or $D = 1$. Therefore, for the models with scene-level observations, we can assume, due to the randomized experiment that was conducted for the simulations, that the covariate values are balanced (pre-treatment characteristics of individuals are distributed similarly). It follows that causal analysis methods to create counterfactual populations where this balance is created are not necessary at this level of analysis. In contrast, we utilize inverse propensity treatment-weighted or doubly robust models for our models which utilized utterance-level observations. For this level of the data and these models, we assume that the covariate distributions have become imbalanced and therefore require causal analysis methods to reweight the observations. This is due to a few possible reasons – unmeasured or partially measured confounders since the randomization was not originally applied to the utterances but rather to the officers, or the introduction of numeric representations of the conversational text; these features in numeric form represent unobserved characteristics (see sections 3.3.1 and 3.3.2).

A.3 Finetuning Details

For our Pipeline II approach, we fine-tuned the Hugging Face Llama 3.1 8B parameter large language model to conduct supervised fine-tuning of the binary classifier with a "one-versus-rest" approach. We develop binary classification models by finetuning the base llama 3.1 8B model and adding a custom classification head composed of a three-layer perceptron, and implement last-token pooling to develop our sequence embedding. We use this machine learning approach to simplify a multi-class problem to two classes in order to utilize binary classification to compute the predicted value of the main estimator $P[D = 1|S, C, M]$.