Prose and Cons: Measuring Policing Disparities with Text Data

David S. Abrams* & Jonathan H. Choi†

March 20, 2025

Preliminary and incomplete; please do not quote, cite, or distribute.

Abstract

Researchers have traditionally run regressions on numerical and categorical data to detect disparities in policing by race, ethnicity, and gender. This approach can only control for a limited set of simple features, leaving significant unexplained variation and raising concerns of omitted variable bias. Using a novel dataset of text from more than a million police stops, we propose a new method applying large language models (LLMs) to incorporate text data into regression analysis of stop outcomes. Conventional methods excluding text data suggest significant disparities in favor of Black and male suspects. Our new LLM-boosted method incorporating text data changes both findings, implying much smaller disparities in favor of male suspects and race disparities statistically indistinguishable from zero.

*University of Pennsylvania Law and Wharton (dabrams@upenn.edu)

†University of Southern California Law School (jonchoi@law.usc.edu). This paper's hypotheses and methods are pre-registered with the Open Science Foundation at: `https://osf.io/jntr6/?view_only=ddaf04facf9941cd97e1155214163647`.

1 Introduction

Our criminal justice system exerts a huge amount of effort on prediction. Which juveniles at risk of committing crimes will benefit most from additional resources? Where should police officers patrol to optimally deter crime? Which defendants will recidivate if given a lenient sentence? In financial terms, these predictions determine the allocation of hundreds of billions of dollars at the many levels of the criminal justice system, involving millions of people. More important are the impacts on lives saved or lost; crimes committed or averted; years spent imprisoned or free.

To make these predictions, researchers and criminal justice decision-makers have traditionally focused on numerical data and data that can be coded numerically. For example, sentencing guidelines typically rely on two numerical dimensions intended to capture the extensiveness of a defendant's criminal history and severity of the crime in question, respectively. In other contexts like parole predictions (see e.g., Berk 2017), numerical scores may be used from psychological assessments, detailed criminal history and demographic data.

In this paper, we add to a long line of work on prediction in the criminal justice system. Specifically, by using LLMs, we can make much more extensive use of text data than has previously been possible. Rather than simply converting text data into categorical variables, we use LLMs to incorporate the full text of police reports, which has the potential to substantially increase explanatory power.

We investigate new data from police stops in Philadelphia, where we use the full text written down by the police officer to describe the circumstances of the stop and potential frisk or arrest. In these "*Terry* stops," officers may make stops or frisks only if there is reasonable suspicion the subject is involved in a crime, or in the case of frisks, carrying a weapon. Discovery of contraband—whether a weapon or evidence of crime—is therefore conventionally used as a primary outcome measure for the success of stops or frisks.

We compare contraband predictions made with numerical and categorical data to those made with the addition of text data. One of our analyses focuses on the impact of race in predicting contraband. There we find that text data have a substantial impact–completely changing the significance and/or sign of the variables for gender, race, and Latinx status.

Without controlling for text data, we find significant disparities favoring Black suspects, significant disparities disfavoring female suspects, and near-zero disparities with respect to Latinx suspects. Specifically, a female suspect has a 2.22 percentage point lower chance of being found with contraband than a male suspect ($p = 0.000$). Following Knowles et al. (2001), a lower likelihood of being found with contraband is interpreted as bias against the group in question (females), because it implies the police are over-policing despite the low likelihood of retrieving contraband. In contrast, a Black suspect otherwise identical to a White suspect on non-text observables has a 1.35 percentage point higher chance of being found with contraband ($p = 0.003$), and a Latinx suspect has a 0.87 percentage point higher chance of being found with contraband than a non-Latinx suspect ($p = 0.160$). The results for female and Black suspects are large in magnitude relative to the average 7.3% likelihood that any suspect is found with contraband, and they are statistically significant at the 99% level for the anti-female and pro-Black disparities.

However, when controlling for text data, these disparities change substantially. The likelihood of a Black suspect being found with contraband compared to a White suspect decreases to only 0.2 percentage points ($p = 0.592$), statistically indistinguishable from zero. The anti-female disparity reduces to 0.89 percentage points ($p = 0.032$), and the very slight disparity favoring Latinx suspects flips to a very slight disparity *dis*favoring Latinx suspects, of 0.11 percentage points ($p = 0.832$). The anti-female disparity is much smaller, and the pro-black disparity disappears almost entirely.

Thus our analysis suggests that text matters–failing to control for text can change both the magnitude and sign of findings about disparities. This suggests that empirical analysis of policing practice should shift toward methods that can take free-form text into account; and it casts doubt on some earlier results regarding disparities that do not take free-form text into account.

In addition, our methods allow us for the first time to analyze what language in police reports best predicts the discovery of contraband, as well as what language in police reports best predicts the legality of stops. We use this analysis to validate our methods and to provide additional insights into policing practice.

Part II of this paper provides background on our setting and prior work on the topic. Part III introduces our data and methods, and Part IV reports our main findings. Part V concludes.

2 Background

2.1 Setting

Our study analyzes a novel dataset of pedestrian *Terry* stops conducted by the Philadelphia Police Department between 2014 and 2023. *Terry* stops are named after the Supreme Court's ruling in *Terry v. Ohio* (1968), which established that police officers may briefly detain a person if they have reasonable suspicion that the individual has been, is, or is about to be engaged in criminal activity. The Court also held that officers may conduct a limited pat-down search, or "frisk," for weapons if they reasonably believe the person may be armed and dangerous. In practice, officers must be able to articulate the specific observations or information that led them to suspect criminal activity or the presence of weapons, thereby justifying the stop and potential frisk.

In our dataset, each time an individual is stopped or frisked, the officer records details about the person stopped, including the location and time of the stop, and the officer conducting it. Crucially for this paper, the officer also provides a narrative intended to establish reasonable and articulable suspicion (RAS) for the stop. These reports constitute the bulk of the data we analyze. Additionally, as part of a monitoring agreement stemming from litigation in *Bailey v. City of Philadelphia* (2011), attorneys code a random sample of stops and arrests with a legal assessment of whether RAS was present in the narratives.

The Philadelphia Police Department provided data from police reports written after stops. Certain variables—like whether the police had reasonable suspicion for stops and frisks—were manually coded by lawyers as part of the monitoring process, on a randomly selected sample. We supplement the police data with demographic, economic, and crime data from the U.S. Census and the Philadelphia Police Department for additional controls.

2.2 Prior Literature

Contraband discovery rates are an important outcome variable in empirical studies on policing, particularly in studying potential racial bias. Knowles et al. (2001) proposed an "outcome test," building on earlier work by Becker (1957), under which an unbiased

police officer should find contraband on suspects of different races, genders, etc. at equal rates. Different rates of contraband discovery suggest bias insofar as "officers driven by racial prejudice will continue to search minority citizens at higher rates despite finding less contraband" (Tillyer and Klahm, 2011).

Studies conducted on this theoretical foundation have had mixed results. Some research has found no evidence of disparities, with statistically equivalent contraband discovery rates between Black and White suspects (Knowles et al., 2001; Persico and Todd, 2006; Hernandez-Murillo and Knowles, 2004). However, other studies have found lower contraband discovery rates for minority suspects, suggesting potential disparities in search practices (Engel and Johnson, 2006; Ridgeway, 2007). Tillyer and Klahm (2011) found that Black suspects were twice as likely as White suspects to be found with contraband in discretionary searches, suggesting disparities in *favor* of Black suspects (at least in discretionary searches; they found equal rates for mandatory searches).

Critics have raised concerns about the assumptions and limitations of using contraband discovery rates as evidence of disparities. Most prominently, various critics have observed the specter of omitted variable bias—for a variety of reasons, the circumstances of police stops may differ depending on the race of the suspect (Anwar and Fang, 2006; Engel, 2008; Engel and Tillyer, 2008).

Police departments often have extensive documentation of the circumstances of police stops that could in theory be used to mitigate omitted variable bias. However, the documentation is natural language not easily converted to structured data. This article addresses this issue by using LLMs to incorporate natural language data in statistical analysis of contraband discovery.

3 Methodology

3.1 Hypotheses

We aim to test several different hypotheses in this paper. [1]

[1] [[Note that we have not yet tested some of the hypotheses below–we will do so soon!]]

First, we test the hypothesis that individual characteristics (particularly race, gender, and ethnicity) will significantly influence the predicted values of dependent variables. If individual characteristics strongly predict outcomes, that may suggest disparities in policing practice. As noted above, if Black suspects are less likely to be found with contraband than White suspects, that would suggest that the police are biased against Black suspects (because they are more likely to subject Black suspects to legally unjustified stops). Or, if Black suspects are more likely to be frisked after being stopped compared to White suspects, even holding the stated circumstances of the stop constant, that would again suggest that the police are biased against Black suspects. Thus regression analysis helps to shed light on potential disparities in policing.

Second, we test the hypothesis that text matters. By conducting prediction both with structured data only (excluding text data) and with all data, including text data, we can assess how important text data are in the story and whether regressions that exclude text data suffers from omitted variable bias. If the results significantly differ, that would cast doubt on a huge swath of the literature and raise the inclusion of text data as an important best practice for future work.

Third, we test the hypothesis that specific pieces of language in police reports predict outcomes. For example, what activities of suspects are most likely to lead to arrests? What activities of suspects are most likely to lead to *legally impermissible* arrests? Shedding light on what factors police care the most about can help us to determine whether these factors are in line with our values, and determining what is most predictive of impermissible arrests provides valuable information for future officer training.

Fourth and finally, we test the hypothesis that different officers have different propensities to reach certain outcomes. Perhaps certain officers are more likely to make impermissible frisks, or more likely to escalate a stop to a frisk or a frisk to an arrest, all else equal. Perhaps certain officers simply behave more unpredictably. Testing this hypothesis will help us to shed light on which officers are potentially problematic, potentially providing a useful tool to police departments to flag problem officers for further training.[2]

[2] [[Note: We have not tested this hypothesis yet.]]

3.2 Data

Our data come from the Philadelphia Police Department and describe pedestrian police stops and frisks from the years 2016 to 2023. Our sample of 67,469 total observations was randomly selected from over a million stops in this time period. A large number of variables are available, including information on outcomes, subjects, officers, and importantly, free text data. Table 5 includes summary statistics on a selection of the variables. We intend to conduct additional analysis on the full dataset, which has closer to a million observations; the full dataset lacks lawyer codings of whether reasonable suspicion was present, but LLMs may be able to impute this information as described below.

The data include information on whether a suspect was frisked, whether contraband was discovered, whether there was reasonable suspicion for the stop or the frisk, and whether the suspect was arrested. There is also information on the type of contraband recovered, including whether it was a gun or other weapon, drugs or something else. There is a great deal of data about the subject, including variables about individual appearance, race, gender, age, and Latinx status.

There is information about the location and time of the stop as well as identifiers for the officer and partner making the stop. We code location based on the Police Service Area (PSA) in which the stop was made. Most importantly, there is a detailed free-text narrative by the police officer explaining the reason for the stop, which is intended to convey evidence of reasonable suspicion. The same information is also available for a frisk if one was made.

Table 5 contains summary statistics for each of the data we used in our analysis.

4 Analysis - Contraband Discovery

4.1 Methods

4.1.1 LLM-Boosted OLS Regression

Applying related research that one of us is currently conducting, we use a new method to incorporate natural language in causal inference by leveraging predictions generated

from a fine-tuned large language model (LLM). Specifcally, we fine-tune an LLM to generate direct predictions of the outcome variable, which are then used as an additional control in OLS regression.

Mathematically, a conventional OLS regression would take the form:

$$Y_i = \beta_0 + \boldsymbol{\beta}^T \mathbf{X}_i + \varepsilon_i \tag{1}$$

Where Y_i is the dependent variable for observation i, β_0 is the intercept term, $\boldsymbol{\beta}$ is a $k \times 1$ vector of coefficients for the independent variables, $\mathbf{X}_i$ is a $k \times 1$ vector of independent variables for observation i, and ε_i is the error term for observation i.

We simply add an additional term to this regression:

$$Y_i = \beta_0 + \boldsymbol{\beta}^T \mathbf{X}_i + \delta p_i^{LLM} + \varepsilon_i \tag{2}$$

Where δ is the coefficient for the LLM-predicted probability, and p_i^{LLM} is the predicted probability generated by the fine-tuned LLM for observation i.

We generate outcome predictions using Llama 3, a decoder-only transformer model developed by Meta. The weights of Llama 3 are open-source to researchers, making it suitable for fine-tuning, unlike closed models like Anthropic's Claude and OpenAI's GPT. This accessibility allows us to fine-tune the model weights to optimize performance for our specific prediction tasks.

To adapt Llama 3 to our specific prediction tasks, we fine-tune by tweaking the weights of the model to optimize its performance on a set of training examples. Following Devlin et al. (2019) and Brown et al. (2020), we minimize the cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \tag{3}$$

where y_i is the empirical outcome variable (here, whether or not the suspect ultimately was found to have contraband) and $\hat{y}_i$ is the model's prediction.[3] After

[3]We specifically employ Low-Rank Adaptation (LoRA), an implementation of Parameter-Efficient

fine-tuning, we evaluate its performance on a set of test examples that was kept segregated during the training process.[4] Bai et al. (2023) show that transformers can implicitly learn and employ statistical models in prediction, suggesting that a fine-tuned LLM might adapt textual inputs to a wide range of functional forms without being explicitly trained with any particular functional form in mind.

We specifically train Llama 3 to predict whether a suspect was discovered to have contraband using the stop narrative produced by the police officer. Then, once we fine-tune Llama 3 to predict whether contraband was discovered, we incorporate the predicted probabilities generated by the LLM[5] as an additional variable in OLS regression. That is, we model Equation 2, where Y_i is whether or not the officer discovered contraband for suspect i (1 if yes, 0 if no), and p_i^{LLM} is the LLM's predicted probability that contraband would be discovered for suspect i. This allows us to use an LLM to reduce the residual variation in our predictive task while relying on OLS assumptions familiar to economists, simply adding an additional control.

We calculate the predictive performance of the OLS model, whether incorporating textual predictions or not, by using R-squared statistics as well as Mean Squared Error (MSE), testing MSE by training the OLS model on the training set and then

Fine-Tuning (PEFT), which fine-tunes a subset of the model's weights in order to reduce memory requirements and training time (Hu et al., 2022; Zhang et al., 2022). In LoRA, the weight update for a layer is represented as the product of two low-rank matrices, significantly reducing the number of trainable parameters compared to full fine-tuning (Zhang et al., 2022).

For a given weight matrix $\mathbf{W}_0$ in the original model, LoRA decomposes the weight update into the product of two low-rank matrices $\mathbf{B} \in R^{d \times r}$ and $\mathbf{A} \in R^{r \times k}$, where r is the chosen rank (typically much smaller than d and k):

$$\mathbf{W} = \mathbf{W}_0 + \alpha \mathbf{B}\mathbf{A} \tag{4}$$

Where $\mathbf{W}_0$ is frozen during training, α is a scaling factor, and only $\mathbf{B}$ and $\mathbf{A}$ are trained. This reduces the number of trainable parameters from $d \times k$ to $r(d + k)$. For example, if $d = k = 1000$ and $r = 8$, this reduces the number of trainable parameters from 1,000,000 to 16,000.

During the forward pass, given an input $\mathbf{x}$, the output is computed as:

$$\mathbf{h} = \mathbf{W}\mathbf{x} = \mathbf{W}_0\mathbf{x} + \alpha(\mathbf{B}\mathbf{A})\mathbf{x} \tag{5}$$

This decomposition allows for efficient fine-tuning while maintaining model performance through the low-rank approximation of the weight updates.

This approach has shown comparable performance to full fine-tuning on various natural language processing tasks while dramatically reducing the number of trainable parameters (Hu et al., 2021).

[4]We use an 80%/20% training/testing split, which is one conventional option for large datasets like ours.

[5]These are straightforwardly generated by exponentiating the LLM's log probabilities.

calculating MSE on the test set.[6]

4.1.2 Why Not Use Only LLMs?

The method described above is just one way to use LLMs to incorporate text data in a prediction task. Another alternative would be to use an LLM as a complete replacement for OLS regression, giving an LLM as input *all* variables, including race, ethnicity, gender, other controls (which could be analyzed using OLS regression), and the police report text, and training the LLM to predict contraband discovery. The specific impact of each variable could then be estimated using perturbation analysis, comparing the average prediction of contraband discovery across the sample when, for example, we perturb the sample so that all suspects are White and all suspects are Black (but all other variables otherwise retain their original sample value). The difference between the average predictions in each perturbation condition would then produce an estimate of the effect of a suspect's being Black instead of White, analogous to regression coefficients. The same perturbation procedure could be repeated to produce estimates for all variables of interest.

The LLM-only method could theoretically always perform at least as well as any method that constrains functional form, like the method applied in this paper. Neural networks (which are at the core of all modern LLMs) are Turing-complete (Siegelmann and Sontag, 1995), meaning it is possible to imitate any alternative computational method with a sufficiently large neural network (Hornik, 1991). If there are significant interaction terms between any of the variables in an OLS regression (or interactions between the interactions, or interactions between interactions between interactions, etc.), the LLM-only method could take these into account as well, which theoretically should improve performance.

One of us has conducted methodological work evaluating the difference in performance and explainability of different techniques to use LLMs in causal inference,

[6]It is also possible to calculate accuracy (what proportion of predictions was correct), precision (what proportion of positive identifications was correct), recall (what proportion of actual positives was identified correctly), and F1 score (the harmonic mean of precision and recall), which are common in the machine learning literature; however, due to significant class imbalance, when we restrict the training data (notably when we drop natural language), in some training samples no inputs are sufficiently predictive of the under-represented training class, leading to zero predicted negatives or positives and an F1 score of zero. Thus we believe that MSE and R-squared are better metrics for performance in this particular setting.

including both the method used in this paper (LLM-boosted OLS) and the LLM-only approach described in the previous two paragraphs. Choi and Connell (2024) finds that the improvement in going from LLM-boosted OLS to LLM-only inference is generally small, while the disadvantages are significant, especially that LLM-only inference is a black box. While regression coefficients can be approximated using perturbation analysis, the true functional form of the inference (for example, the importance of the implicit interactions between variables) remains unknown. This opacity contrasts sharply with the interpretability of OLS coefficients, which are well understood and extensively analyzed in existing econometric literature.

Moreover, Choi and Connell (2024) finds that coefficients imputed using the perturbation process above tend to be inaccurate in synthetic settings with known coefficient values. In contrast, incorporating an additional control into OLS is relatively secure and unlikely to bias estimates of other variables, subject to the potential issues discussed in Subsection 6.[7]

4.1.3 Pre-Processing Police Reports to Address Proxying, Multicollinearity, and Attenuation Bias

Officer descriptions of stop events often include details that might be problematic for our analysis. First, many of the police reports contain details not only of the events leading up to the stop but also the outcome of the stop (for example, "Marijuana was found..."). Including this text when fine-tuning our LLM would cause attenuation bias on all other variables in our OLS regression, because the LLM prediction variable p_i^{LLM} would capture some of the variation that should be attributed to other variables (Frost, 2020).

Second, many of the descriptions explicitly describe demographic variables about the suspect that are already included in the OLS regression (for example, "Male was found..."). Including this information in the text used to fine-tune the LLM would

[7]There are other, more technical, considerations that make LLM-only inference unappealing. First, by converting numerical data to tokens, an LLM conducting direct prediction does not optimize a numerical model. This can make its behavior potentially unpredictable. For example, continuous variables may not monotonically affect predicted outcome probabilities, complicating interpretation. Second, the specific fine-tuning method used (PEFT with LoRA) may underfit or overfit on the data. Decisions about hyperparameters are more consequential than when only using LLM prediction to generate an additional variable. This potentially increases researcher degrees of freedom, raising concerns about the robustness and replicability of findings.

cause multicollinearity between our variables of interest and p_i^{LLM}. This too would cause attenuation bias in the estimate of the coefficients on our variables of interest (for example, the variable representing gender).

Third and more subtly, even if we remove explicit references to race, ehtnicity, and gender, *proxies* for key demographic variables might remain–for example, female suspects might be more likely to be described as wearing dresses, and male suspects might be more likely to be described as having beards. Because wearing a dress or having a beard is not relevant to contraband discovery except to the extent it proxies for gender, including this sort of information in the text of police reports would also cause multicollinearity and attenuation bias.

We address all of these issues by pre-processing the datasets using an LLM to exclude the problematic information described above. Detailed information about the model, model hyperparameters, and prompts we used is available in Section B.3 of the Appendix. We first generate a version of the dataset redacting any information that could not have been known prior to the beginning of the stop (for example, simply removing the sentence beginning "Marijuana was found..."). We then redact any explicit mention of demographic variables of interest from the police reports (for example, replacing "male suspect" with "individual").

After conducting these initial redactions, the concern about proxy multicollinearity still remains. To identify whether the remaining text after the initial redactions would allow the LLM to proxy for race, gender, or ethnicity, we use a logistic regression model with LASSO regularization (Tibshirani, 1996) to estimate which specific words are most predictive of a suspect's race, gender, and ethnicity.

Specifically, let d_{ij} represent the frequency of word j in document i. The term frequency (TF) is given by:

$$\text{TF}_{ij} = \frac{d_{ij}}{\sum_k d_{ik}} \tag{6}$$

The inverse document frequency (IDF) for word j is calculated as:

$$\text{IDF}_j = \log \frac{N}{|\{i : d_{ij} > 0\}|} \tag{7}$$

where N is the total number of documents and $|\{i : d_{ij} > 0\}|$ is the number of

documents containing word j. The final TF-IDF score is:

$$\text{TF-IDF}_{ij} = \text{TF}_{ij} \times \text{IDF}_j \tag{8}$$

Using these TF-IDF vectors as inputs, we estimate separate LASSO-regularized logistic regression models for race, ethnicity, and gender. For each demographic variable $y_i \in \{0, 1\}$, the probability of the positive class is modeled as:

$$P(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + e^{-(\beta_0 + \sum_j \beta_j x_{ij})}} \tag{9}$$

where $\mathbf{x}_i$ is the TF-IDF vector for document i, and the β_j coefficients are estimated by minimizing the penalized negative log-likelihood:

$$\min_{\beta_0, \boldsymbol{\beta}} \left[-\sum_{i=1}^{N} \left(y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right) + \lambda \sum_{j=1}^{J} |\beta_j| \right] \tag{10}$$

where $p_i = P(y_i = 1|\mathbf{x}_i)$, λ is the regularization parameter, and the L_1 penalty term $\lambda \sum_{j=1}^{J} |\beta_j|$ encourages sparse solutions by shrinking some coefficients exactly to zero. The regularization parameter λ is selected through 5-fold cross-validation to maximize prediction accuracy while maintaining model interpretability.

In our initial analysis, we find that a number of important proxies are present in the text data: the suspect's *name* (predictive of Latinx ethnicity), *attire* (predictive of gender, including accessories like purses), *hair* (predictive of gender, including facial hair), and *location* (predictive of race). Information about name, attire, and hair has little plausible predictive value except as a proxy for demographic variables, so we redact them. Location information is more ambiguous–indeed, location is usually treated as an appropriate control and we do separately control for location in our OLS regression, following convention. Out of an abundance of caution, we redact geographic data as well, relying on the controls already present in the OLS variables.

After completing all redactions, we run a final LASSO-regularized logistic regression to inspect the words that are most predictive of race, ethnicity, and gender. Table 1 shows the performance statistics for the estimate of race, ethnicity, and gender respectively.

Metric	Race	Ethnicity	Gender
Overall Accuracy	0.7180	0.9103	0.8553
Macro Precision	0.6239	0.4552	0.5944
Macro Recall	0.5116	0.5000	0.5003
Macro F1-Score	0.4508	0.4765	0.4618

Table 1: Performance statistics for logistic regression models predicting demographic variables (*Race*, *Ethnicity*, *Gender*) from redacted police report text. The models were trained to assess whether the pre-processed text data still contained proxies for demographic information after redactions. Metrics reported include Overall Accuracy, Macro Precision, Macro Recall, and Macro F1-Score. The dataset was split into 80% training and 20% testing sets. The metrics are reported for the test set.

Table 1 shows that the model has limited ability to accurately predict demographic variables based on the text. As Table 5 shows, our sample is 70.2% Black, 9.6% Latinx, and 14.1% female. Thus a model that guessed that all suspects were Black, non-Latinx, and male would have an accuracy of 0.702, 0.904, and 0.859, respectively. The actual accuracy statistics from the prediction model barely exceed these base rates.

Table 2 shows the top 10 most predictive features for each demographic variable, along with their corresponding importance scores (in parentheses; these are the coefficients from the logistic regression, but they have no clear interpretation because of the vectorization and tokenization process we use, described in Section B.4 of the Appendix).

Race	Ethnicity	Gender
syringes (-3.5739)	swimming (3.0685)	17 (3.8111)
identify (3.3943)	spoke (-2.7244)	flag (3.3598)
syringe (-3.0125)	crisis (2.6462)	baggy (3.0576)
dock (-2.8367)	24th (2.1374)	verify (2.3874)
tracks (-2.7470)	sexual (2.0967)	hanging (2.3330)
spotted (2.6971)	detained (2.0826)	entryway (2.0856)
carts (-2.4620)	fence (2.0328)	staggering (1.9590)
15 (2.3055)	k2 (1.9692)	help (1.9516)
h37 (-2.1986)	grant (1.9604)	electric (1.9217)
median (-2.1901)	cleared (1.9453)	assaulted (-1.8568)

Table 2: Top 10 most predictive words (features) for logistic regression models predicting each demographic variable (*Race*, *Ethnicity*, *Gender*) from redacted police report text. The numbers in parentheses are the coefficients from the logistic regression models (importance scores), indicating the strength and direction of association between each word and the demographic variable. Positive scores indicate words more associated with White, Latinx, or male suspects, while negative scores are associated with Black, non-Latinx, or female suspects. The features were extracted using Term Frequency-Inverse Document Frequency (TF-IDF) vectorization with LASSO regularization.

Qualitative inspection of the words most predictive of our demographic variables suggests that we have appropriately redacted proxies for variables of interest. The remaining predictive features are largely related to the nature of the stop or the behavior observed, rather than direct proxies for demographic characteristics.

This analysis provides confidence that our redaction process has significantly reduced the presence of demographic proxies in the text data, minimizing the risk of multicollinearity and attenuation bias in our subsequent analyses.

4.2 Regression Equations and Model Performance

We conducted four different OLS regressions:

1. Key Variables:

$$y_i = \beta_0 + \beta_1 f_i + \beta_2 b_i + \beta_3 l_i + \varepsilon_i \tag{11}$$

This regression includes only the key variables of interest: f_i represents gender

(female), b_i is an indicator variable for whether the suspect is Black, and l_i represents whether the individual is Latinx.

2. Basic Controls:

$$y_i = \beta_0 + \beta_1 f_i + \beta_2 b_i + \beta_3 l_i + \boldsymbol{\beta}^T \mathbf{X}_i + \varepsilon_i \tag{12}$$

This regression adds control variables in vector $\mathbf{X}_i$, which includes age (in years), height (in feet), weight (in pounds), build (thin, medium, heavy, stocky, small, tall, muscular, or some combination thereof),[8] whether a partner was present (yes or no), a linear time control (minutes since the year 2000), and the month[9] and time of day when the stop occurred.[10]

3. Full OLS Controls:

$$y_i = \beta_0 + \beta_1 f_i + \beta_2 b_i + \beta_3 l_i + \boldsymbol{\beta}^T \mathbf{X}_i + \sum_{o=1}^{O} \gamma_o o_{oi} + \sum_{p=1}^{P} \delta_p p_{pi} + \varepsilon_i \tag{13}$$

This regression includes all previous variables and adds o_{oi} for officer fixed effects and p_{pi} for PSA (Police Service Area) fixed effects.

4. Full OLS Controls + LLM:

$$y_i = \beta_0 + \beta_1 f_i + \beta_2 b_i + \beta_3 l_i + \boldsymbol{\beta}^T \mathbf{X}_i + \sum_{o=1}^{O} \gamma_o o_{oi} + \sum_{p=1}^{P} \delta_p p_{pi} + \lambda p_i^{LLM} + \varepsilon_i \tag{14}$$

This final regression includes all previous variables and adds p_i^{LLM}, which represents the LLM-predicted probability of contraband discovery for observation i.

Table 6 and Figures 1 and 2 present performance statistics for each of the regressions. As is conventional in the computer science literature, we test performance by randomly splitting our dataset into a training subset and a testing subset (80% of the sample for training, 20% for testing). We "train" the OLS model by fitting its coefficients based on the training subset, and we calculate R-squared and MSE by plugging

[8] We used fixed effects for the specific combination of build types, although the vast majority of suspects were identified with only one build type.

[9] We used fixed effects for each month.

[10] We included fixed effects for time of day, splitting the day between midnight to 8 AM, 8 AM to 4 PM, and 4 PM to midnight.

in predicted values for y_i using the coefficients generated when estimating OLS on the training subset. Thus, for the "train" R-squared and MSE in the table and figures, we calculated residual variation using coefficients from the training subset, variable values from the training subset, and true values from the training subset; but for the "test" R-squared and MSE, we calculated residual variation using coefficients from the training subset, variable values from the *testing* subset, and true values from the *testing* subset.

R-squared and MSE will almost always be better (higher and lower, respectively) in training than in test. The difference in these statistics between training and test is known as generalization error and is often a sign of overfitting. As the table and figures show, generalization error significantly increases when we add controls for PSA and officer. This is to be expected, because there is substantial variation in the cases that any particular officer will deal with; the officer's propensity to find contraband in the training set may not carry over to the testing set, causing overfitting.

The key result from the table and figures is that R-squared and MSE dramatically improve when adding predictions based on natural language. Explanatory power in the absence of text data is quite poor even with full controls and fixed effects, with test R-squared of 0.0584 and test MSE of 0.0625. With the addition of text data, test R-squared jumps to 0.3368 and test MSE falls to 0.0440, substantial improvements in both cases.

This suggests that important residual variation is captured when we include text data, and reinforces the concern that there may be omitted variable bias when text data are excluded.

4.3 Examples of Changed Predictions

The inclusion of LLM predictions in our regression analysis led to significant changes in contraband prediction probabilities for individual cases. Here are two illustrative examples:

- **Example 1:**
  Police Report: "During a narcotics investigation, police went into store and observed an individual offender place green leafy substance in multiple vials and packets underneath the chip rack."

LPM probability without LLMs: 4.64%
LPM probability with LLMs: **90.47%**

- **Example 2:**
  Police Report: "Police observed a green Honda parked running unattended with two children inside. The individual came from inside of the store and stated that the unoccupied vehicle was indeed theirs."

  LPM probability without LLMs: 32.07%
  LPM probability with LLMs: **0.32%**

These examples demonstrate how the inclusion of text data can lead to more context-aware predictions, correcting for biases or oversimplifications in models relying solely on structured data.

4.4 Results

Table 7 and Figures 3, 4, and 5 present the results of the regression analysis. The model including only structured data suggests that female suspects were 2.22 percentage points less likely to be found with contraband than male suspects ($p = 0.000$); Black suspects were 1.35 percentage points more likely to be found with contraband than White suspects ($p = 0.003$); and Latinx suspects were 0.87 percentage points more likely to be found with contraband than non-Latinx suspects ($p = 0.160$).

Using the model of contraband discovery in Knowles et al. (2001), these results suggest strong anti-female disparities, moderate pro-Black disparities, and near-zero disparity with respect to Latinxs. The results regarding anti-female and pro-Black disparities were significant at the 99% level. These effects are substantial when compared to the average 7.3% likelihood that any suspect is found with contraband.

However, when text data are included in the training and test datasets, coefficient estimates dramatically shift in magnitude. When controlling for text data, female suspects were 0.89 percentage points less likely to be found with contraband than male suspects ($p = 0.032$); Black suspects were only 0.2 percentage points more likely to be found with contraband than White suspects ($p = 0.592$); and Latinx suspects were 0.11 percentage points less likely to be found with contraband than non-Latinx suspects ($p = 0.832$).

This suggests near-zero disparity regarding Black suspects, substantially less (though still statistically significant) disparity regarding females versus males, and near-zero disparity regarding Latinx suspects. The anti-female disparity, while still present, is much smaller than in the structured-data-only model, and the pro-Black disparity nearly vanishes.

In summary, while the models trained only on structured data show notable policing disparities on the demographic variables we tested, particularly regarding female and Black suspects, the inclusion of free-form text data substantially reduces these disparity estimates. This suggests that the disparities observed with structured data alone may be partially explained by additional contextual information captured in text.

5 Analysis - Reasonable Suspicion

In addition to contraband discovery, a second outcome variable that we can test is whether or not the police had a reasonable suspicion that a crime was being or had been committed, as required under *Terry v. Ohio*. Using ablation to explain the predictions of a fine-tuned LLM, we find that the language most associated with legal stops indicates the clear commission of a specific crime, whereas the language most associated with illegal stops is circumstantial evidence (e.g., the suspect's standing in an area where narcotics use is common) and misunderstandings about the suspect's conduct (e.g., the police believing a can was open when it was closed).

5.1 LLM Fine-Tuning and Ablation

Under the settlement with the Philadelphia Police Department from which the data derives, lawyers manually read each police report from a random subset of stops and code the stop as either having reasonable suspicion or lacking reasonable suspicion. After dropping observations which could not be clearly coded, the random subset comprises 37,082 observations of stops that were coded as clearly having or lacking reasonable suspicion. The lawyers coding the stops were given the full text of the police report, without redaction; however, they were instructed that the police should have had sufficient reasonable suspicion prior to the stop (meaning that a stop lacking

reasonable suspicion could not become justified ex post, merely because the suspect turned out to be engaging in a crime).

We fine-tuned Llama 3 to predict the human evaluations of reasonable suspicion, using the same fine-tuning methods described in Subsection 4.1.1 Performance statistics for the model are below in Table 3 (testing set only). Like the humans lawyers coding manually, Llama was given access to the full text of the police reports.

Table 3: Performance Statistics for Reasonable Suspicion Prediction Model

Metric	**Value**
Accuracy	0.872
Precision	0.807
Recall	0.777
F1 Score	0.791
Efron's R^2	0.432
Mean Absolute Deviation	0.168
Mean Squared Deviation	0.092

After fine-tuning the model, we used ablation to identify what parts of the police reports were most predictive of a human judgment that the stop either had or lacked reasonable suspicion.[11] To conduct ablation, we split each police report into phrases delimited by periods, exclamation marks, question marks, commas, semicolons, and colons. We then used the fine-tuned model to predict the probability that a stop would be coded as having reasonable suspicion after sequentially removing each phrase. To illustrate, each colored phrase would be analyzed separately in the following example:

ABOVE MALE OBSERVED URINATING ON HIGHWAY, STOPPED FOR INVEST, NCIC/PCIC CHECKED, NO WANTS, CBI

The ablation impact score Δ_i for each phrase i is calculated as:

$$\Delta_i = \log P(y = 1|x) - \log P(y = 1|x_{-i}) \tag{15}$$

[11] We also tested LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017), and LASSO-regularized logistic regression with TF-IDF (Tibshirani, 1996). We found that LIME and SHAP produced unintuitive and unexplainable results (with high scores assigned to seemingly random words). We found that logistic regression produced more intuitive results that were qualitatively similar to the results from ablation analysis, but the use of n-grams failed to take into account important contextual information.

$$\Delta_i^{\text{linear}} = 100 \cdot (P(y = 1|x) - P(y = 1|x_{-i})) \tag{16}$$

where x is the full text and x_{-i} is the text with phrase i removed. This difference for each of the phrases in the example is as follows:

Table 4: Impact of Phrase Removal on Model Predictions

Phrase	Δ_i **(log)**	Δ_i^{linear} **(%)**
ABOVE MALE OBSERVED URINATING ON HIGHWAY	5.000	99.28
STOPPED FOR INVEST	0.021	2.00
NCIC/PCIC CHECKED	0.010	1.00
NO WANTS	0.009	0.79
CBI	0.007	0.75

As Table 4 shows, the phrase describing the suspect urinating on the highway is responsible for the vast majority of the model's prediction that the officer had reasonable suspicion that the suspect was committing a crime (i.e., urinating in public). With that phrase, the model's prediction that the report will be coded to give rise to reasonable suspicion is near 100%; without it, the prediction is near 0%. As one would intuitively expect, the other phrases have near-zero predictive power.

5.2 Results

Table 8 contains the 20 phrases associated with the largest increases in the predicted likelihood that a report will be coded as having reasonable suspicion, and Table 9 contains the inverse, the 20 phrases associated with the largest decreases. Note that the phrases in both tables are the unprocessed, unredacted contents of police reports (including any oddities in formatting in the original reports), because this was the information provided to the human coders.

Table 8 reveals a variety of clearly illegal conduct seen to justify a stop: smoking indoors, public urination, drinking from open containers on public roads or highways, or engaging in other dangerous conduct (gambling, wandering intoxicated) on the highway. Table 9 mostly contains descriptions of pretextual stops based on generalized suspicions–for example, suspects merely being present "IN A VACANT LOT

KNOWN FOR NARCOTICS USE," or merely sitting in an area where an ordinance about sidewalk behavior is posted. In addition, some of the observations were cases where the officer's suspicions turned out to be unjustified–as when a suspect was thought to be drinking alcohol but turned out to be "DRINKING AN ENERGY DRINK." The latter case is potentially a problem with the human coding, since in theory the presence or absence of reasonable suspicion is supposed to be determined solely based on information available before a stop is made.

To evaluate the LLM's performance at predicting reasonable suspicion, we employ Receiver Operating Characteristic (ROC) analysis, a standard method for assessing binary classifiers. The ROC curve in Figure 6 plots the true positive rate (sensitivity) against the false positive rate (1 - specificity) at various classification thresholds. The Area Under the Curve (AUC) of 0.902 indicates that the model has strong predictive power—for any randomly chosen pair of stops, one with reasonable suspicion and one without, the model will correctly rank them 90.2% of the time.

The ability of algorithms to do well at classifying police narratives as having or lacking reasonable suspicion has important implications. Most immediately, if the algorithms can achieve a high rate of reliability, they could save substantial time on a task that typically requires some legal training. If sufficiently reliable, our method could provide live feedback to officers before making stops, advising them of whether the stops are likely illegal.

Ultimately, the decision of whether to employ an algorithm will depend on what levels are deemed acceptable by those charged with making the assessments.[12] Even if it is not possible to classify all narratives with a high enough level of confidence, it may be possible to reduce the workload for human coders substantially by identifying narratives that can be classified with very high confidence.

One can also imagine a second and more ambitious use of police narrative classification algorithms. Police officers could describe their justification for a stop in real time, and algorithm could provide immediate feedback about the legality of the stop. This could help prevent stops or frisks where the officer is unsure whether the facts rise to the level of reasonable suspicion. It will not likely be available in all cases, as in a swiftly developing or chaotic situation. But even if only applicable some of the time, it could help eliminate a large proportion of illegal stops.

[12]In future work, we plan to explicitly explore the efficacy of various algorithms on this task and compare error rates to inter-rater disagreement.

One crucial question barrier to practical implementation, however, is whether we can identify some subset of cases for which the model is sufficiently accurate that police departments would feel comfortable relying on it. Figure 7 provides additional insight into the model's performance by examining how accuracy varies with prediction confidence.

To generate Figure 7, we first fit a quadratic model of prediction error using the LLM's predicted probability of reasonable suspicion. The quadratic specification captures the intuition that prediction confidence should be lowest near 50% probability and increase toward both extremes. We then ranked cases by their predicted error and plotted how accuracy changes as cases with lower predicted accuracy are added to the sample. The resulting concave, downward-sloping curve demonstrates that the model can maintain 95% accuracy while automatically classifying nearly three-quarters (73.9%) of cases. This suggests substantial potential for automated classification in practice—the model could reliably handle the majority of straightforward cases, providing important context to officers otherwise prone to much higher rates of error (20.5% in our sample).

6 Limitations and Robustness Checks

6.1 Confounders in Police Report Descriptions

Because so much of our analysis relies on descriptions written by police officers, a natural concern is that the descriptions themselves exhibit bias that confounds our estimates of disparities. It could be, for example, that when police see a suspect smoking something hand-rolled, they might describe it as "likely tobacco" if the suspect is White and "likely marijuana" if the suspect is Black. Or, to take another example, police might be concerned about being accused of gender bias and therefore devote extra care to making their report sound suspicious when stopping a woman. In either case, bias in natural language descriptions would serve as a confounder in our analysis.

We can test this possibility in the OLS regression that incorporates predicted probabilities, by interacting the predicted probability with each of the variables of interest—i.e., generating an interaction between the indicator variable for "Female",

"Black", etc. with the predicted probabilities.[13] Table 7 shows the results of this analysis. The coefficient for female suspects is statistically significant at 99% (p = 0.005), while the coefficients on the interactions for Black and Latinx suspects are statistically insignificant (p = 0.122 and p = 0.400, respectively). The magnitude of the coefficients is -0.0779 for female, 0.0337 for Black, and 0.0241 for Latinx, respectively; note that this term is a multiplier against the value of the prediction coefficient which itself has a mean value of 0.075.

A negative coefficient on the interaction term for female suspects implies that the same report is less likely to lead to contraband discovery for female suspects. We do not, however, know the underlying cause of this effect. It could be that female suspects are inherently less likely to carry contraband even on an objectively neutral accounting of the facts of a given stop; it could also be that police write reports that make female suspects appear more suspicious on an objectively equivalent set of facts. Whatever the reason, incorporating an interaction term further attenuates the finding of disparities based on gender; the (non-interaction) coefficient falls from -0.0089 ($p = 0.032$) to -0.0040 ($p = 0.374$) when including interaction terms. Thus, with respect to the gender disparity findings, our main results might be treated as a lower bound on the amount of attenuation that should occur when incorporating natural language controls.

6.1.1 Just-So Reports and Coefficient Attenuation

Even if police reports exhibit no disparities, controlling for their content could inappropriately attenuate coefficient estimates in contraband discovery if the police tend to write just-so reports, reshaping the narrative in their police reports after the fact to make the stop seem justifiable.[14]

Note that this is not a concern if police merely engage in across-the-board puffery–

[13] [[Another potential method to assess bias, which we have not conducted yet, is to fine-tune a model on the contents of police reports to see if it is possible to predict the race of suspects (or any other characteristic hypothesized to be subject policing disparities). We can facially evaluate whether the contents of the reports are predictive; if they are, we can then apply explainability tools like ablation, LIME, and SHAP to identify which characteristics of the reports are most predictive. In doing so, we can see whether there is telltale or "dog whistle" language that police are more likely to use in reports for certain kinds of suspects.]]

[14] [[Note: We would welcome any suggestions about methods to evaluate the extent to which this is happening!]]

for example, if the police were always to make events sound 10 percentage points more likely to lead to contraband discovery, the fine-tuned LLM would account for this, since its predictions are ultimately rooted in actual contraband results and it would simply discount for the 10 percentage point puffery in its predicted probabilities. Because our dataset consists only of cases where stops were made, police have a consistent incentive to give the appearance of reasonable suspicion, which would tend to give rise to level bias controllable by the LLM's training.

But a more severe problem exists when different sorts of stops are differently misreported. Here, too, certain directions of misreporting are less problematic. If police were to take greater care to make stops seem justified when no contraband is ultimately found (because they might think the contraband speaks for itself in cases where it is found), this would simply reduce the predictive power of the model and make it a less effective control. On the other hand, if police were to distort their reporting to make stops seem more justified in cases where contraband was found (i.e. in cases where stops really *were* justified), that would essentially turn the LLM control into an "over-control" and lead to attenuation of the magnitude of other coefficients in the OLS regression.

There is some evidence that police either do not try to or are not very good at misreporting in general, like the frequency with which police make stops that are later judged to lack reasonable suspicion (20.5%), and much of the existing literature (which often extracts simple controls from stop narratives) operates under the same assumption that stop narratives accurately reflect what happened. Moreover, the fact that we do not see attenuation across the board (the coefficient for the Latinx indicator variable increases in magnitude) is some evidence against the potential for attenuation from just-so reporting. However, the possibility remains.

To test the possibility that certain police reports might implicitly signal that a suspect was carrying contraband, and to validate the redactions in Subsection 4.1.3, we conducted the same ablation analysis described in Subsection 5.1, identifying the phrases in the redacted police reports that were most predictive of contraband discovery. The results in Table 10 show the 20 phrases among all of the police reports whose removal decreased the predicted likelihood of contraband discovery the most. The results are sensible. Some reflect a type of policing associated with high likelihood of recovering contraband, like "working the Eagles detail with L and I"–this is a reference to the Licenses and Inspections patrol, which confiscates counterfeit and un-

licensed merchandise as contraband. The others generally involve visual descriptions of contraband, like seeing an individual discard a firearm (firearms are considered contraband) or exchanging marijuana for money.

Although this analysis might miss subtler signals of stop outcomes, it at least suggests that obvious information about stop outcomes has been removed, and it does not suggest officer misreporting.

7 Conclusion

This paper demonstrates the significant impact that LLMs can have when analyzing police stop data for potential racial disparities. By leveraging the full text descriptions provided by officers rather than just numerical and categorical data, LLMs can produce substantially different results, in this case causing apparent policing disparities to decrease or disappear. This finding highlights the importance of considering free-form text in analyses of policing practices and casts doubt on some prior conclusions regarding policing disparities that did not incorporate such textual information.

References

Anwar, S. and Fang, H. (2006). An alternative test of racial prejudice in motor vehicle searches: Theory and evidence. *American Economic Review*, 96(1):127–151.

Bai, Y., Chen, F., Wang, H., Xiong, C., and Mei, S. (2023). Transformers as Statisticians: Provable In-Context Learning with In-Context Algorithm Selection. In *37th Conference on Neural Information Processing Systems*.

Becker, G. S. (1957). *The economics of discrimination*. University of Chicago Press.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Choi, J. H. and Connell, P. (2024). Llms for causal inference. Working Paper.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

Engel, R. S. (2008). Revisiting the use of citizen demeanor by police officers: An empirical assessment. *Crime Delinquency*, 54(4):600–636.

Engel, R. S. and Johnson, R. (2006). Toward a better understanding of racial and ethnic disparities in search and seizure rates. *Journal of Criminal Justice*, 34(6):605–617.

Engel, R. S. and Tillyer, R. (2008). A critique of the "outcome test" in racial profiling research. *Justice Quarterly*, 25(1):1–36.

Frost, J. (2020). Attenuation bias in regression. *Statistics By Jim*.

Hernandez-Murillo, R. and Knowles, J. (2004). Racial profiling or racist policing? bounds tests in aggregate data. *International Economic Review*, 45(3):959–989.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., and Chen, W. (2022). Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 4(6):506–518.

Knowles, J., Persico, N., and Todd, P. (2001). Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy*, 109(1):203–229.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30:4765–4774.

Persico, N. and Todd, P. E. (2006). Racial profiling? detecting bias using statistical evidence. *Annual Review of Economics*, 1:229–254.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.

Ridgeway, G. (2007). Analysis of racial disparities in the new york police department's stop, question, and frisk practices. *RAND Corporation*.

Siegelmann, H. T. and Sontag, E. D. (1995). Computational capabilities of recurrent narx neural networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 27(2):208–215.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tillyer, R. and Klahm, C. F. (2011). Searching for contraband: Assessing the use of discretion by police officers. *Police Quarterly*, 14(2):166–185.

Zhang, J., He, C., Dai, Z., Gu, Q., Carpenter, B., Dyer, C., and Neubig, G. (2022). Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

A Tables and Figures

A.1 Tables

Variable	Mean	Standard Deviation
Contraband Discovered	0.073	0.254
Reasonable Suspicion for Stop	0.570	0.495
Male	0.859	0.348
Black	0.702	0.458
White	0.281	0.450
Latinx	0.096	0.294
Age	33.268	13.274
Height	5.517	0.425
Weight	170.543	33.598
Year	2018.890	2.402
Month	5.891	3.052
Evening	0.566	0.496
Daytime	0.314	0.464
Night	0.120	0.325
Officer with Partner	0.744	0.436
LLM Contraband Prediction	0.075	0.176

Table 5: Summary statistics for key variables in the dataset of Philadelphia police stops from 2016 to 2023. The table reports the mean and standard deviation for the following variables: *Contraband Discovered* (indicator variable equal to 1 if contraband was found, 0 otherwise), *Reasonable Suspicion for Stop* (indicator variable equal to 1 if reasonable suspicion was present, 0 otherwise), *Gender* (*Male* indicator variable equal to 1 if the suspect is male, 0 otherwise), *Race* (*Black* and *White* indicator variables), *Ethnicity* (*Latinx* indicator variable equal to 1 if the suspect is Latinx, 0 otherwise), *Age* (in years), *Height* (in feet), *Weight* (in pounds), *Year* (calendar year of the stop), *Month* (month of the stop), *Time of Day* (*Evening*, *Daytime*, and *Night* indicator variables), *Officer with Partner* (indicator variable equal to 1 if the officer had a partner during the stop, 0 otherwise), and *LLM Contraband Prediction* (the predicted probability of contraband discovery from the LLM model). The sample includes 67,469 observations.

Regression Type	R-squared		MSE	
	Train	Test	Train	Test
Key Variables	0.0026	0.0006	0.0707	0.0769
Basic Controls	0.0207	0.0165	0.0689	0.0764
Full OLS Controls	0.1559	0.0263	0.0594	0.0756
Full OLS Controls + LLM	0.4259	0.2368	0.0404	0.0593

Table 6: Performance metrics (R-squared and Mean Squared Error) for different OLS regression models predicting contraband discovery. The models include: (1) **Key Variables** only, which includes indicator variables for *Female* (1 if suspect is female, 0 otherwise), *Black* (1 if suspect is Black, 0 otherwise), and *Latinx* (1 if suspect is Latinx, 0 otherwise); (2) **Key Variables plus Basic Controls**, which adds controls for *Age* (in years), *Height* (in feet), *Weight* (in pounds), *Build* (categorical variable for body type), *Officer with Partner* (indicator variable), *Time* (minutes since the year 2000), *Month* (month fixed effects), and *Time of Day* (fixed effects for *Evening*, *Daytime*, and *Night*); (3) **Full OLS Controls**, which adds *Officer Fixed Effects* and *Police Service Area (PSA) Fixed Effects*; and (4) **Full OLS Controls with LLM Predictions**, which includes the LLM-predicted probability of contraband discovery based on text data. Performance metrics are reported for both training and test datasets. The training set includes 80% of the data; the test set includes the remaining 20%.

	Key Variables	**Key + Basic Control**	**All Structured Variables**	**All Structured Variables + Predicted**	**All Structured Variables + Predicted + Interactions**
Female	-0.0280***	-0.0285***	-0.0222***	-0.0089**	-0.0040
	(0.0044)	(0.0049)	(0.0050)	(0.0041)	(0.0045)
Black	0.0181***	0.0129***	0.0135***	0.0020	-0.0005
	(0.0038)	(0.0038)	(0.0045)	(0.0037)	(0.0040)
Latinx	0.0262***	0.0204***	0.0087	-0.0011	-0.0028
	(0.0058)	(0.0059)	(0.0062)	(0.0051)	(0.0057)
Female × Prediction					-0.0779***
					(0.028)
Black × Prediction					0.0337
					(0.022)
Latinx × Prediction					0.0241
					(0.029)

$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table 7: OLS regression results for the likelihood of contraband discovery during police stops, using different model specifications. The dependent variable is an indicator equal to 1 if contraband was discovered, 0 otherwise. The table reports coefficient estimates and standard errors (in parentheses) for key demographic variables (*Female*, *Black*, *Asian*, *Latinx*), with different sets of controls: (1) **Key Variables** only, which includes the demographic indicators; (2) **Key Variables plus Basic Controls**, which adds controls for *Age* (in years), *Height* (in feet), *Weight* (in pounds), *Build* (categorical variable for body type), *Officer with Partner* (indicator variable), *Time* (minutes since the year 2000), *Month* fixed effects, and *Time of Day* fixed effects (*Evening*, *Daytime*, *Night*); (3) **Full OLS Controls**, which adds *Officer Fixed Effects* and *Police Service Area (PSA) Fixed Effects*; (4) **Full OLS Controls with LLM Predictions**, which includes the LLM-predicted probability of contraband discovery based on text data; and (5) **Full OLS Controls with LLM Predictions and Interactions**, which adds interactions between the LLM predictions and the demographic variables. Statistical significance levels are indicated by *, **, and *** for $p < 0.1$, $p < 0.05$, and $p < 0.01$, respectively.

Table 8: Ablation Results for Reasonable Suspicion - Positive

Phrase	Δ_i **(log)**	Δ_i^{linear} **(%)**
POLICE OBSERVED MALE INSIDE OF THE CHINESE STORE AT ABOVE LOCATION SMOKING A CIGARETTE INSIDE AND WAS NOT BUYING ANYTHING.	5.998	99.60
MALE OBSERVED URINATING ON THE HIGHWAY AT ABV LOC.	5.500	99.55
MALE WAS OBSERVED BY POLICE URINATING ON THE HWY AT ABOVE LISTED LOCATION.	5.500	99.55
POLICE OBSERVED LISTED MALE ALONG WITH ANOTHER MALE SITTING ON SIDEWALK DRINKING 16 OZ CANS OF HURRICANE MALT LIQUOR.	5.624	99.52
MALE OBSERVED WITH OPEN CAN OF BEER ON HIGHWAY.	5.499	99.50
POLICE OBSERVED MALE DRINKING AN OPEN CONTAINER OF BEER ON THE HIGHWAY AT ABOVE LOCATION,	5.623	99.42
POLICE OBSERVED MALE THROWING TRASH ON THE HIGHWAY.	5.623	99.42
POLICE OBSERVED LISTED MALE URINATING BEHIND THE WALL AT ABOVE LOCATION.	5.249	99.40
MALE OBSERVED GAMBLING ON THE HIGHWAY.	5.498	99.37
abv female was seen on the hwy with a open container of beer drinking.	5.373	99.37
OFFENDER OBSERVED URINATING ON THE HWY AT ABV LOCATION.	5.125	99.36
POLICE OBSERVED MALE URINATING ON THE HIGHWAY.	5.124	99.35
W/M WAS OBSERVED BY POLICE URINATING ON THE HIGHWAY AND WAS STOPPED FOR INVESTIGATION.	5.124	99.31
MALES OBSERVED GAMBLING ON THE HIGHWAY.	5.497	99.31
ABV MALE WAS SEEN FINISHING URINATING OUT IN PUBLIC ON THE SIDE OF THE BUILDING TOWARDS THE ALLEY WAY AT ABV LOC,	5.124	99.30
police observed male publically urinating on the highway at listed location.	5.000	99.30
ABOVE MALE OBSERVED URINATING ON HIGHWAY,	5.000	99.28
male observed urinating in alleyway.	5.248	99.28
POLICE OBSERVED FEMALE APPEARING INTOXICATED ON HIGHWAY.	5.746	99.27
POLICE OBSERVED LISTED MALE URINATING INSIDE THE ALLEY WAY.	5.124	99.27

Table 8: 20 phrases from police reports most strongly associated with lawyers coding the stop as having reasonable suspicion, ranked by their impact on model predictions. The table reports results from ablation analysis of a fine-tuned large language model trained to predict whether lawyers would code a stop as having reasonable suspicion. For each phrase, Δ_i reports the log difference in predicted probability when the phrase is removed, and Δ_i^{linear} reports the linear (percentage point) difference. The model was trained on 37,082 police reports that were manually coded by lawyers as part of a settlement agreement monitoring the Philadelphia Police Department. Higher values indicate that removing the phrase more strongly decreased the model's prediction that the stop would be coded as having reasonable suspicion. Phrases are shown exactly as they appeared in the original police reports, including any capitalization or formatting irregularities.

Table 9: Ablation Results for Reasonable Suspicion - Negative

Phrase	Δ_i **(log)**	Δ_i^{linear} **(%)**
POLICE OBSERVED A GROUP OF PEOPLE IN A VACANT LOT KNOWN FOR NARCOTICS USE.	-2.770	-89.80
MALE WAS DRINKING AN ENERGY DRINK.	-2.700	-83.29
P/O [OFFICER] AND PARTNER P/O [OFFICER] WORKING AS N120 OBSERVED TWO H/M MALES ARGUING ON THE HIGHWAY.	-2.686	-82.05
POLICE FOUND CAN TO BE SEALED.	-2.795	-81.38
LISTED WAS SITTING ON THE STEPS OF A CLOSED BUSINESS.	-2.561	-81.26
UPON ARRIVAL THE MALE WALKED AWAY APPROX 30 MIN LATER THE SAME MALE WAS OBSERVED SITTING AT THE SAME LOCATION ND WAS STOPPED FOR INVESTIGATION,	-1.677	-79.20
POLICE FOUND CAN TO BE SEALED.	-2.309	-76.77
AS A RESULT OF MULTIPLE COMPLAINTS FROM STORE OWNERS OF CROWDS OF NE'ER-DO-WELLS CONVENING INSIDE AND AROUND THEIR STORES FIGHTING,	-1.614	-73.31
MATCHED FLASH FOR PER WITH GUN.	-1.965	-73.30
MALE CBI	-2.686	-72.39
POL REC RC FOR MALES SELLING NARCOTICS ON HWY AT ABV LOC.	-1.771	-71.93
THROUGH INVESTIGATION MALE WAS DRINKING AN ICED TEA.	-3.576	-71.02
ABOVE FEMALE ALONG WITH ANOTHER FEMALE WERE SITTING ON STEPS OF ABOVE PROPERTY WHICH WAS GETTING NUMEROUS COMPLAINTS AND OF DRUG USE AND PROSTITUTION.	-1.274	-69.35
HE FIT THE FLASH AND WAS INVESTIGATED.	-1.411	-69.22
UPON FURTHER INVESTIGATION IT WAS OBSERVED TO BE A ARIZONA FRUIT DRINK.	-1.814	-68.47
OBJECT REVEALED TO BE AN ANKLE BRACELET.	-2.092	-68.11
POLICE OBSERVE MALE AND FEMALE SITTING ON TOP OF HILL AT ABOVE LOCATION WHICH IS BETWEEN [STREET] AND [AVENUE] WHERE CITY CODE 10-611 IS POSTED.	-1.899	-67.89
MALE CBI	-2.188	-67.02
POLICE OBSERVE A GROUP OF MALES SITTING ON TOP OF HILL AT ABOVE LOCATION WHICH IS BETWEEN [STREET] AND [AVENUE] WHERE CITY CODE 10-611 IS POSTED.	-1.953	-64.79
MALE OBSERVED FOR HOURS IN FRONT OF [SPECIFIC] CHINESE FOOD STORE AND WAS ADVISED TO MOVE ALONG MULTIPLE TIMES AS ITS A HIGH CRIME/DRUG AREA AND HE WASN'T MAKING A PURCHASE.	-2.338	-63.82

Table 9: 20 phrases from police reports most strongly associated with lawyers coding the stop as lacking reasonable suspicion, ranked by their impact on model predictions. The table reports results from ablation analysis of a fine-tuned large language model trained to predict whether lawyers would code a stop as having reasonable suspicion. For each phrase, Δ_i reports the log difference in predicted probability when the phrase is removed, and Δ_i^{linear} reports the linear (percentage point) difference. The model was trained on 37,082 police reports that were manually coded by lawyers as part of a settlement agreement monitoring the Philadelphia Police Department. More negative values indicate that removing the phrase more strongly increased the model's prediction that the stop would be coded as lacking reasonable suspicion. Phrases are shown exactly as they appeared in the original police reports, including any capitalization or formatting irregularities.

Table 10: Ablation Results for Contraband Discovery

Phrase	Δ_i **(log)**	Δ_i^{linear} **(%)**
While working Eagles detail with L and I,	3.226	90.26
Police obs the individual holding a green ziplock packet containing marijuana.	2.541	88.68
observed selling unlicensed NFL merchandise on the lot.	2.321	87.53
Police observed an individual discard a firearm and then jump over a wall.	2.733	86.41
While working the game with L and I,	2.517	86.41
At that time police observed the above listed offender turn their body away from police and pull a firearm from their waist band placing the firearm on the stairs to a residence.	3.891	86.28
Police obs the individual holding a red tinted tube containing marijuana.	2.479	82.88
As police got closer to the pickup truck an individual began to kneel down with their left hand raised and their right hand lowered with a tan colored gun in it.	3.370	82.26
Individual observed by plainclothes police ([officer number]) buying narc.	2.452	80.49
While working the Eagles detail with L and I,	1.678	79.46
While working the Eagles game with L and I,	1.678	79.46
Police then observed that individual toss two clear bags with white substance inside.	2.652	79.19
Above individual was observed by police engaged in narc.	3.087	78.02
Above individual was observed by police handing marijuana to individual #2 in exchange for USC.	2.419	77.61
Above was obs by plainclothes officer in illegal narc trans.	2.419	77.61
Police could see into the front hoodie pocket and notice a black firearm inside.	3.278	74.80
While on patrol I came upon an individual standing in the highway in possession of marijuana in their right open palm.	2.378	74.17
P/O [officer] and [another officer] observed an individual who was in the driver seat of a [car] smoking a marijuana cigarette and observed an individual put something in the glove compartment.	2.797	72.99
while chasing police observed the individual reaching in their waistband grabbing baggies containing a green leafy substance alleged marijuana.	1.602	72.22
P/O observed individual sitting in a [car] while on the cell phone and a clear plastic bag stamped "Donald Trump".	3.705	71.31

Table 10: 20 phrases from police reports most strongly associated with contraband discovery, ranked by their impact on model predictions. The table reports results from ablation analysis of a fine-tuned large language model trained to predict whether contraband would be discovered during a stop. For each phrase, Δ_i reports the log difference in predicted probability when the phrase is removed, and Δ_i^{linear} reports the linear (percentage point) difference. The model was trained on pre-processed police reports with demographic information and post-stop outcomes redacted. Higher values indicate that removing the phrase more strongly decreased the model's prediction that contraband would be discovered. Phrases are shown exactly as they appeared in the redacted police reports, including any capitalization or formatting irregularities after pre-processing.

A.2 Figures

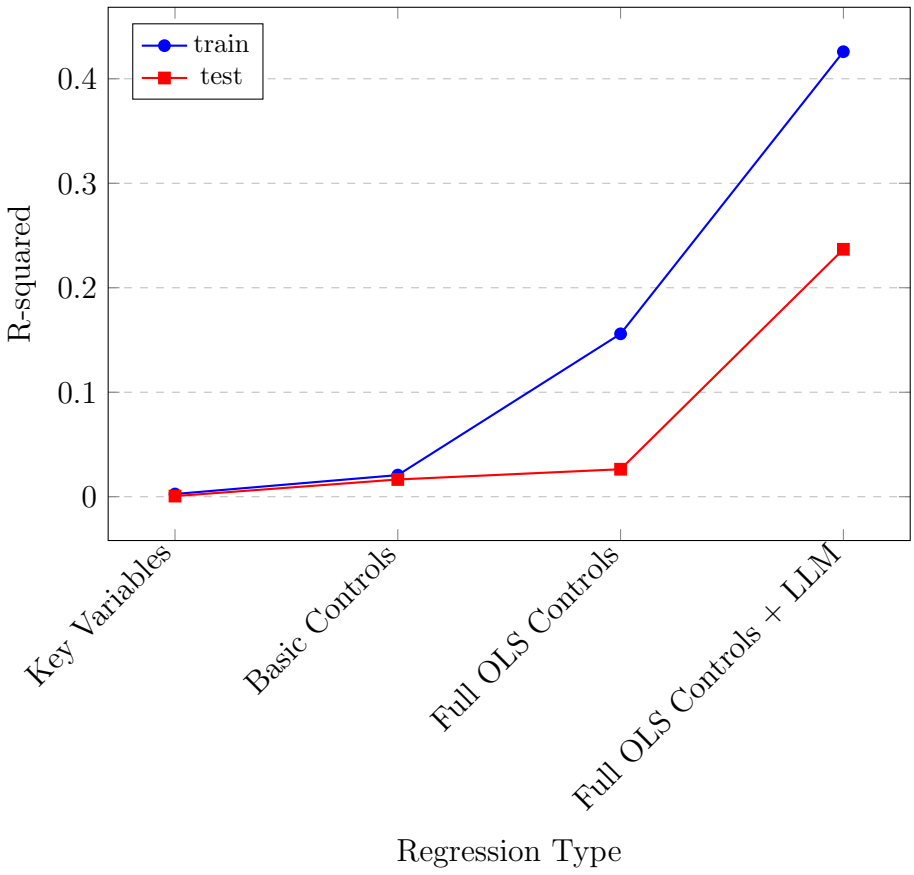


Figure 1: Comparison of R-squared values for different OLS regression models predicting contraband discovery, showing performance on both training and test datasets. The regression models include: (1) **Key Variables** only, which includes indicator variables for *Female*, *Black*, and *Latinx*; (2) **Key Variables plus Basic Controls**, which adds controls for *Age*, *Height*, *Weight*, *Build* (categorical variable), *Officer with Partner* (indicator variable), *Time* (minutes since the year 2000), *Month* fixed effects, and *Time of Day* fixed effects (*Evening*, *Daytime*, *Night*); (3) **Full OLS Controls**, which adds *Officer Fixed Effects* and *Police Service Area (PSA) Fixed Effects*; and (4) **Full OLS Controls with LLM Predictions**, which includes the LLM-predicted probability of contraband discovery based on text data. The training set includes 80% of the data; the test set includes the remaining 20%.

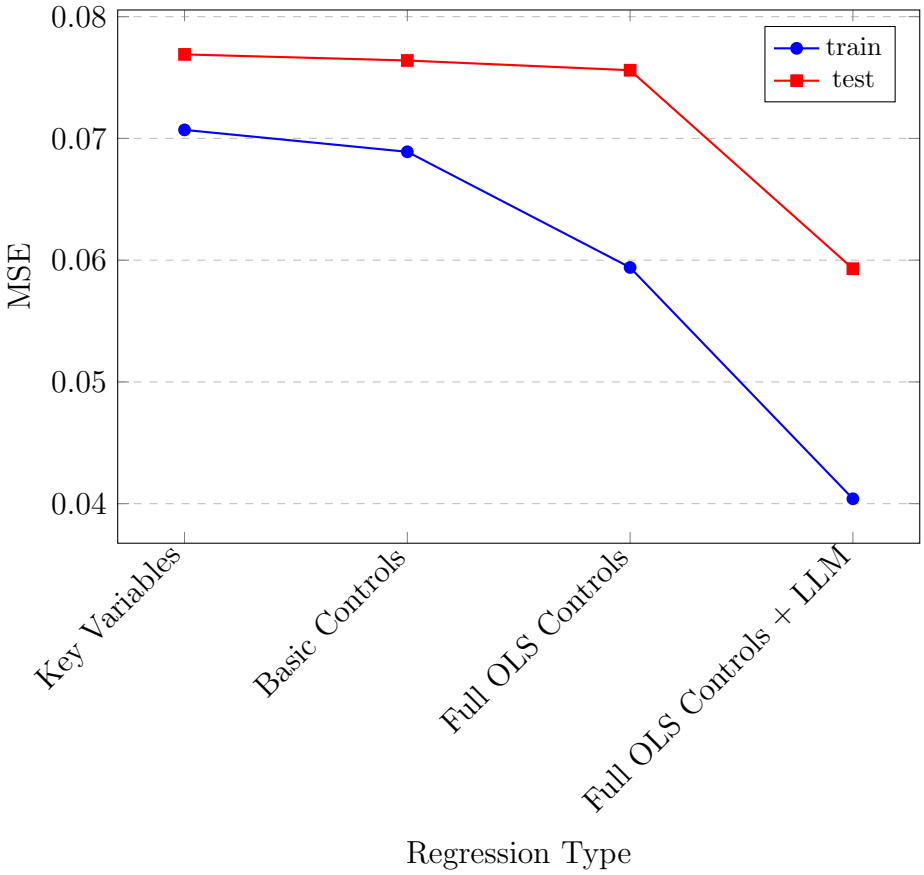


Figure 2: Comparison of Mean Squared Error (MSE) for different OLS regression models predicting contraband discovery, showing performance on both training and test datasets. The regression models include: (1) **Key Variables** only, which includes indicator variables for *Female*, *Black*, and *Latinx*; (2) **Key Variables plus Basic Controls**, which adds controls for *Age*, *Height*, *Weight*, *Build* (categorical variable), *Officer with Partner* (indicator variable), *Time* (minutes since the year 2000), *Month* fixed effects, and *Time of Day* fixed effects (*Evening*, *Daytime*, *Night*); (3) **Full OLS Controls**, which adds *Officer Fixed Effects* and *Police Service Area (PSA) Fixed Effects*; and (4) **Full OLS Controls with LLM Predictions**, which includes the LLM-predicted probability of contraband discovery based on text data. Lower MSE values indicate better predictive accuracy.

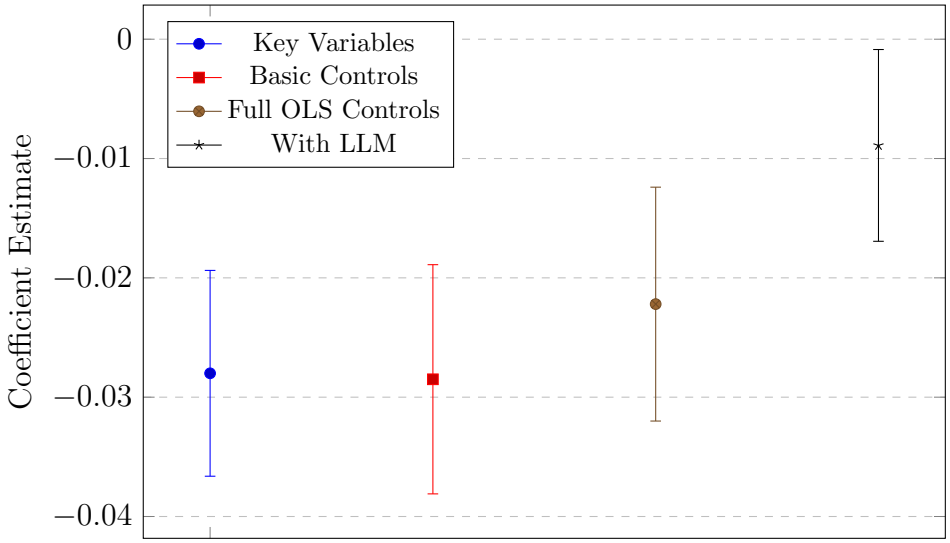


Figure 3: Coefficient estimates and 95% confidence intervals for the *Female* indicator variable in OLS regressions predicting contraband discovery, across different model specifications. The models include: (1) **Key Variables** only, which includes indicator variables for *Female*, *Black*, and *Latinx*; (2) **Key Variables plus Basic Controls**, which adds controls for *Age*, *Height*, *Weight*, *Build* (categorical variable), *Officer with Partner* (indicator variable), *Time* (minutes since the year 2000), *Month* fixed effects, and *Time of Day* fixed effects (*Evening*, *Daytime*, *Night*); (3) **Full OLS Controls**, which adds *Officer Fixed Effects* and *Police Service Area (PSA) Fixed Effects*; and (4) **Full OLS Controls with LLM Predictions**, which includes the LLM-predicted probability of contraband discovery based on text data.

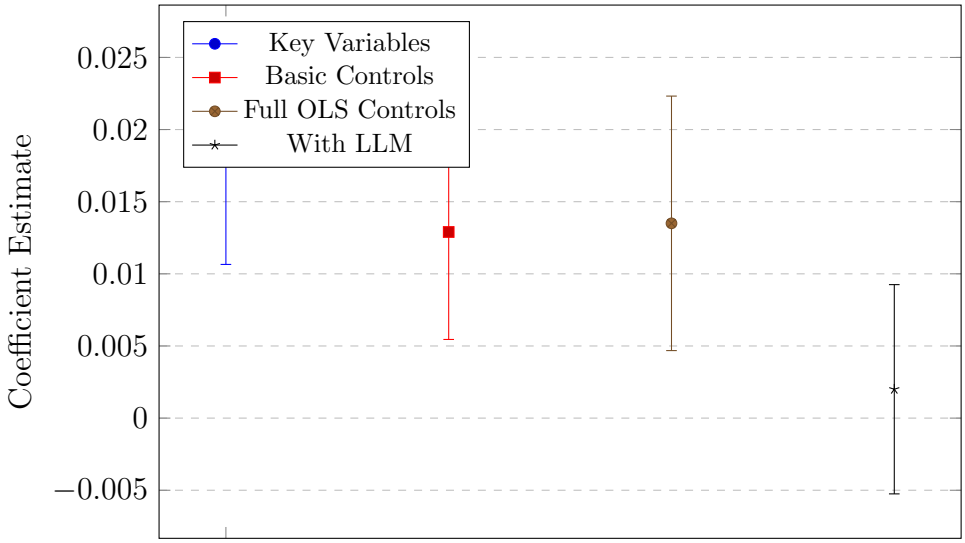


Figure 4: Coefficient estimates and 95% confidence intervals for the *Black* indicator variable in OLS regressions predicting contraband discovery, across different model specifications. The models include: (1) **Key Variables** only, which includes indicator variables for *Female*, *Black*, and *Latinx*; (2) **Key Variables plus Basic Controls**, which adds controls for *Age*, *Height*, *Weight*, *Build* (categorical variable), *Officer with Partner* (indicator variable), *Time* (minutes since the year 2000), *Month* fixed effects, and *Time of Day* fixed effects (*Evening*, *Daytime*, *Night*); (3) **Full OLS Controls**, which adds *Officer Fixed Effects* and *Police Service Area (PSA) Fixed Effects*; and (4) **Full OLS Controls with LLM Predictions**, which includes the LLM-predicted probability of contraband discovery based on text data.

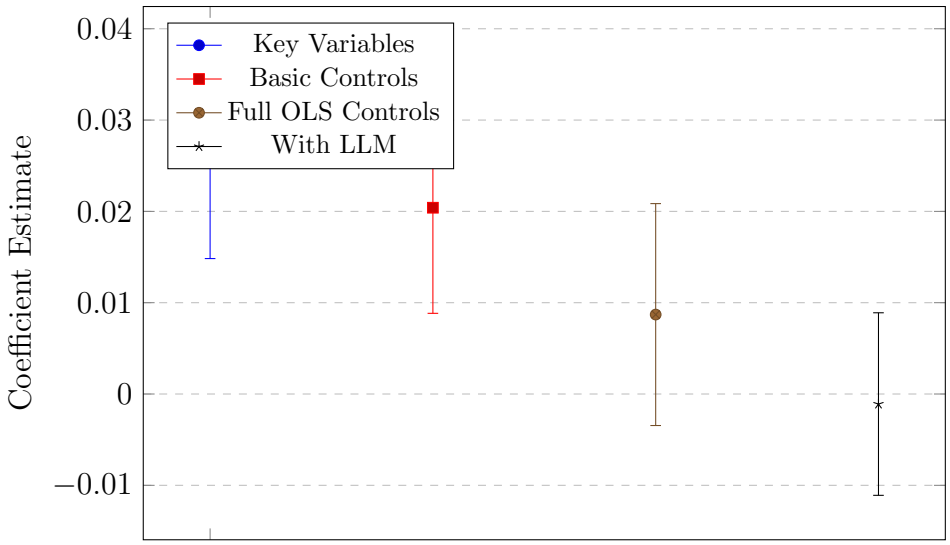


Figure 5: Coefficient estimates and 95% confidence intervals for the *Latinx* indicator variable in OLS regressions predicting contraband discovery, across different model specifications. The models include: (1) **Key Variables** only, which includes indicator variables for *Female*, *Black*, and *Latinx*; (2) **Key Variables plus Basic Controls**, which adds controls for *Age*, *Height*, *Weight*, *Build* (categorical variable), *Officer with Partner* (indicator variable), *Time* (minutes since the year 2000), *Month* fixed effects, and *Time of Day* fixed effects (*Evening*, *Daytime*, *Night*); (3) **Full OLS Controls**, which adds *Officer Fixed Effects* and *Police Service Area (PSA) Fixed Effects*; and (4) **Full OLS Controls with LLM Predictions**, which includes the LLM-predicted probability of contraband discovery based on text data.

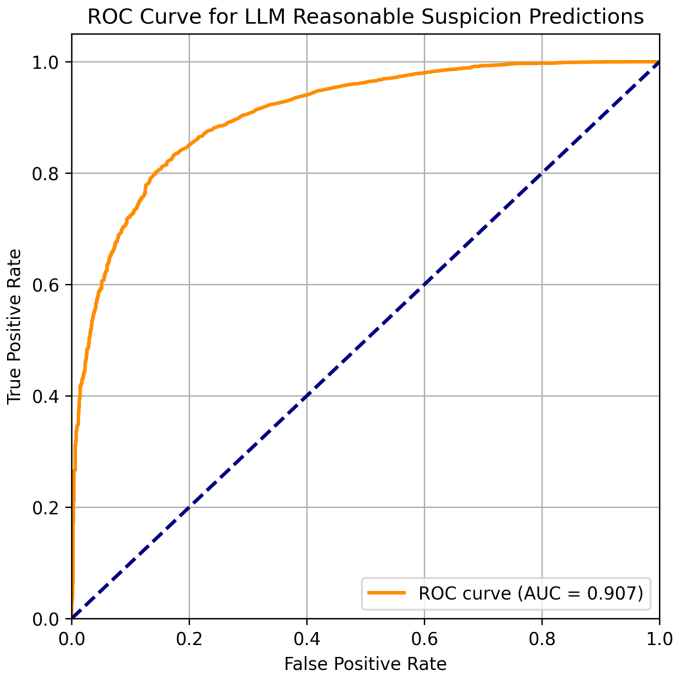


Figure 6: Receiver Operating Characteristic (ROC) curve for LLM predictions of reasonable suspicion. The curve plots the true positive rate against the false positive rate at various classification thresholds. The Area Under the Curve (AUC) of 0.907 indicates strong predictive performance, significantly better than random chance (which would have an AUC of 0.5). The curve was plotted based on performance in the validation test set.

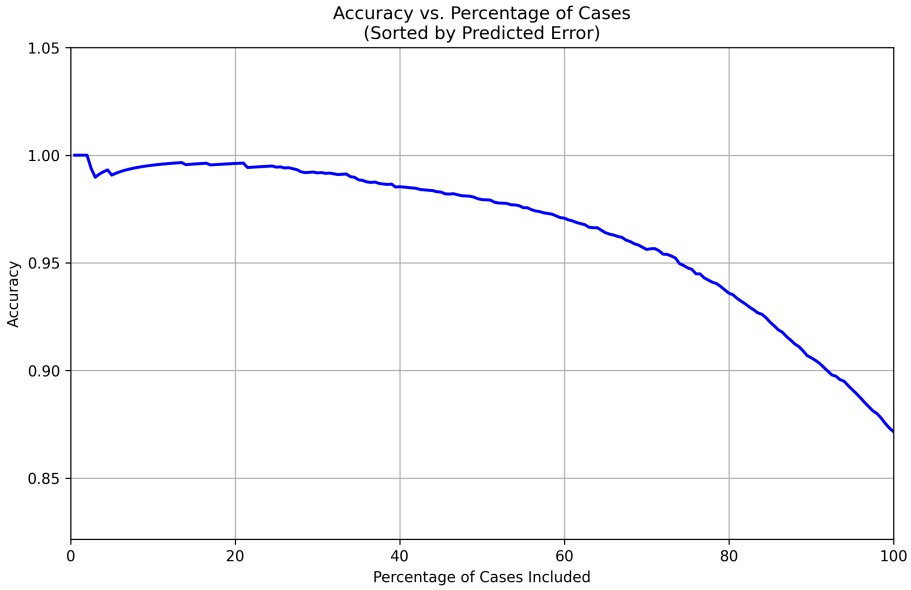


Figure 7: Accuracy of LLM predictions as a function of the percentage of cases included, sorted by predicted error. The model achieves 95% accuracy on approximately 73.9% of cases when predictions are ranked by confidence level, suggesting strong potential for automated classification of a substantial portion of cases. Accuracy was calculated with respect to the validation test set.

B Appendix

B.1 Data Processing

The dataset underwent several preprocessing steps to ensure data quality and consistency. The following procedures were applied:

1. **Standardizing Variables**: Variables were standardized to decrease the number of discrete values (e.g., mapping "Y", "Yes", "yes" to "Yes").
2. **Stop Reason**: The 'STOPRS' column was converted to uppercase and mapped to 'Yes', 'No', or 'N/A'.
3. **Age Processing**:
    - Ages were converted to numeric values.

- Observations with ages less than 9 or greater than 99 were dropped.
- Negative ages and age 0 were dropped.

4. **Physical Characteristics**:
    - Observations with height values outside the range of 4'6" to 7' were dropped.
    - Observations with weight values less than 50 lbs or greater than 500 lbs were dropped.
    - Observations with specific invalid weights (501, 511, 601, 999) were dropped.
    - Observations with unrealistic weight-build combinations (e.g., 'Thin' with weight greater than 300 lbs) were dropped.
5. **Date and Time**:
    - A 'minutes_since_2000' variable was created as a linear time control.
    - A Month variable was created.
    - Time of day was categorized as 'Night' (0:00-8:00 hours), 'Day' (8:00-16:00 hours), or 'Evening' (16:00-24:00 hours).
6. **Location Information**:
    - Police Service Area (PSA) 7700 was removed.
7. **Text Cleaning**:
    - 'apos;' was replaced with apostrophes in 'REASON_DESC'.
    - Newline characters were replaced with spaces in 'REASON_DESC'.
8. **Rare Value Handling**:
    - Observations with rare values (frequency ¡ 0.5%) in 'EYE_COLOR' and 'BUILD' were dropped.
    - Observations involving officers with fewer than 10 stops were dropped.
9. **Duplicate Removal**: Duplicate rows were removed based on the following columns: "LOCATION", "OFFICER_PAYROLL", "FURTHER_DESC", "SEX", "AGE", "HEIGHT".

B.2 LLM Fine-Tuning Hyperparameters

We used the following hyperparameters and configurations to fine-tune Llama 3:

- **Contraband Discovery Prompt:** "Here is a police report provided by a Philadelphia police officer containing details about a pedestrian stop: POLICE REPORT. Based on the description in the police report above, predict whether the police officer found contraband on the suspect. Contraband includes weapons (legal or illegal), illegal drugs, stolen property, and large amounts of cash. Answer only 'Yes' or 'No'."
- **Reasonable Suspicion Prompt:** "Here is a police report provided by a Philadelphia police officer containing details about a pedestrian stop: POLICE REPORT. Based on the description in the police report above, did the officer have a reasonable suspicion that the suspect had committed, was committing, or was about to commit a crime, as required by Terry v. Ohio? Cases considered to give rise to reasonable suspicion include but are not limited to (1) curfew stops of persons 23 or younger and (2) stops where the suspects have contraband (contraband includes guns or other weapons, whether legal or illegal, illegal drugs, stolen property, and large amounts of cash). Answer only 'Yes' or 'No'."
- **Base Model:** We used a 4-bit quantized version of the Llama 3 8B Instruct model, which allows for faster loading and reduced memory usage.
- **Sequence Length:** The maximum sequence length was set to 2048 tokens.
- **Quantization:** We employed 4-bit quantization to reduce memory usage and enable faster training.
- **LoRA Configuration:**
  - Rank (r): 16
  - LoRA Alpha: 32
  - LoRA Dropout: 0
  - Bias: "none"
- **Training Configuration:**

  - Batch Size: 16 per device
  - Gradient Accumulation Steps: 1
  - Warmup Steps: 100
  - Number of Epochs: 3
  - Learning Rate: 0.0001
  - Optimizer: AdamW (8-bit)
  - Weight Decay: 0.01
  - Learning Rate Scheduler: Cosine

- **Precision:** We used mixed precision training, automatically selecting between FP16 and BF16 based on hardware support.
- **Gradient Checkpointing:** We used Unsloth for gradient checkpointing.

B.3 Pre-Processing Police Reports: LLM Model, Hyperparameters, and Prompts

All edits to the reports are made using GPT-4o-mini, with a final edit to ensure the data are not contaminated with post-stop information using GPT-4o (a slightly more powerful model). We use gpt-4o-mini-2024-07-18 and gpt-4o-2024-11-20, respectively, with 4096 maximum output tokens, temperature 0.000001, and all other hyperparameters set to defaults. We initially experimented with using GPT-4o to conduct all of the reformatting and redaction at once, but ultimately found that the redactions are more thorough when broken up into multiple steps, even using a cheaper model for some of the simpler steps.

We both reformat the police reports so that they are consistent (instead of some being all capitalized and some being sentence-cased), and then conduct the redactions. The following prompts were used in the following order (i.e., the output from applying GPT-4o-mini with each prompt was used as the input using the next prompt):

1. **Convert to Sentence Case:**

You are being given the contents of a police report describing a police stop.

If the report is in all-caps, convert it into sentence case. Return only the modified text, without any additional explanations or comments. Do not change the text at all except as specified above.

If no modifications are required, return the original text. If no text is given, just reply "N/A".

2. **Remove Post-Stop Information:**

   You are being given the contents of a police report describing a police stop.

   Edit the text so that it contains only the information the police would have known before making the stop. To do this, remove any discussion of the outcome of the police stop as well as any information the police would not have known before making the stop. For example, convert "Suspect appeared to be smoking marijuana, upon further investigation it was only tobacco" to "Suspect appeared to be smoking marijuana", because "further investigation" likely involved a stop. As another example, the sentence "Search of suspect revealed marijuana" should simply be deleted, because searches and arrests can only occur after a stop.

   Return only the modified text, without any additional explanations or comments. Do not change the text at all except as specified above. If no changes are required, just reply with the original text.

3. **Remove Gender Information:**

   You are being given the contents of a police report describing a police stop.

   Remove any information about the gender of the suspect. For example, convert "Male was found..." to "Individual was found...".

   Also replace any gendered pronouns with gender-neutral pronouns.

   Return only the modified text, without any additional explanations or comments. Do not change the text at all except as specified above. If no changes are required, just reply with the original text.

4. **Remove Race Information:**

   You are being given the contents of a police report describing a police stop.

   Remove any information about the race of the suspect. For example, convert "Black individual found..." to "Individual found...".

   Return only the modified text, without any additional explanations or comments. Do not change the text at all except as specified above. If no changes are required, just reply with the original text.

5. **Remove Hispanic/Latinx Information:**

   You are being given the contents of a police report describing a police stop.

   Remove any information about whether the suspect was Hispanic/Latinx or not. For example, convert "Hispanic suspect found..." to "Suspect found...".

   Return only the modified text, without any additional explanations or comments. Do not change the text at all except as specified above. If no changes are required, just reply with the original text.

6. **Remove Age Information:**

   You are being given the contents of a police report describing a police stop.

   Remove any information about the suspect's age. For example, convert "Young suspect found..." to "Suspect found...".

   Return only the modified text, without any additional explanations or comments. Do not change the text at all except as specified above. If no changes are required, just reply with the original text.

7. **Remove Name Information:**

   You are being given the contents of a police report describing a police stop.

   Remove any information about the name of the suspect, UNLESS it is directly relevant to the commission of the crime.

Return only the modified text, without any additional explanations or comments. Do not change the text at all except as specified above. If no changes are required, just reply with the original text.

8. **Remove Geographic Information:**

   You are being given the contents of a police report describing a police stop.

   Remove any information about geography and/or location, UNLESS it is directly relevant to the commission of the crime.

   Return only the modified text, without any additional explanations or comments. Do not change the text at all except as specified above. If no changes are required, just reply with the original text.

9. **Remove Appearance Information:**

   You are being given the contents of a police report describing a police stop.

   Remove any information about the attire, hairstyle, and/or facial hair of the suspect, UNLESS it is directly relevant to the commission of the crime.

   Return only the modified text, without any additional explanations or comments. Do not change the text at all except as specified above. If no changes are required, just reply with the original text.

10. **Remove Post-Stop Information (Final Check) with GPT-4o:**

    You are being given the contents of a police report describing a police stop.

    Edit the text so that it contains only the information the police would have known before making the stop. To do this, remove any discussion of the outcome of the police stop as well as any information the police would not have known before making the stop. For example, convert "Suspect appeared to be smoking marijuana, upon further investigation it was only tobacco" to "Suspect appeared to be smoking marijuana", because "further investigation" likely involved a stop. As another example, the sentence "Search of suspect revealed

marijuana" should simply be deleted, because searches can only occur after a stop. A stop occurs when a reasonable person would not feel free to leave or end the encounter. However, any information the police could have learned before the stop actually begins should be retained.

Output only the modified text, without any additional explanations or comments. Do not change the text at all except as specified above. If no changes are required, just output the original text.

B.4 Testing for Proxy Multicollinearity with Explainable Race Predictions from Report Text

We employ slightly older but more explainable Natural Language Processing (NLP) techniques to investigate the relationship between police reports and race, ethnicity, and gender. Our analysis addresses the concern that the content of the police reports might inappropriately proxy for race, causing attenutation of coefficient estimates for race in the regressions above. Our process can be summarized as follows:

1. **Text Pre-processing:** We pre-processed the text in the police reports as follows:
    - Conversion to lowercase
    - Removal of punctuation
    - Tokenization
    - Removal of stop words
2. **Feature Extraction:** We employed Term Frequency-Inverse Document Frequency (TF-IDF) vectorization to convert the pre-processed text into numerical features, with a maximum of 5000 features.
3. **Model Training:** For each demographic variable, we trained a separate logistic regression model with LASSO regularization. The data was split into 80% training and 20% testing sets.

4. **Model Evaluation:** We evaluated each model using standard classification metrics, including precision, recall, and F1-score for each category within the demographic variables.

5. **Feature Importance Analysis:** We extracted and ranked the most important features (words) for predicting each demographic variable. Additionally, we identified the top features associated with each category within each demographic variable.