Prose and Cons: Evaluating the Legality of Police Stops with Large Language Models

By David S. Abrams and Jonathan H. Choi[*]

In the near future, algorithms may assist law enforcement with real-time legal advice. We take a step in this direction by evaluating how well current AI can perform legal analysis of the decision to stop or frisk pedestrians, comparing multiple algorithmic and non-algorithmic approaches. We find that large language models (LLMs) can accurately assess reasonable suspicion under Fourth Amendment standards. Using 41,332 attorney-coded police stops from 2014 to 2024, we fine-tune an LLM to identify whether stops are illegal based on the content of police report narratives. The LLM identifies illegal stops with 88% accuracy and is well-calibrated in its confidence estimates. Importantly, it achieves 95% accuracy on the 75% of cases about which the model is most confident. Using topic modeling and LLM annotation, we also identify the primary justifications for stops and develop decision rules that police could implement to reduce illegal stops most effectively. We show that no set of rules is as effective as a fine-tuned LLM. These findings demonstrate LLMs' potential both to efficiently audit police practices and to provide real-time guidance.

I. Introduction

Law enforcement officers must routinely make split-second decisions with potentially fatal consequences. One very common such decision is the determination of whether to stop, frisk or search a suspect. While police shootings receive more attention, police stops are far more frequent, in some cities occurring more than 1,000 times a day (New York Civil Liberties Union, 2025; Hausman and Kronick, 2023). Officers making these stops must carefully navigate between controlling crime and respecting the constitutional rights of suspects.

In a stop, the officer often has little to go on beyond brief observation of the individual, the environment, and perhaps radio information about criminal suspects. Recent technological advances have the potential to transform this common police practice. Body-worn cameras already digitize some of the data officers observe,

[*] Abrams: University of Pennsylvania, dabrams@upenn.edu. Choi: Washington University School of Law, thechoipolloi@gmail.com. Acknowledgements: David Rudovsky, Julian Nyarko, Mike Livermore, Anthony Niblett, Ben Chen, Participants at: OWCAL, SMU Dedman Faculty Workshop, USC CLASS Workshop, U Toronto Econ Summer Workshop, NBER L&E and Crime Workshops, Fordham AI Workshop, Wharton Urban & Real Estate Workshop, CrimFest, ETH Zurich Law and Economics Workshop. This paper's hypotheses and methods are pre-registered with the Open Science Foundation at: `https://osf.io/jntr6/?view_only=ddaf04facf9941cd97e1155214163647`.

but there is currently little analysis of this data. In this paper, we propose and test new AI algorithms that could process data and make recommendations on the legality of police stops in real time. While this application may not come into use immediately due to technological, regulatory or inertial hurdles, we also demonstrate an immediate application of our algorithms—for ex-post legal review of police stops and frisks. These types of reviews are common in many jurisdictions, especially in cities that have been subject to litigation about or investigation of their policing practices. These reviews currently require large amounts of work by trained individuals (typically lawyers) at high expense.

This paper presents analysis of a novel dataset with over 40,000 observations of police stops in a major U.S. city, each of which includes a narrative of the rationale for the stop. In these "*Terry* stops," officers may make stops only if there is reasonable and articulable suspicion (RAS) that the subject is involved in a crime, and may conduct frisks only if there is RAS that the person is armed and dangerous. Importantly, a team of expert attorneys coded each stop and frisk in our dataset for whether it met the legal standard for RAS. We use these data to train various machine learning (ML) models to identify whether stops and frisks have RAS.

We find that algorithms can perform this core legal analysis at high levels of accuracy compared to attorneys, and far exceeding that of police officers and law students. In our data, the top-performing algorithm (fine-tuned Llama 3) has 87.9% accuracy and 0.901 AUC-ROC. Importantly, the models quantify prediction confidence for each stop, which we use to rank them (calibrating confidence on the train set and evaluating it out-of-sample on the test set). When keeping only the top half of the observations about which the model is most confident, accuracy jumps to 98.2%. We also show that we can keep 74.6% of the data and still maintain 95% accuracy.

Llama 3's performance compares favorably to various human benchmarks. Its accuracy significantly exceeds the officers themselves, whose narratives justify reasonable suspicion in 80.3% of stops. It also exceeds the performance of law student RAs trained to assess RAS, who score 77.0% on accuracy and 0.759 on AUC-ROC. Finally, Llama 3's accuracy roughly matched the rate of agreement between expert attorneys hired by plaintiffs and by the city itself (87.0% agreement). These comparison points, discussed in further detail in Section II.D, suggest that it might be difficult to improve Llama 3's performance further due to subjective disagreement on RAS between experts.

In addition to Llama 3, we use OpenAI's o3 model to evaluate RAS, with a detailed prompt but without fine-tuning.[1] We also compare Llama 3's performance with several older ML techniques, including random forest, logistic regression, and linear probability models (LPMs).[2] While Llama 3 has the best performance, ran-

[1] For replicability, we use the o3-2025-04-16 model checkpointed with a knowledge cutoff of May 31, 2024.

[2] We also tested support vector machines, but they performed significantly more poorly than the other

dom forests and logistic regression also perform well and are well-calibrated. The advantage of Llama 3 is most noticeable in the share of stops it can correctly classify with 95% accuracy. There it exceeds all other methods tried by over 20 percentage points. We also compare performance with different inputs, finding that Llama 3's performance improves only very slightly when adding additional covariates beyond the text narrative.

While the algorithmic approach shows excellent performance, there might be technological and logistical barriers to full implementation. As an alternative, one could create simple decision rules based on textual analysis that could help decrease stops lacking RAS without omitting too many legal stops. We formulate such guidelines using topic modeling with OpenAI's o3 model to identify topics in each police report and then use o3 to identify the main reason for each stop. We are then able to estimate how much omitting stops with each main reason would reduce overall stops and those lacking RAS.

We find that the distribution of stop reasons has a long tail, with many stops justified by a rare main reason. This means a simple decision-rule approach, with 10 or fewer rules of thumb, would only eliminate a small share of stops lacking RAS. In addition, decision rules are much less efficient than the algorithmic approach, in that they require far more stops to be eliminated to achieve a similar reduction in the false positive rate compared to fine-tuned Llama 3.

The rest of the paper proceeds as follows: Section II provides a brief legal background about stop and frisk policing along with summary statistics for our jurisdiction. In Section III we present the methods employed to predict RAS, compute confidence in rankings, and perform topic modeling. Section IV includes our main results in each of those areas, and Section V concludes.

II. Background and Data

A. *Literature Review*

Inspired by recent advances in ML, a huge literature compares the performance of humans and LLMs at legal tasks. Kleinberg et al. (2018)'s seminal paper showed that simple predictive models could outperform judges in bail decisions, and many scholars have built on this early work by demonstrating that LLMs perform well at legal analysis, whether as substitutes (Martin et al., 2024; Savelka and Ashley, 2023; Nay et al., 2024) for or complements (Shao et al., 2025; Choi, Monahan and Schwarcz, 2024; Choi and Schwarcz, 2025) to human analysis. More broadly, researchers have compared the performance of LLM and human annotation in domains such as content analysis (Bojić et al., 2025) and medicine (Lin et al., 2023; Goh et al., 2024; Nori et al., 2025). The literature suggests that LLMs can competently produce sophisticated analysis in a variety of domains.

At the same time, a substantial theoretical and empirical literature has explored approaches to detecting racial disparities in police stops. (Knowles, Persico and

classic ML models and were therefore excluded from the final results.

Todd, 2001) launched this literature with a simple optimizing model of police and offender behavior and proposed an outcome test where a comparison of “hit rates” - the fraction of frisks that yield contraband - by racial group could reveal disparities. This paper inspired various papers focused on largely on refining models of racial disparity in police stops, including important contributions by (Anwar and Fang, 2006), (Durlauf, 2006), (Dharmapala and Ross, 2004), and (Antonovics and Knight, 2009). More recent work by (Gelbach, 2021) and (Feigenberg and Miller, 2022) among others has continued to broaden both models of racial disparities in police stops and their empirical application.

While the work discussed above focuses on racial disparities in police stops, fewer papers have attempted to model the policing optimization problem more broadly. (Chalfin and McCrary, 2018) model the optimal *number* of police but not the activities they engage in. (**?**) propose a “police production function” where contraband discovery is the primary objective. (Campbell, 2025) considers the optimization problem of a police captain choosing how to allocate personnel between making stops and other potentially productive activities, such as patrol. (Rivera and Ba, 2025) consider how police oversight impacts misconduct.

Our work takes on new significance given ongoing attempts to automate police monitoring. Axon’s Draft One, for example, is an LLM-driven tool that generates police reports based on audio from body-worn cameras (Ferguson, 2024). Tools like these could generate raw data that could subsequently be fed into a model like the ones that we train in this paper.

Due to the paucity of available data on this topic and the novelty of the required LLM tools, few attempts to computationally evaluate RAS exist. To our knowledge, the only other attempt was Oliver et al. (2024), which annotates factors relating to RAS from court opinions involving vehicular stops and then uses these factors to predict judicial rulings on RAS. However, their study differed from ours in important respects. Because they use case law rather than stop narratives, their sample is highly selected and likely unrepresentative of real-world policing practice. Moreover, they predict judicial decisions *using text written by the judges themselves*—this potentially contaminates their prediction models, since judges likely frame case facts selectively to support their judgments. Our dataset does not suffer from this problem, since officers presumably believe there is RAS in every case and the judgment we are predicting is a separate evaluation by legal experts.

B. Conceptual Framework

Before proceeding to the empirical heart of this paper, it is worth a brief discussion of police incentives and the legal background. At a high level, a police department is tasked with (among other things) reducing crime while respecting citizens’ constitutional rights. This is accomplished in myriad ways within and across police departments. Our focus is on Stop and Frisk policing, an extremely common tactic that became popular in the early 2000’s.Many police leaders continue to believe it has a substantial deterrent effect on crime, although no study has yet validated this belief.

Although we do not formally model police incentives (as does (Campbell, 2025)), one way to conceive of it is as a series of principal-agent problems: that between citizens and the police chief/city mayor and that between the police chief/ leadership and patrol officers. In each of these relationships, some of the most visible performance measures are crime rates (particularly violent), "clearances" - the share of crimes where an arrest is made, expenditures, and instances of violations of constitutional rights, particularly those that receive public attention. In short, this is a constrained optimization problem where safety and liberty are the primary objects of interest.

A general model of these agency problems is beyond the scope of this paper. But it is worth describing police officer incentives to make stops and frisks, to record them accurately, and to ensure they are legally justified, along with how each of these may vary by type of suspected crime. [3]

Police officers in our jurisdiction, as with many others, routinely stop and question individuals when they have reasonable suspicion that they may be involved in the commission of a crime or are planning criminal conduct. These are known as *Terry* stops, after the Supreme Court ruling in *Terry v. Ohio* which established the reasonable suspicion standard. *Terry v. Ohio* also established that, for their protection, police 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.

Incentives of officers to make stops have varied over time as public attitudes to the practice have changed. There have never been clear, numerical targets for officers to stop or frisk, and in fact there is immense variation across officers in how much they do so.

Data from police stops must be entered into an electronic database. Original entry typically occurs through an application on a mobile data computer (MDC), a laptop installed in a police vehicle. These are subsequently checked for completeness and compliance with Fourth Amendment standards by the supervising sergeant. There has typically been no consequence to officers who fail to state reasonable suspicion for a stop or frisk. Very recently, there has been an effort to incentivize only stops and frisks with RAS. Officers who report stops or frisks without RAS and Sergeants who fail to correct the officers are subject to retraining, oral and written warnings and for repeat offenders to discipline in the form of lost vacation days. However, the vast majority of repercussions have been in the form of verbal warnings and either none or an extremely small number of vacation days have been lost to date.

Of course, each interaction between police officers and citizens has the potential

[3] We omit the impact here of intrinsic incentives, which is explored by (Chalfin and Gonçalves, 2023).

to go poorly. This is one reason why the reasonable suspicion standard exists. It also means that potential gains from stops and frisks should be balanced by a social planner against these potential harms. Potential gains—in deterred or detected crimes—will be larger for more serious crimes. But in this paper we do not distinguish between potential crime severity for three reasons. First, the initial rationale for the stop is often different from a crime the individual is arrested for. Second, the reasonable suspicion standard is independent from crime type, and thus from a legal perspective must be followed uniformly, even if this may not lead to the optimal solution to the constrained optimization problem (although it may be optimal more broadly).

C. Data Description

Our study analyzes a novel dataset of pedestrian *Terry* stops conducted by a major metropolitan police department between 2014 and 2024. Our dataset contains 41,332 observations randomly selected from over a million stops during this period. For each stop, the officer records details about the circumstances of the stop, the individual stopped, the location and time of the stop, and officer identifiers. Outcomes of the stop include whether the individual was frisked, searched or arrested, and whether contraband was discovered (and if so, what kind). Crucially, the officer provides a narrative intended to establish reasonable and articulable suspicion (RAS) for the stop. These narratives constitute the bulk of the data we analyze.

As part of a monitoring agreement stemming from litigation, a small group of plaintiff's attorneys with expertise in these matters manually read each police narrative in the random subset of stops and code each stop as either having or lacking reasonable suspicion (each lawyer makes only one assessment). For cases where a frisk occurred, they separately evaluate whether the frisk had reasonable suspicion. After dropping observations which could not be clearly coded, we have 41,332 observations where stops were coded as having or lacking reasonable suspicion and 7,795 observations where frisks were coded as having or lacking reasonable suspicion. [4] While the attorneys coding the stops had access to the full text of the police narrative without redaction, the RAS determination was made based solely on ex ante information, meaning that a stop lacking reasonable suspicion could not become justified ex post, merely because the suspect turned out to have engaged in a crime.

Table 1 presents summary statistics for the key variables in our dataset. The data reveal several important patterns relevant to our analysis. First, 80.3% of stops in the lawyer-coded sample were assessed as having reasonable suspicion,

[4] In our current analysis, we only analyze stops coded as having or lacking reasonable suspicion. Thus we drop stops that were incorrectly classified as such (such as arrests). We include only frisks coded as having or lacking RAS where the stop preceding the frisk had RAS. Thus we exclude frisks that occur despite the preceding stop lacking reasonable suspicion - this is considered "fruit of the poisonous tree". We also exclude miscoded frisks, e.g. those that were actually searches incident to arrest.

suggesting that a substantial proportion of stops were legally unfounded. Frisks occurred in 19.7% of stops, with 73.6% of these frisks coded as having reasonable suspicion when the initial stop was lawful. As in many criminal justice datasets from large cities, the population of individuals stopped by police is largely male (85.9%) and non-White (70.3% Black and 9.4% Hispanic). The majority (57%) of stops occur between 4PM and midnight and three-quarters of stops are made when an officer has a partner. Contraband is discovered in 7.7% of stops - this may include weapons, drugs, or stolen property.

Table 1—Summary Statistics for Police Stops, 2014-2024

Variable	**Mean**	**Standard Deviation**
Reasonable Suspicion for Stop	0.803	0.398
Individual Frisked	0.197	0.398
Reasonable Suspicion for Frisk	0.736	0.441
Male	0.859	0.348
Black	0.703	0.457
White	0.280	0.449
Hispanic	0.094	0.292
Age	33.2	13.2
Height	5.5	0.4
Weight	170.5	33.7
Year	2019.14	2.346
Evening	0.57	0.50
Daytime	0.31	0.46
Night	0.12	0.33
Officer with Partner	0.748	0.434
Contraband Discovered in Stop	0.077	0.266

Notes: This table presents summary statistics for key variables in the dataset of 41,332 police stops from 2014 to 2024. The table reports the mean and standard deviation for the following variables: *Reasonable Suspicion for Stop* (indicator variable equal to 1 if reasonable suspicion was present, 0 otherwise, calculated only for cases coded as "Yes" or "No"), *Individual Frisked* (indicator variable equal to 1 if the suspect was frisked, 0 otherwise), *Reasonable Suspicion for Frisk* (indicator variable equal to 1 if reasonable suspicion was present for the frisk, 0 otherwise, calculated only for cases coded as "Yes" or "No"), *Gender* (*Male* indicator variable equal to 1 if the suspect is male, 0 otherwise), *Race* (*Black* and *White* indicator variables), *Ethnicity* (*Hispanic* indicator variable equal to 1 if the suspect is Hispanic, 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* (4 PM to midnight), *Daytime* (8 AM to 4 PM), and *Night* (midnight to 8 AM) indicator variables), *Officer with Partner* (indicator variable equal to 1 if the officer had a partner during the stop, 0 otherwise), and *Contraband Discovered* (indicator variable equal to 1 if contraband was found, 0 otherwise).

D. Ground Truth and Intercoder Reliability

Before proceeding, it is worth clarifying a key issue in this paper (and life): what is truth? How does one know when a stop truly lacked RAS based on the written narrative? When training and evaluating the performance of our ML models (including Llama 3), we treat the assessment of the attorney coders as ground truth. Thus, one may interpret the performance of the algorithms (and humans) that we report in this paper as measured by how close they come to these human experts.

Of course, legal evaluations are subjective, and even expert attorneys could disagree in specific cases on whether RAS is present or absent. Many scholars have observed the problem that human evaluations treated as ground truth might in fact be subjective (Plank, 2022; Chen, Mermel and Liu, 2021). One standard method to relax the assumption that human codings are ground truth is to have multiple humans evaluate the same case and then calculate intercoder agreement rates (Movva, Koh and Pierson, 2024). Then, we could simply compare intercoder agreement between the humans against intercoder agreement between the humans and the LLM.

In our study, we have three different human codings to compare against the baseline generated by plaintiff's attorneys. First, we have a subset of randomly selected cases from 2024 that were coded both by the plaintiff's attorneys and the attorney for the city where the police stops occurred. Out of 1311 randomly selected stops, when coding for the presence of RAS, the attorneys agreed on 1140 and disagreed on 171, giving an intercoder agreement rate of 87.0%. This is lower than the intercoder agreement rate between our best model (fine-tuned Llama 3) and the plaintiff's attorneys of 87.9% (which we present below as the accuracy of the fine-tuned Llama 3 model). This might suggest that our best model is approaching the theoretical limit for performance, above which determinations become subjective. On the other hand, the plaintiff's and defendant's attorneys have opposing incentives in litigation, which could make their agreement rate artificially low.

A second basis for human comparison is the agreement rate between the plantiff's attorneys and our human RA. This agreement rate was even lower, 77.0% for stops and 52.0% for frisks, far lower than the agreement rate for fine-tuned Llama 3 (76.9% for frisks). Again this might suggest task subjectivity, although it might also suggest that the task is simply too difficult for a human RA to tackle.

A third basis for comparison is the rate of agreement between the plaintiff's lawyers and the police officers themselves. On the sample visible to us (which excludes both true and false negatives, where the police did not believe RAS was present), the plaintiff's attorneys found that 80.3% of stops had reasonable suspicion, and 73.6% of frisks. This again implies a rate of agreement lower than the rate of agreement between fine-tuned Llama 3 and the plaintiff's attorneys. However, this comparison is especially fraught, since the point of the litigation and settlement was a belief that police officers were violating the law and making stops when RAS was absent—our priors do not suggest that the disagreement between officers and the lawyers was simply due to the subjectivity of RAS.

In the absence of recodings within the team of plaintiff's attorneys, the above offer only suggestive evidence of the extent of the subjectivity issue. Meanwhile, the remainder of the paper still treats the plaintiff's attorneys codings as ground truth, but our performance statistics could be treated as lower bounds in light of the subjectivity of RAS determinations.

III. Empirical Framework

A. Predicting RAS with Different Models

We evaluate three distinct approaches: fine-tuned large language models (Llama 3), classic machine learning algorithms, and a top-performing off-the-shelf large language model (o3).

Our primary approach involves fine-tuning separate Llama 3 models to predict the human evaluations of reasonable suspicion for stops and frisks independently. Specifically, we generate predictions of whether there was RAS for the stop and, separately, whether there was RAS for the frisk using Llama 3, a decoder-only transformer model developed by Meta.[5] 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 and o-series models. This accessibility allows us to fine-tune the model weights to optimize performance for our specific prediction tasks.

We tested each model with three different sets of inputs: only the narrative police report, the police report plus information that can easily be accessed pre-stop, and all the information available (including post-stop data). We focus on the narrative-only prompt in our analyses for reasons of legal alignment, interpretability, and fairness. In addition, we found that including additional information did not to improve the performance of the models, which is consistent with the human experts' appropriately making their determinations on the contents of the police narratives alone. See Appendix B for details on the development and testing of these variations of variables included in the models.

To adapt Llama 3 to our specific prediction tasks, we employ Low-Rank Adaptation (LoRA), a parameter-efficient fine-tuning method that modifies only a small subset of the model's weights.[6] Following standard practice for binary classifica-

[5] A decoder-only transformer is a neural network architecture that processes text sequentially from left to right, predicting each next word based on all previous words. Unlike encoder-decoder models used for translation tasks, decoder-only models are optimized for text generation and completion tasks. They use self-attention mechanisms to capture long-range dependencies in text, making them particularly effective for understanding complex narratives like police reports.

[6] We specifically employ Low-Rank Adaptation (LoRA), an implementation of Parameter-Efficient Fine-Tuning (PEFT), which fine-tunes a subset of the model's weights in order to reduce memory requirements and training time (Ding et al., 2023; Hu et al., 2022). Appendix Section A.A2 describes LoRA fine-tuning in greater detail.

tion tasks in natural language processing (Devlin et al., 2019; Brown et al., 2020), we minimize the cross-entropy loss:

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

where y_i is the outcome of interest, a dummy variable indicating whether there was RAS for the stop or frisk and $\hat{y_i}$ is the model's continuous prediction, bounded to be between 0 and 1. We use cross-entropy loss rather than squared error loss because it is the standard loss function for binary classification tasks, providing better gradient properties when the model outputs are interpreted as probabilities.[7] After fine-tuning, we evaluate its performance on a set of test examples that was kept segregated during the training process.[8]

As alternative approaches, we also train classic machine learning models as a baseline for comparison. Specifically, we implement three standard algorithms: random forest classifiers with 100 estimators, logistic regression with L2 regularization (also known as ridge regression), and simple ordinary least squares regression with no regularization (also known as a linear probability model (LPM)). For these models, we convert the police narratives into bag-of-words features using count vectorization with a maximum of 5,000 features and English stop words removed. Each model is trained on the same training split (55% of the data) used for the Llama fine-tuning, with hyperparameters tuned on the validation set (15% of the data). Appendix C provides additional information about and formal mathematical definitions of each of the baseline ML models.

Additionally, we evaluate OpenAI's state-of-the-art o3 model in an off-the-shelf configuration without fine-tuning. We prompt engineered both a short and a detailed system prompt for o3 explaining the legal framework for reasonable suspicion under *Terry v. Ohio* as applied in our jurisdiction. The text for these prompts was previously developed to train law students to perform the RAS evaluation. The model is asked to provide predictions on a 0-100 scale. We convert this score to a dummy for RAS, using a threshold of 50 for comparison with the binary classifications. We find that the longer, more detailed prompts outperform the shorter prompts on the validation set, and therefore we use the longer prompts as our main specification. Appendix Sections A.A4 and C.C3 contain additional details on the prompts used and full performance statistics.

[7] Cross-entropy loss is preferred over squared error loss for binary classification because it penalizes confident wrong predictions more heavily and provides stronger gradients when predictions are far from the true labels, leading to faster and more stable convergence during training.

[8] We use a 55%/15%/30% train/validation/test split, which allows us to tune hyperparameters on the validation set while maintaining a completely held-out test set for final evaluation.

B. Ranking Confidence in Predictions

In addition to simply optimizing prediction, it may be possible to identify some subset of cases where the reasonable suspicion models can be used with high confidence to make live recommendations to officers.

To identify cases where the model predictions are most reliable, we use a confidence scoring mechanism based on each algorithm's predicted probability of reasonable suspicion. Specifically, we define the confidence score for observation i as:

$$c_i = |P(y_i = 1|x_i) - 0.5| \tag{2}$$

where $P(y_i = 1|x_i)$ is the model's predicted probability that observation i has reasonable suspicion.[9] This metric captures the model's decisiveness: predictions closer to the decision boundary of 0.5 indicate greater uncertainty, while predictions near 0 or 1 indicate higher confidence. By ranking observations according to c_i and selecting those with the highest confidence scores, we can identify a subset of cases where the model's predictions are most reliable. The confidence scores are calibrated on the train set and validated out-of-sample on the test set.

C. Topic Modeling

In order to understand what words or phrases from the narratives were most predictive of RAS, we perform topic modeling. Rather than analyzing individual words,[10] we developed a method to extract semantically meaningful topics from police narratives. We employed OpenAI's o3 model to identify generalizable topics within each police report that reflect the rationale or justification for the stop.

For example, a report stating "Police observed male walking on the highway at listed location. Police had prior knowledge that male had an open warrant for assault" would yield two distinct topics: "Walking on Highway" and "Known Warrant." This approach allows us to identify semantically meaningful components of police narratives rather than arbitrary text segments, providing clearer insights into which types of observations drive legal assessments.

Complete details on our topic modeling method are described in Appendix Section E.

[9] Certain algorithms, specifically LLMs (like Llama 3 and o3) and logistic regression, natively generate log probabilities of outcomes rather than linear probabilities. In these cases, we exponentiate the log probabilities to generate linearized probability values. Thus, if $\log P(y_i = 1|x_i)$ is the log probability, then $P(y_i = 1|x_i) = \exp(\log P(y_i = 1|x_i))$.

[10] We also tested alternative interpretability methods including LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations), but found they produced unintuitive results with high importance scores assigned to seemingly random words. LIME, SHAP, and other older methods evaluate the effects of single words in text, ignoring important context in the process. Our topic-based approach preserves semantic context and produces more legally interpretable results.

D. Evaluating Importance of Topics

While multiple topics may appear in a single police report, we posit that officers typically have a primary justification—a "main reason"—that motivates their decision to conduct a stop or frisk. To identify this main reason, we again employed OpenAI's o3 model, providing it with the full police report and the list of canonical topics present in that report. The model was instructed to identify which single topic best represents the primary reason the stop or frisk was initiated, considering factors such as the chronological order of events, what triggered the initial police contact, and the most serious or significant reason if multiple justifications exist. Appendix E and E provide complete details on how main reasons were identified.

This main reason identification allows us to simulate targeted policy interventions. Specifically, we can calculate what would happen to the false positive rate—the proportion of stops or frisks that lack reasonable suspicion—if police were instructed to discontinue stops or frisks when certain topics constituted the main reason. By systematically evaluating the impact of removing each topic as a main reason for stops or frisks, we can rank potential policy interventions by how much they reduce unconstitutional stops or frisks while minimizing the impact on legitimate police work. Appendix E.E1 formally describes the equations used to calculate false positive rates and determine decision rules.

IV. Empirical Results

A. Model Comparison

We use fine-tuned Llama 3, classic machine learning algorithms (including logistic regression, LPM, and random forest), and OpenAI's off-the-shelf o3 model to predict the presence of reasonable suspicion. We compare these to two types of human evaluations, those conducted by the police officer who wrote the report, as well as by a law student given similar instructions to the o3 model (the longer version). Table 2 presents the performance metrics for each approach on the test set for stop predictions. Like the expert lawyers lawyers coding manually, the models were given access to the full text of the police reports. By comparison, the police themselves were 80.3% accurate in making stops over our sample and 73.6% accurate in making frisks, in both cases excluding unobserved false negatives (since we do not observe the suspects the police chose not to stop).[11]

To provide a benchmark against which to compare our model results, we also instructed a law student to assess whether there was RAS for each stop and frisk, on a 10-point scale. Similar to the approach with the algorithms, we converted those scores to a binary variable for RAS if the score was 5 or above. These

[11] The performance statistics reported here are for the narrative-only prompt configuration. As detailed in Appendix B, we also tested configurations that included pre-stop metadata and all available metadata, but found that the simpler narrative-only approach provided the most consistent performance.

results are also included in Table 2 to show the performance of a person with some legal background but no direct experience with police stops and to provide a basis for comparison that includes both false positives and false negatives, like the ML models.

The "Acc @ 50%" column shows the model's accuracy when evaluating only the 50% of cases for which the model is most confident, based on the predicted probability of reasonable suspicion (determined based on the model's training without access to the outcome variable in the test set). The "Max Subset w/ 95% Acc" column indicates the maximum percentage of cases that can be automatically classified while maintaining at least 95% accuracy, again after ranking the observations based on model confidence. For example, the fine-tuned Llama 3 model achieves 98.2% accuracy on its most confident 50% of predictions and can maintain 95% accuracy while classifying 74.6% of all cases. These metrics reflect both the performance of each model and each model's calibration in estimating the confidence of its predictions; ceteris paribus, better-calibrated models will perform better on these metrics. For all metrics in the table, a higher value is better.

Table 2—Comparison of Model Performance for Stop Reasonable Suspicion Prediction

Model	**AUC-ROC**	**F1**	**Acc**	**Acc @ 50%**	**Max Subset w/ 95% Acc**
Fine-tuned Llama 3	0.901	0.927	0.879	0.982	74.6%
Logistic Regression	0.822	0.901	0.834	0.951	50.7%
Random Forest	0.814	0.907	0.836	0.943	43.0%
LPM	0.778	0.886	0.806	0.942	39.8%
o3	0.776	0.887	0.815	0.925	17.0%
Human Coding	0.759	0.853	0.770	0.935	41.9%

Notes: This table compares the performance of different models in predicting reasonable suspicion for police stops. Models include fine-tuned Llama 3, classic machine learning algorithms (Logistic Regression, Linear Probability Model, Random Forest), OpenAI's o3 model, and human coding by a law student. Performance metrics are: AUC-ROC (area under the Receiver Operating Characteristic curve), F1 score, overall accuracy (Acc), accuracy on the most confident 50% of predictions (Acc @ 50%), and the maximum percentage of cases that can be classified while maintaining at least 95% accuracy (Max Subset w/ 95% Acc). Higher values indicate better performance for all metrics.

The fine-tuned Llama 3 model generally achieved the best performance, with AUC-ROC of 0.901, F1 score of 0.927, and accuracy of 87.9%. Notably, this represents a substantial improvement over human coding performance, which achieved an accuracy of 77.0%, F1 score of 0.853, and AUC-ROC of 0.759. While the human RA achieved solid performance when highly confident (93.5% accuracy at 50%), he could only maintain 95% accuracy on 41.9% of cases, compared to 74.6% for the fine-tuned Llama 3 model.

OpenAI's o3 model, despite being a more advanced architecture than Llama 3, achieved lower performance (AUC-ROC of 0.776, F1 score of 0.887, and 81.5%

accuracy) when used off-the-shelf with prompting alone (rather than fine-tuning, as we did for Llama 3). The classic ML models—the random forest, logistic regression, and LPM—showed varying performance, with random forest and logistic regression performing reasonably well.

The superior performance of the fine-tuned Llama 3 model can be attributed to several factors. First, fine-tuning allows the model to adapt specifically to the legal nuances and writing style of our police reports (unlike the human coding and o3). Second, the transformer architecture can capture complex contextual relationships in the narratives that bag-of-words approaches miss (unlike the classic ML models). Third, the model benefits from pre-training on a large corpus of text, providing a strong foundation for understanding language patterns (unlike the classic ML models).

Similar patterns emerged for frisk predictions, as shown in Table 3. The fine-tuned Llama 3 model again outperformed all alternatives, though the performance gap was slightly smaller for this more challenging prediction task.

In practice, empiricists generally prefer AUC-ROC as a measure of performance because it is not sensitive to the cutoff chosen for binary classification, unlike the other metrics. (The metrics that rely on binary classification tend to disproportionately penalize poorly calibrated models.) To illustrate the Receiver Operating Characteristic (ROC) analysis, we plot the true positive rate (sensitivity) against the false positive rate (1 - specificity) at various classification thresholds for each model. The fine-tuned Llama 3 model achieves the highest Area Under the Curve (AUC) of 0.901, indicating strong predictive power—for any randomly chosen pair of stops, one with reasonable suspicion and one without, the model will correctly rank them 90.1% of the time. This substantially outperforms the other approaches, with Logistic Regression achieving an AUC of 0.822, Random Forest 0.814, and Human Coding 0.759.

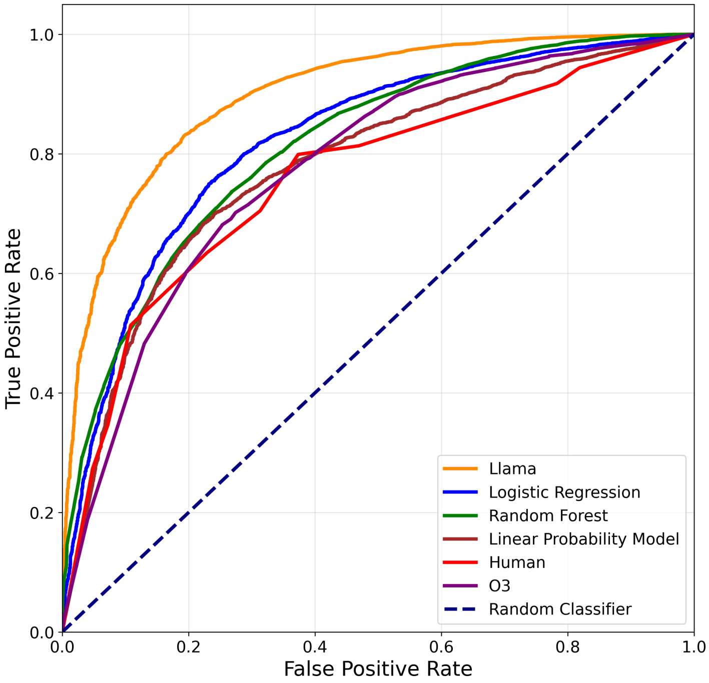

Figure 1. Receiver Operating Characteristic (ROC) Curves for Multiple Models Predicting Reasonable Suspicion in Police Stops

Notes: This figure displays the Receiver Operating Characteristic (ROC) curves comparing multiple models' predictions of reasonable suspicion in police stops. The curves plot the true positive rate (sensitivity) against the false positive rate (1 - specificity) at various classification thresholds using the test dataset. The diagonal dashed line represents random chance (AUC = 0.5).

In practical terms, the ROC curves reveal significant performance differences among the models. For the fine-tuned Llama 3 model, choosing a threshold that limits the false-positive rate (where the model falsely identifies a stop as having reasonable suspicion when it does not) to 5.1% still yields a true-positive rate of 60%, and relaxing the false-positive rate to 10% pushes the true positive rate to 72.4%. Put differently, if the Llama 3 model were calibrated to designate only one stop in twenty as lawful when attorneys would disagree, it could nevertheless automatically identify well over half of the truly lawful cases. The other models show lower true positive rates at these same false positive thresholds, confirming that fine-tuned Llama 3 had the best overall performance.

The frisk prediction results follow a similar pattern. The fine-tuned Llama 3 model achieved the best performance on AUC-ROC, Accuracy at 50%, and Maximum Subsample at 95% Accuracy (although the Random Forest has a higher F1 score and Accuracy), outperforming classic ML approaches, the off-the-shelf o3 model, and the human RA. This task proved more challenging than stop predictions across all models. The o3 model struggled particularly with frisk predictions, achieving only 64.1% accuracy even with detailed legal prompting, suggesting that domain-specific fine-tuning is especially valuable for this more nuanced legal determination.

Table 3—Comparison of Model Performance for Frisk Reasonable Suspicion Prediction

Model	**AUC-ROC**	**F1**	**Acc**	**Acc @ 50%**	**Max Subset w/ 95% Acc**
Fine-tuned Llama 3	0.793	0.849	0.769	0.913	29.6%
Random Forest	0.746	0.861	0.769	0.874	16.1%
Logistic Regression	0.746	0.833	0.747	0.875	20.7%
LPM	0.546	0.637	0.547	0.799	0.0%
o3	0.699	0.725	0.641	0.851	21.8%
Human Coding	0.567	0.636	0.520	0.680	1.9%

Notes: This table compares the performance of different models in predicting reasonable suspicion for police frisks. Models include fine-tuned Llama 3, classic machine learning algorithms (Random Forest, Logistic Regression, LPM), and OpenAI's o3 model. Performance metrics are the same as in Table 2: AUC-ROC, F1 score, overall accuracy, accuracy on the most confident 50% of predictions, and maximum percentage of cases classifiable at 95% accuracy. Higher values indicate better performance for all metrics.

Figure 2 illustrates the ROC curves for frisk predictions. The fine-tuned Llama 3 model achieves the highest AUC of 0.793, followed by Random Forest (0.746) and Logistic Regression (0.746). While the overall performance is lower than for stops, the relative ranking of models remains consistent, with fine-tuned language models outperforming traditional approaches.

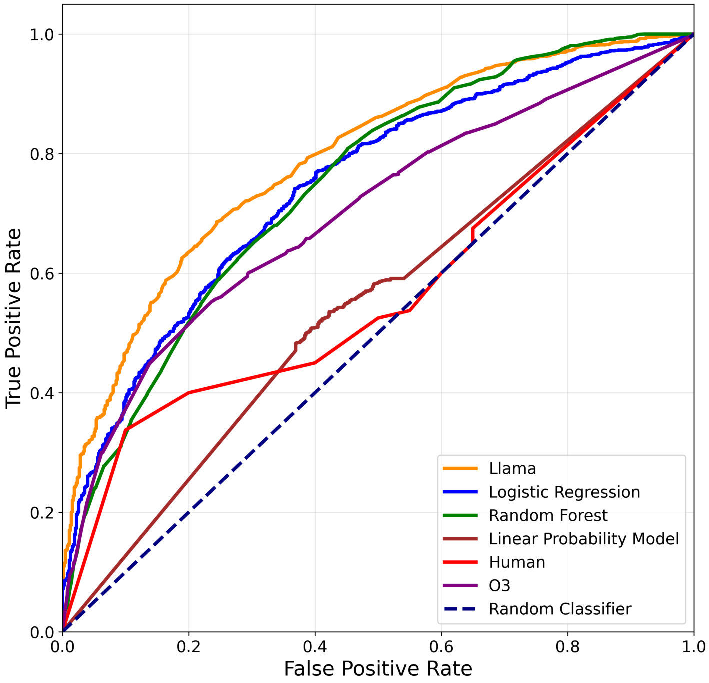

FIGURE 2. RECEIVER OPERATING CHARACTERISTIC (ROC) CURVES FOR MULTIPLE MODELS PREDICTING REASONABLE SUSPICION IN POLICE FRISKS

Notes: This figure displays the Receiver Operating Characteristic (ROC) curves comparing multiple models' predictions of reasonable suspicion in police frisks. The curves plot the true positive rate (sensitivity) against the false positive rate (1 - specificity) at various classification thresholds using the test dataset. Models include fine-tuned Llama 3 (AUC = 0.793), Random Forest (AUC = 0.746), Logistic Regression (AUC = 0.746), Linear Probability Model (AUC = 0.546), and OpenAI's o3 model (AUC = 0.699). The diagonal dashed line represents random chance (AUC = 0.5).

Why is model performance substantially worse on frisks than on stops? First, identifying reasonable suspicion for frisks may be inherently more difficult than for stops—the police themselves perform at 73.6% accuracy for frisks versus 80.3% for stops (excluding unobserved false negatives). Second, our train set for frisks is much smaller than for frisks, which would tend to decrease model performance even on a task of equivalent difficulty.

B. Ensemble Methods

To further improve prediction accuracy, we implemented several ensemble methods that combine predictions from multiple models. Ensemble techniques have been shown to reduce overfitting and improve generalization by combining the strengths of different models (Dietterich, 2000).

We tested several methods of ensembling predictions from different models: (1)

simple averaging of model probabilities, (2) weighted averaging based on individual model performance, (3) median voting, (4) stacking with logistic regression as a meta-learner, and (5) a novel neural stacking technique using a DistilBERT-based architecture that combines text embeddings with base model predictions. The neural stacking approach allows the meta-learner to access both the raw police narrative and the models' predictions, potentially allowing the predictions to be combined in a more nuanced way. Technical details of these ensemble methods are provided in Appendix D.

Despite their theoretical advantages, our ensemble methods provided only marginal improvements over the fine-tuned Llama 3 model alone. As detailed in Tables D1 and D2 in the Appendix, the best-performing ensemble method (stacking with logistic regression) achieved AUC-ROC improvements of only 0.001 for stop predictions and 0.021 for frisk predictions. This suggests that the fine-tuned language model already captures most of the predictive signal in the police narratives, leaving little room for improvement through model combination.

C. Ranking Confidence in Predictions

Ultimately, the decision of whether to employ an algorithm will depend on what error rates are deemed acceptable by those charged with making the assessments. Even if it is not possible to classify all narratives with a high enough level of confidence, it may be possible to use LLMs to audit police practice if we can identify 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. Real-time algorithmic feedback might not be available in all cases, for example in a fast-moving or chaotic situation. But even if only applicable some of the time, it could help eliminate a large proportion of illegal stops.

One crucial barrier to practical implementation is whether we can identify some subset of cases for which the models are sufficiently accurate that police departments would feel comfortable relying on them. As discussed above in Section III.B, one simple way to accomplish this is to test the ex ante confidence assigned by each of the models to its prediction that reasonable suspicion is or is not present. Because each model generates a probability in the range $[0, 1]$, we can simply treat the distance in each model's predicted probability from 0.5 as its level of confidence, with 0 or 1 reflecting the highest level of confidence and 0.5 representing the lowest. Having produced a ranking of cases based on model confidence, we can generate subsets of cases comprising the top $n_p\%$ of cases for which the model is most confident.

We then plot accuracy for each subset of cases as follows:

$$\text{Accuracy}(p) = \frac{1}{n_p} \sum_{i \in S_p} \mathbb{I}(\hat{y}_i = y_i)$$

Where p is the percentage of cases included, S_p represents the subset of cases with highest predicted confidence, $n_p = |S_p|$ is the number of cases in that subset, $\hat{y}_i$ is the predicted label (using the 0.5 threshold), y_i is the true label, and $\mathbb{I}$ is the indicator function.

Thus the actual accuracy of the test subset is plotted on the y-axis of the Figure, and the $\frac{n_p}{n}\%$ of cases with the lowest predicted error used as the test subset to determine the accuracy on the y-axis is plotted on the x-axis.

Figure 3 shows how reasonable suspicion model accuracy changes as we include more cases, sorted by their predicted error.

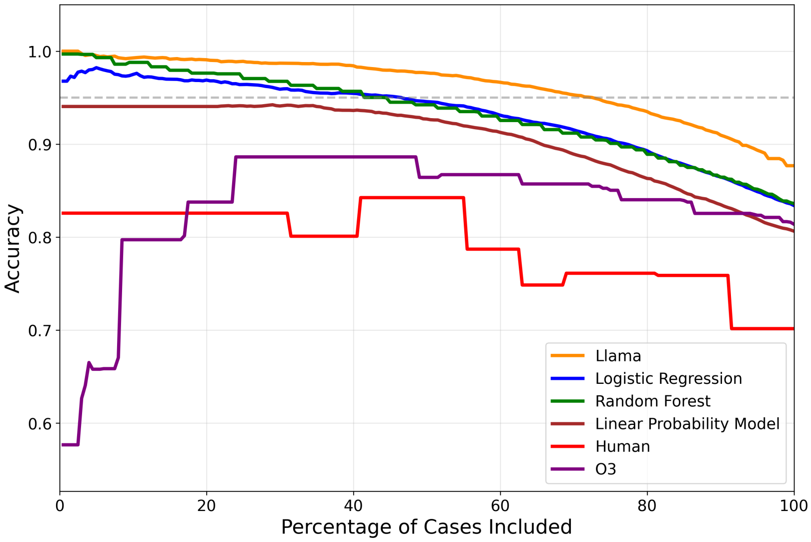

Figure 3. Accuracy of Multiple Models' Predictions by Confidence Ranking for Stops

Notes: This figure shows how the accuracy of different models' predictions of reasonable suspicion for stops varies as a function of the percentage of cases included, sorted by each model's confidence ranking. The x-axis represents the percentage of test cases included (from most to least confident), while the y-axis shows the accuracy achieved on that subset. The fine-tuned Llama 3 model (orange) maintains 95% accuracy (horizontal dashed line) while automatically classifying 74.6% of cases, substantially outperforming other approaches. When restricted to its most confident 50% of predictions, the Llama 3 model achieves 98.2% accuracy. Other models shown include Logistic Regression (blue), Linear Probability Model (brown), Random Forest (green), Human Coding (red), and OpenAI's o3 model (purple).

The resulting curves demonstrate clear performance differences among the models. The fine-tuned Llama 3 model achieves the best performance, maintaining 95% accuracy while automatically classifying nearly three-quarters (74.6%) of cases. In comparison, Logistic Regression can maintain 95% accuracy for 50.7%

of cases, while Random Forest achieves this threshold for 43.0% of cases. This suggests substantial potential for automated classification in practice—the Llama 3 model could reliably handle the majority of straightforward stops, providing important context to officers otherwise prone to much higher rates of error (19.7% in our sample).

Note that the models trained on actual data are well-calibrated, in the sense that the accuracy curve monotonically declines as more cases are included. This is not true either of the o3 model (which peaks in accuracy near 50% of cases, suggesting that the model is overconfident about its predictions at extremes) or the human RA's coding (which has multiple spikes in accuracy).

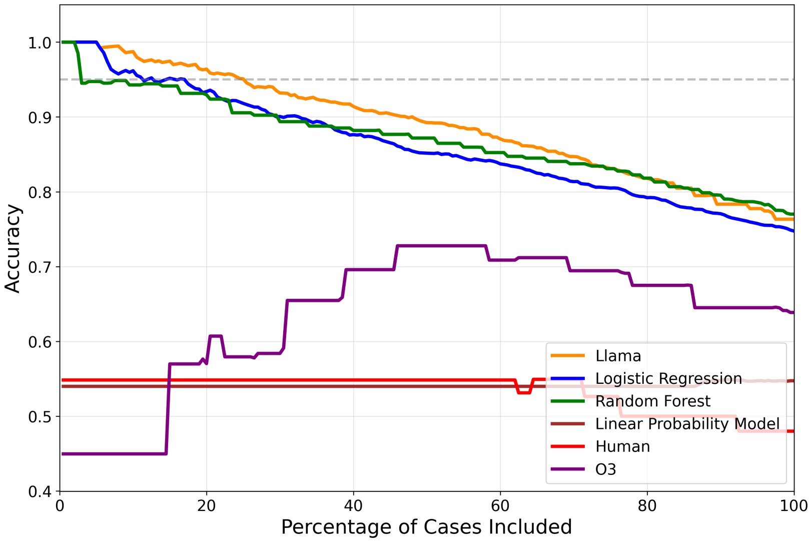

Figure 4. Accuracy of Multiple Models' Predictions by Confidence Ranking for Frisks

Notes: This figure shows how the accuracy of different models' predictions of reasonable suspicion for frisks varies as a function of the percentage of cases included, sorted by each model's confidence ranking. Similar to Figure 4 for stops, the x-axis represents the percentage of test cases included (from most to least confident), while the y-axis shows accuracy. The fine-tuned Llama 3 model maintains 95% accuracy while classifying only 29.6% of frisk cases, substantially lower than the 74.6% achieved for stops. This reduced performance across all models suggests that predicting reasonable suspicion for frisks is inherently more challenging than for stops.

On the other hand, the results of Figure 4 are less promising across all models. The fine-tuned Llama 3 model performs best but can only maintain 95% accuracy for 29.6% of cases, while Logistic Regression achieves this threshold for 20.7% of cases. The consistently lower performance across all models suggests that predicting frisks is a more difficult task than stops, a problem compounded by the smaller train set of frisks. Again, the models trained on actual data are well-calibrated, but the o3 model is not.

To demonstrate model calibration, Figures 5 and 6 show how model accuracy varies across different predicted probability ranges for stops and frisks, respectively. A well-calibrated model should exhibit a V-shaped curve in these plots (the black dashed lines represent perfect calibration), with high accuracy at both extremes (near 0 and 1) where the model is most confident, and lower accuracy near 0.5 where predictions are most uncertain.

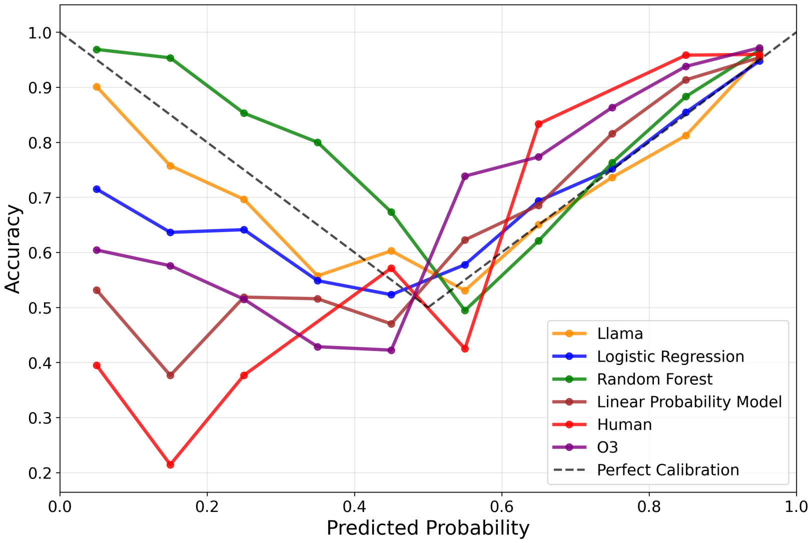

Figure 5. Model Accuracy by Predicted Probability Range for Stop Reasonable Suspicion

Notes: This figure shows how accuracy varies across different predicted probability ranges for each model when predicting reasonable suspicion in police stops. The x-axis represents the predicted probability of reasonable suspicion (binned into 10 equal intervals), while the y-axis shows the accuracy within each bin. The dashed black line represents perfect calibration.

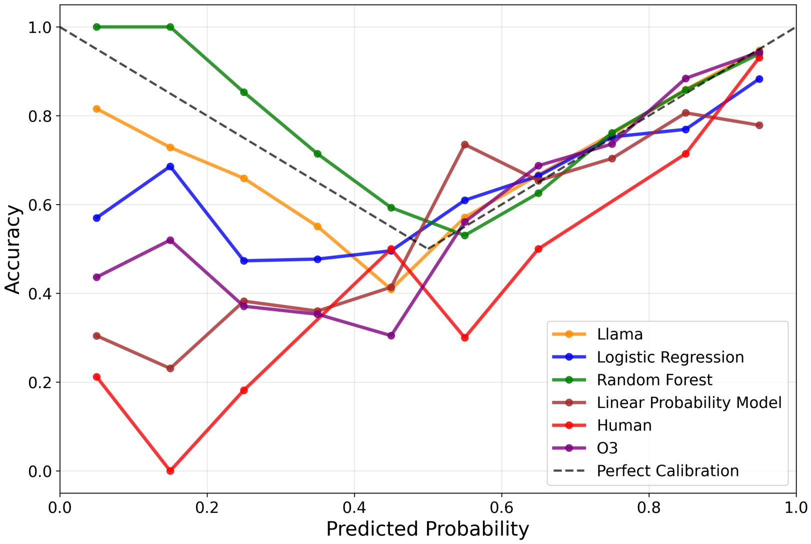

Figure 6. Model Accuracy by Predicted Probability Range for Frisk Reasonable Suspicion

Notes: This figure shows how accuracy varies across different predicted probability ranges for each model when predicting reasonable suspicion in police frisks. Similar to Figure 5, the x-axis represents the predicted probability of reasonable suspicion (binned into 10 equal intervals), while the y-axis shows the accuracy within each bin. The dashed black line represents perfect calibration.

For both stops and frisks, the fine-tuned Llama 3 model and classic ML approaches display the expected V-shaped pattern, achieving high accuracy when predicting probabilities far from 0.5. However, the o3 model and human coder deviate significantly from this ideal, particularly at the extremes. For frisks (Figure 6), these patterns are even more pronounced: the human coder's performance is notably poor across all probability ranges, with accuracy often below 70% even at the extremes where confidence should be highest. These calibration issues explain why the o3 and human models fail to produce monotonic accuracy curves when cases are ranked by confidence, as in Figures 3 and 4. Figures C1 and C2 in Appendix C.C4 provide additional information about calibration along the same lines.

These results are largely to be expected; the classic ML models and the LLM were trained on empirical data that improve calibration. In contrast, the o3 model and the human coders did not receive direct feedback about their miscalibration and therefore did not have an opportunity to correct for it. It is possible that their performance would have improved had they received such feedback (Mellers et al., 2014).

D. Topic Modeling and Decision Rules for Stops

Using the canonical topics identified through our topic modeling, we analyzed which topics most frequently serve as the primary justification for stops and how targeted policy interventions could reduce false positive rates.

Distribution of Main Reasons

The distribution of main reasons reveals important patterns in police stop justifications. Table 4 presents the five most common main reasons for stops. Notably, 15% of stops were justified primarily by the top 5 topics, and 25% of stops were justified primarily by the top 10 topics, suggesting that police stops are driven by a relatively concentrated set of circumstances.

Table 4—Top Five Main Reasons for Police Stops

Main Reason	**Count**	**Percentage of Stops**
Suspect Matches Description	431	3.75%
Person with Gun Call	357	3.10%
Theft in Progress Call	315	2.74%
Odor of Marijuana	308	2.68%
Public Urination	300	2.61%

Notes: This table presents the five most common main reasons identified as primary justifications for police stops in our test set. The main reason represents the single most important topic that motivated the officer's decision to conduct the stop, as determined by the o3 model's analysis of police narratives. These top five reasons account for 14.87% of all stops in the test set. The counts represent the number of stops where each topic was identified as the main reason.

The distribution of police stops across canonical topics identified as main reasons follows an approximately exponential distribution, as shown in Figure E1 in Appendix E; a small number of topics account for a large proportion of stops.

Simulated Policy Interventions

To evaluate potential policy interventions, we simulate the impact of instructing officers not to make stops when certain topics constitute the main reason for the stop. We rank topics by their effect on decreasing the aggregate false positive rate. This balances the false positive rate associated with each stop justification against the frequency of that justification; if some justification has a very high false positive rate but is very rare, telling officers to discontinue stops based on that justification will have little effect in reducing overall false positives. We assume perfect officer compliance with the policy interventions.

Table 5 presents the cumulative impact of progressively discontinuing stops based on the most problematic main reasons. The results demonstrate that discontinuing stops based on the single most problematic main reason topic would reduce the false positive rate from 19.5% to 18.9% while eliminating 3.1% of all stops. When extending this policy to the top five main reason topics, the false positive rate decreases to 17.6% while eliminating 8.7% of stops.

Table 5—Cumulative Impact of Discontinuing Stops by Main Reason Topics

Number of Topics Removed	**False Positive Rate**	**Percentage of Stops Removed**
0	19.5%	0.0%
1	18.9%	3.1%
2	18.4%	4.4%
3	18.1%	7.0%
4	17.9%	8.0%
5	17.6%	8.7%
10	16.8%	11.3%
15	15.8%	15.1%
20	15.5%	17.6%

Notes: This table shows the cumulative effect of removing stops where progressively more topics serve as the main reason, ranked by their efficiency in reducing false positive rates. The false positive rate represents the proportion of remaining stops that lack reasonable suspicion. The analysis assumes perfect officer compliance with policy interventions.

Figure E3 in the Appendix visualizes these results, showing how the false positive rate decreases as more topics are removed as valid reasons for stops, plotted against the percentage of total stops that would be affected by such policies.

Table 6 presents the five most effective topics to target for policy intervention—those that would reduce false positives the most.

Table 6—Top Five Topics for Reducing False Positive Rates in Police Stops

Topic	**Topic FPR**	**FPR Reduction**	**% of Stops**
Person with Gun Call	38.4%	3.10%	3.10%
Suspicious Conduct/Behavior	54.2%	2.25%	1.25%
Odor of Marijuana	30.2%	1.51%	2.68%
Panhandling	58.5%	1.44%	0.71%
Loitering	76.9%	1.00%	0.34%

Notes: This table identifies the five most effective topics to target for reducing false positive rates in police stops, based on the train set. Topics are ranked by their efficiency in reducing false positive rates—the reduction in unconstitutional stops relative to the total proportion of stops affected. Topic FPR represents the false positive rate within each topic category (proportion of stops lacking reasonable suspicion). FPR Reduction shows the relative percentage decrease in the overall false positive rate if stops based on this topic were discontinued (based solely on the train set). The analysis assumes perfect officer compliance with policy interventions.

These findings are further illustrated by comparing our decision rule approach with an alternative method using the Llama model's predicted probabilities directly. Figure 7 shows the false positive rate achieved when retaining different percentages of stops, comparing two approaches: (1) our decision rules based on removing stops with specific main reasons, and (2) ranking stops by the Llama model's predicted probability of reasonable suspicion and removing those with the lowest probabilities. The blue line represents the false positive rate when implementing decision rules to discontinue stops, where the decision rules are determined using the train set and implemented on the test set. This represents a realistic, cross-validated application of decision rules. The green line represents the false positive rate when decision rules are determined using the test set and implement on the test set. While not realistic, this represents the theoretical optimal performance from a set of decision rules.

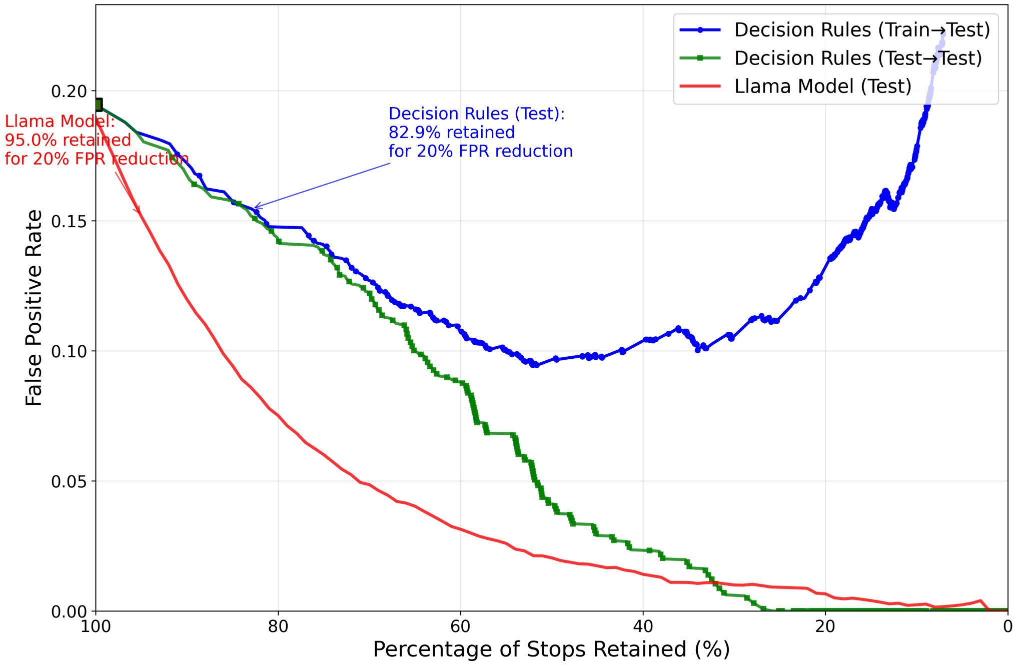

Figure 7. Comparison of False Positive Rate Reduction Strategies

Notes: This figure compares two approaches for reducing false positive rates in police stops. The x-axis shows the percentage of stops retained, while the y-axis shows the resulting false positive rate. The red line shows the performance of the Llama model, if we remove stops with the lowest predicted probabilities of reasonable suspicion (having calibrated the Llama model using the train set and then tested it using the test set). The blue and green lines represent a decision rule approach that progressively removes stops based on problematic main reason topics, ranked by their ability to reduce false positives. The blue line reflects the false positive rate when decision rules are determined using the train set and tested using the test set, and the green line represents the false positive rate when decision rules are both determined using the test set and tested using the test set. The green line therefore represents the theoretical optimal performance of a decision-rule approach, whereas the blue line represents a realistic implementation. The decision rules do not monotonically decrease the false positive rate in the blue line due to overfitting; certain rare justifications for stops have a high false positive rate in the train set and seem promising to remove but turn out to have a lower false positive rate in the test set.

The comparison reveals that while both approaches successfully reduce false positive rates, the Llama model consistently outperforms the realistic implementation (blue line) and also outperforms the theoretically optimal implementation of decision rules (green line) except at very high levels of stops discontinued (around 70% of stops discontinued). Table **??** quantifies this difference at key retention levels.

Stops Retained (%)	**Decision Rules FPR**	**Llama Model FPR**	**Difference**
90	0.172	0.120	+0.053
80	0.148	0.075	+0.073
70	0.127	0.049	+0.079
60	0.107	0.031	+0.076
50	0.097	0.020	+0.077

Table 7—Performance Comparison of Decision Rules vs. Llama Model at Key Retention Percentages

Notes: This comparison was generated by ranking all stops either by their main reason topic (for decision rules) or by their Llama-predicted probability of reasonable suspicion. The table shows false positive rates (FPR) at different retention levels. The "Difference" column shows the decision rules FPR minus the Llama model FPR; positive values indicate the Llama model achieves lower false positive rates. Results are based on a test dataset of 11,501 stops, and the decision rules were determined based on the train set and then evaluated on the test set (the blue line in the prior figure). For the decision rule approach, stops were progressively removed based on the ranking of main reason topics by how much their removal decreased the false positive rate of the remaining stops. For the Llama model approach, stops were ranked by their predicted probability of having reasonable suspicion, and those with the lowest probabilities were progressively removed.

This comparison was generated using the same test dataset of 11,501 stops. For the decision rule approach, stops were progressively removed based on the ranking of main reason topics by how much their removal would reduce false positives for the remaining set of stops. For the Llama model approach, stops were ranked by their predicted probability of having reasonable suspicion, and those with the lowest probabilities were progressively removed. At each retention level, we calculated the false positive rate among the remaining stops.

The superior performance of the Llama model approach suggests that while simple decision rules based on main reasons can reduce unconstitutional stops, the neural network captures more nuanced patterns that enable better discrimination between lawful and unlawful stops. However, the decision rule approach offers advantages in terms of interpretability and ease of implementation—officers can be given clear, actionable guidance about which types of stops to avoid, whereas the Llama model operates as a "black box" that would be more difficult to explain and implement in practice. We report the same analysis for frisks in Appendix F.

V. Discussion

A. Limitations

While our results suggest potential promising applications for LLMs in providing real-time feedback to law enforcement officers, there are a number of potential issues that could arise in implementation.

One broad issue is that the data that we use to train models were collected in a particular setting (ordinary policing without LLM feedback), and introducing LLM feedback could change that setting in a manner that affects the performance of the model. For instance, officers currently exercise their own judgment and plausibly only make stops when they believe that RAS is present. Live AI advice might change that decisionmaking dynamic—officers might ask the AI for advice about borderline cases where they believe that RAS is not actually present, in order to get a second opinion. Because none of these cases are presently in the train set, this could decrease the performance of our AI models when applied in the field. The extent of the problem is difficult to anticipate—it depends on the sorts of issues that would arise in these marginal cases and how much they differ from the current set of cases used to train the LLM.

Relatedly, the effect of AI feedback depends on behavioral questions in how officers respond to that feedback. Ideally, officers would now have a tight feedback loop from live advice and would quickly improve their instincts about the sorts of scenarios that have or lack RAS, and they would use these improved instincts to make fewer unconstitutional stops. But there are two potential alternatives, both problematic.

First, officers might simply ignore the AI's feedback and make stops as they were before. Whether this occurs or not depends on details of implementation—for instance, the specific instruction officers receive about how to use AI feedback and the informal emphasis placed by the department on compliance with AI instructions.

Second, officers could use this information not to make fewer unconstitutional stops, but to manipulate their descriptions of stops to trick their superiors and courts into allowing stops that are truly unconstitutional. That is, providing this feedback to officers could simply drive unconstitutional behavior underground. However, if AI advice is complemented by body-worn cameras, or if body-worn camera content is itself used to generate the AI advice, this potential problem may be less likely. Descriptions of stops produced by body-worn cameras are likely more difficult (although not impossible) to manipulate.

A final issue with our current approach is it the lack of explanation may meet resistance by officers in the field. Currently, our models simply produce a classification of whether RAS is present or not; it would be better to accompany these assessments with an explanation of *why* a stop has or lacks RAS. These could also be generated by an LLM and could draw on a database to identify the most likely legal issues with a given police stop.

We hope to conduct a follow-on field experiment providing AI feedback to actual law enforcement officers, in which case we can develop more explainable ML techniques and attempt to quantify the impact of the potential issues discussed above.

Note that none of the limitations discussed above apply in the context of using LLMs to audit police practice, as is currently done in many cities, making it a safe and useful starting point to deploy LLMs for classification of RAS.

B. Alternative Hypotheses

There are other police stop issues of substantial interest that we do not address here. Most notably, there is great interest (see subsection II.A) in assessing racial disparities in policing, and incorporating text data could potentially allow substantial improvement. Predicting which frisks most likely to result in contraband discovery is another topic of strong interest.

After some initial work on these topics, we choose not to include them because both suffer from what may be termed an endogeneity problem. Police officers may vary in how they describe a situation based on the race of the individual stopped or frisked. For example, consciously or unconsciously, they may use different wording, include or exclude certain facts, or vary their tone depending on the race of the individual stopped. If related to the outcome of interest (e.g. RAS) a good LLM will pick up on these differences.

As a simple example, assume an officer always uses the word “loiter” to describe a Black individual and “linger” for White individuals. If RAS is less common for Blacks than Whites, *loiter* will predict RAS. Including it as a control variable in an analysis of differential racial effects will lead to incorrectly underestimating racial differences.

A similar problem is present if trying to predict contraband detection. Since the officer knows whether or not contraband has been detected before writing the report, there could also be subtle differences in *wording*—but not in any real world actions—that the LLM would pick up as predictive of contraband. The officer may, for example, be more likely to mention a “bulge” when a weapon was in fact found, but omit it when not. Then *bulge* would be associated with contraband in the text, but not in reality. Since it is real-world predictors of contraband that would be of interest, this would not be helpful.

Importantly, neither of these issues is a concern in our task, predicting RAS. This is because what is of interest is the *narrative text itself.* If words in the narratives predict RAS for whatever reason, this is useful, since this is exactly the task of attorneys—to determine RAS based on text. There is somewhat greater concern in considering the usefulness of the predictions for the potential future applications of giving real-time advice - narratives that suffer from ex-post bias. However, the presence of RAS is crucially different from contraband discovery in that *officers believe there is RAS in 100% of stops.* Thus, there should not be conscious or unconscious wording variation based on their RAS assessment if they believe it to always be present.

Still, application of this methodology for real-time legal assessments would certainly require training with video and audio data from body-worn cameras. Relying primarily or exclusively on objective sources of information would reduce or eliminate any incentives for officers to "shade" narratives in ways that are ex-post favorable. There are potential costs from this approach, however, in loss of information and judgment from officers, as well as potential decreased job satisfaction from reduced officer autonomy.

VI. Conclusion

We evaluated multiple approaches, including classic ML algorithms (random forest, logistic regression, and LPM), OpenAI's state-of-the-art o3 model with detailed legal prompting, and fine-tuned Llama 3 models. Our results demonstrate that fine-tuned Llama 3 models can effectively predict the legality of both stops and frisks, with the stop model achieving 87.9% accuracy and the frisk model achieving 76.9% accuracy. Moreover, fine-tuned Llama 3 was even more accurate on subsets for which it had high confidence, with 95% accuracy on 74.6% of stops and 29.6% of frisks. Fine-tuned Llama 3 models consistently outperformed all ML models, as well as our human RA.

The fact that fine-tuned Llama 3 outperformed o3, even though o3 is a much stronger model absent fine-tuning, suggests the importance of fine-tuning in order to capture domain-specific knowledge in criminal law. The fact that Llama 3 substantially simpler ML models, like random forests and logistic regression, suggests that understanding context and nuance (i.e., Llama 3's ability to understand the relationships between words rather than treating narratives as bags of words) is also important.

In addition, the fact that justifications for stops are so diverse, and the difficulty in formulating simple decision rules that can substantially reduce unconstitutional stops (especially when compared to Llama 3) underscores the complexity of RAS determinations in the field.

The ability of algorithms to classify police narratives as having or lacking reasonable suspicion for both stops and frisks has important practical implications. Most immediately, our methods enable low-cost audits of the legality of police departments' current practices. Rather than relying on small samples or costly large-scale audits involving human lawyers, departments could use LLMs to efficiently identify problems in policing, and plaintiff's lawyers could use LLMs to decrease the costs of litigation.

Contingent on details of implementation, our method could also provide live feedback to officers before making stops or conducting frisks, advising them whether their actions are likely to be deemed illegal. This is particularly important for frisks, where officers must meet the additional burden of articulating why they believe a suspect is armed and dangerous. By providing real-time guidance, these algorithms could help reduce unconstitutional police encounters while maintaining public safety.

REFERENCES

Antonovics, Kate, and Brian G. Knight. 2009. "A New Look at Racial Profiling: Evidence from the Boston Police Department." *The Review of Economics and Statistics*, 91: 163–177.

Anwar, Shamena, and Hanming Fang. 2006. "An Alternative Test of Racial Prejudice in Motor Vehicle Searches: Theory and Evidence." *The American Economic Review*, 96: 127–151.

Bojić, Ljubisa, Olga Zagovora, Asta Zelenkauskaite, et al. 2025. "Comparing large language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm." *Scientific Reports*, 15: 11477.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. "Language Models are Few-Shot Learners." *Advances in Neural Information Processing Systems*, 33: 1877–1901.

Campbell, Romaine A. 2025. "What Does Federal Oversight Do to Policing and Public Safety? Evidence from Seattle."

Chalfin, Aaron, and Felipe M Gonçalves. 2023. "Professional Motivations in the Public Sector: Evidence from Police Officers."

Chalfin, Aaron, and Justin McCrary. 2018. "Are U.S. cities underpoliced? Theory and evidence." *Review of Economics and Statistics*, 100: 167–186.

Chen, Po-Hsuan Cameron, Craig H Mermel, and Yun Liu. 2021. "Evaluation of artificial intelligence on a reference standard based on subjective interpretation." *Lancet*, 3(11): e693–e695. Comment.

Choi, Jonathan H., Amy B. Monahan, and Daniel Schwarcz. 2024. "Lawyering in the Age of Artificial Intelligence." *Minnesota Law Review*, 109(1): 147–218.

Choi, Jonathan H., and Daniel Schwarcz. 2025. "AI Assistance in Legal Analysis: An Empirical Study." *Journal of Legal Education*, 73(2): 384–.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*, 4171–4186.

Dharmapala, Dhammika, and Stephen L. Ross. 2004. "Racial Bias in Motor Vehicle Searches: Additional Theory and Evidence." *Contributions to Economic Analysis Policy*, 3: Art. 12.

Dietterich, Thomas G. 2000. "Ensemble Methods in Machine Learning." Vol. 1857 of *Lecture Notes in Computer Science*, 1–15. Berlin, Heidelberg:Springer.

Ding, Ning, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. 2023. "Parameter-efficient fine-tuning of large-scale pre-trained language models." *Nature Machine Intelligence*, 5(3): 220–235.

Durlauf, Steven N. 2006. "Assessing Racial Profiling." *The Economic Journal*, 116: F402–F426.

Feigenberg, Benjamin, and Conrad Miller. 2022. "Would Eliminating Racial Disparities in Motor Vehicle Searches have Efficiency Costs?" *Quarterly Journal of Economics*, 137: 49–113.

Ferguson, Andrew Guthrie. 2024. "Generative Suspicion and the Risks of AI-Assisted Police Reports." *Northwestern Law Review.* forthcoming.

Gelbach, Jonah B. 2021. "TESTING ECONOMIC MODELS OF DISCRIMINATION IN CRIMINAL JUSTICE."

Goh, Ethan, Robert Bunning, Elaine Khoong, et al. 2024. "Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial." *JAMA Network Open*, 7(10): e2440969.

Hansen, Lars Kai, and Peter Salamon. 1990. "Neural Network Ensembles." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10): 993–1001.

Hausman, David, and Dorothy Kronick. 2023. "The illusory end of stop and frisk in Chicago?" *Science Advances*, 9: eadh3017.

Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. "LoRA: Low-Rank Adaptation of Large Language Models." Poster.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics*, 133(1): 237–293.

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

Lin, Jennifer J, Kabisha Malla, Natasha Cartwright, et al. 2023. "Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology." *Scientific Reports*, 13: 18562.

Martin, Lauren, Nick Whitehouse, Stephanie Yiu, Lizzie Catterson, and Rivindu Perera. 2024. “Better Call GPT: Comparing Large Language Models Against Lawyers.” arXiv:2401.16212.

Mellers, Barbara, Lyle Ungar, Jonathan Baron, Jaime Ramos, Burcu Gurcay, Katrina Fincher, Sydney E Scott, Don Moore, Pavel Atanasov, Samuel A Swift, et al. 2014. “Psychological strategies for winning a geopolitical forecasting tournament.” *Psychological Science*, 25(5): 1106–1115.

Movva, Rajiv, Pang Wei Koh, and Emma Pierson. 2024. “Annotation alignment: Comparing LLM and human annotations of conversational safety.”

Nay, John J., David Karamardian, Sarah B. Lawsky, Wenting Tao, Meghana Bhat, Raghav Jain, Aaron Travis Lee, Jonathan H. Choi, and Jungo Kasai. 2024. “Large Language Models as Tax Attorneys: A Case Study in Legal Capabilities Emergence.” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 382(2270).

New York Civil Liberties Union. 2025. “Stop-and-Frisk Data.” *https://www.nyclu.org/data/stop-and-frisk-data*, Accessed: 2025-06-27.

Nori, Harsha, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P Lungren, Bay Gross, Peter Hames, Mustafa Suleyman, Dominic King, and Eric Horvitz. 2025. “Sequential Diagnosis with Language Models.”

Oliver, Wesley M., Morgan A. Gray, Jaromir Savelka, and Kevin D. Ashley. 2024. “Computationally Assessing Suspicion.” *University of Cincinnati Law Review*, 92(4): 1108.

Pham, Chau Minh, Alexander Hoyle, Simeng Sun, and Mohit Iyyer. 2023. “TopicGPT: A Prompt-based Topic Modeling Framework.”

Plank, Barbara. 2022. “The “Problem” of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation.” 10671–10682. Abu Dhabi, United Arab Emirates:Association for Computational Linguistics.

Polikar, Robi. 2006. “Ensemble based systems in decision making.” *IEEE Circuits and Systems Magazine*, 6(3): 21–45.

Rivera, Roman G, and Bocar A. Ba. 2025. “The Effect of Police Oversight on Crime and Misconduct Allegations: Evidence from Chicago *.” *Review of Economics and Statistics.*

Savelka, Jaromir, and Kevin D Ashley. 2023. “The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts.” *Frontiers in Artificial Intelligence*, 6: 1279794.

Shao, Peizhang, Linrui Xu, Jinxi Wang, Wei Zhou, and Xingyu Wu. 2025. "When Large Language Models Meet Law: Dual-Lens Taxonomy, Technical Advances, and Ethical Governance." *ACM Computing Surveys*, 37(4): 111.

Appendix A: Data Processing, Fine-Tuning, and Prompt Engineering

A1. 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) **Age Processing**:
   - Ages were converted to numeric values.
   - Observations with ages younger than 9 or older than 99 were dropped.
   - Negative ages and age 0 were dropped.
3) **Physical Characteristics**:
   - Height values outside the range of 4’6” to 7’ were dropped.
   - Weight values lower than 50 lbs or greater than 500 lbs were dropped.
   - Unrealistic weight-build combinations (e.g., ’Thin’ with weight greater than 300 lbs) were dropped.
4) **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).
5) **Location Information**:
   - Police Service Area (PSA) 7700 was removed.
6) **Text Cleaning**:
   - ’apos;’ was replaced with apostrophes in the stop narratives.
   - Newline characters were replaced with spaces in the stop narratives.
7) **Rare Value Handling**:
   - Observations with rare values (frequency $< 0.5\%$) Eye Color and Build were dropped.
   - Observations involving officers with fewer than 10 stops were dropped.
8) **Duplicate Removal**: Duplicate rows were removed if they matched based on the following variables: location, officer ID, further suspect description, suspect sex, suspect age, suspect height.

A2. LoRA Fine-Tuning

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 (Hu 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 $\mathbf{B} \in \mathbb{R}^{d \times r}$ and $\mathbf{A} \in \mathbb{R}^{r \times k}$, where r is the chosen rank (typically much smaller than d and k):

$$\mathbf{W} = \mathbf{W}_0 + \alpha \mathbf{B}\mathbf{A} \tag{A1}$$

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:

$$\mathbf{h} = \mathbf{W}\mathbf{x} = \mathbf{W}_0\mathbf{x} + \alpha(\mathbf{B}\mathbf{A})\mathbf{x} \tag{A2}$$

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., 2022).

A3. LoRA Fine-Tuning Hyperparameters

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

- **Base Model:** We used a 4-bit quantized version of the Llama 3.2-8B Instruct model, which allows for faster loading and reduced memory usage.
- **Sequence Length:** The maximum sequence length was set to 2048 tokens.
- **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.

A4. OpenAI o3 Model Prompts

We evaluated OpenAI's o3 model using two different prompt styles for each prediction task: a short prompt and a detailed prompt that incorporates additional information about the relevant legal framework. The detailed prompts performed better and were used for the results reported in the main text.

Stop Reasonable Suspicion Prompts

Short Prompt:

> You are being given the text of a police report describing a police stop in [CITY]. Determine whether the police had reasonable suspicion to make the stop, as required under the Supreme Court's decision in Terry v. Ohio, as interpreted under the laws of [CITY] and [STATE].
>
> Express your answer on a scale from 0 (meaning there was definitely no reasonable suspicion) to 100 (meaning there was definitely reasonable suspicion). Give only the number in your answer.

Detailed Prompt:

> You are an expert on Fourth-Amendment "stop-and-frisk" law in [STATE].
>
> Below you will receive the narrative portion of a police department stop form.
>
> Your single task is to decide whether, **based only on what the officer actually wrote**, the officer had *reasonable suspicion to stop* the person(s) under **Terry v. Ohio**, as that doctrine is applied in [CITY] and [STATE].

LEGAL FRAMEWORK – APPLY *ONLY* TO THE STOP (not frisk, search, or arrest)

1. **Three required elements of reasonable suspicion (RS)**

- **Person-specific** – facts tie the suspicion to the particular individual stopped.
- **Reliability** – the information is reasonably trustworthy. Direct observation ≥ identified citizen > known informant > anonymous tip.
- **Crime-suggestive** – facts indicate past, present, or imminent criminal activity (mere "suspicious" behavior is not enough).

2. **Sources of information and their weight**

- **Officer's own observations** – high reliability; ordinarily satisfy both person-specific and reliability.
- **Flash / radio / third-party report** – medium; require (a) a concrete description that the stopped person matches **and** (b) an identified or clearly reliable source.
- **Anonymous information alone** – low; *never* enough unless independently corroborated by the officer's own observations.

3. **Examples that *can* justify a stop (when all three elements are met)**

- Observed drug transaction ("observed transaction").
- Specific description plus corroboration of a violent crime suspect (e.g., "orange shirt, camo jacket, brown pants, breaking car windows with a hammer").
- Report of person with a gun *plus* matching description or flight or visible bulge.
- Observed open container, apparent intoxication, curfew/truancy when age plausible (<18/<17).
- Observed trespass inside abandoned property, casing cars, aggressive panhandling (with conduct described).
- Officer personally knows subject has an outstanding warrant before the stop.
- Immediate reaction to **gunshots** in the vicinity (stops made moments after shots are treated as RS).

4. **Examples that *never* alone justify a stop (prohibited by the PPD consent decree)**

   - "Loitering," "loitering in a high-crime/drug area."
   - "High-crime area" standing alone.
   - "Suspicious," "furtive movements," "investigation of person," "acting nervous."
   - Anonymous "man with gun" (no corroboration).
   - Merely riding in a vehicle stopped for a traffic infraction (passenger).
   - Vague "involved in a disturbance" unless the disturbance is described (e.g., fight witnessed).
   - General driver/traffic infractions are assumed valid for the driver; do **not** create RS to detain a pedestrian or passenger unless additional facts link that person to a crime.

5. **Special notes**

   - If the narrative plainly records something that is *not* really a "stop" (e.g., executing an arrest warrant, assisting a complainant), output **"n/a"** instead of a number.

6. **Scoring scale – output must be a single integer 0-100**

   - 0–10: Clearly no RS (only prohibited or vague reasons).
   - 11–30: Very weak; maybe one element partially present.
   - 31–49: Insufficient; at least one element missing.
   - 50: Borderline; elements just barely satisfied.
   - 51–70: More likely than not that RS exists.
   - 71–89: Strong RS; all three elements well supported.
   - 90–100: Definitive RS; detailed, specific, reliable facts of criminal activity.

7. **Method**

   - Read only what is written. Do not invent facts or discount ones that are written, even if implausible.
   - Apply sections 1–4 to decide whether all three elements are met.
   - Choose the score from section 6 that best reflects the strength of RS.
   - Respond with only:
     - the integer score, or

- "n/a" if no stop analysis is required.

WHEN YOU RESPOND, PROVIDE ONLY THE INTEGER (or "n/a"). NO WORDS, NO EXPLANATION.

Frisk Reasonable Suspicion Prompts

Short Prompt:

You are being given the text of a police report describing a police stop. Determine whether the police had reasonable suspicion to frisk the person, as required under the Supreme Court's decision in Terry v. Ohio, as interpreted under the laws of [CITY] and [STATE].

Express your answer on a scale from 0 (meaning there was definitely no reasonable suspicion to frisk) to 100 (meaning there was definitely reasonable suspicion to frisk). Give only the number in your answer.

Detailed Prompt:

You are an expert on Fourth-Amendment "stop-and-frisk" law in STATE. Below you will receive the narrative portion of a police department stop form. Your single task is to decide whether, **based only on what the officer actually wrote**, the officer had *reasonable suspicion to frisk* the person(s) under **Terry v. Ohio**, as that doctrine is applied in [CITY] and [STATE].

LEGAL FRAMEWORK – APPLY *ONLY* TO THE FRISK (not stop, search, or arrest)

---

1. **Three required elements of reasonable suspicion (RS)**
   - **Person-specific** – facts tie the suspicion to the particular individual frisked.
   - **Reliability** – the information is reasonably trustworthy. Direct observation $\geq$ identified citizen $>$ known informant $>$ anonymous tip.
   - **Weapon or danger indicator** – facts suggest the person is armed and presently dangerous.
2. **Sources of information and their weight**
   - **Officer's own observations** – high reliability; still need a weapon/danger indicator.

- **Identified victim or eyewitness** – medium-high; officer must note how the source is identifiable.
- **Known informant with a track record** – medium; officer must say the informant has been reliable before.
- **Anonymous tip or vague "flash"** – low; needs independent corroboration by the officer.

3. **Examples that *can* justify a frisk (when all three elements are met)**

- Visible weapon or unmistakable bulge consistent with a weapon.
- Refusal to remove hands from pockets after the officer directs the person to do so.
- Nature of suspected crime is violent (robbery, assault, gun offense).
- Reliable report of a gun plus match to description and corroborating behavior (flight, visible bulge, etc.).
- Violent disturbance in progress (fight, domestic assault) observed or reliably reported.
- Gunshots just heard nearby; officer stops persons fleeing or lingering at the scene.
- Specific description of an armed suspect matched by the stopped person.
- Officer knew, before the stop, of an outstanding warrant for a violent or weapons offense.

4. **Examples that *never* alone justify a frisk (prohibited by the PPD consent decree)**

- High-crime or high-drug area by itself.
- "Furtive movements," "acting suspiciously," "belligerent," "investigation of person."
- Loitering or loitering in a high-crime area.
- Anonymous unverified "man with gun."
- Drug possession or transaction alone (unless other violent-crime indicators are present).
- Merely placing the person in a patrol car for officer "safety."
- General nervousness, ordinary hands-in-pockets, or being a passenger in a traffic stop.

5. **Special notes**

- The officer must have specific, articulable facts that would lead a reasonable officer to believe the person is armed and presently dangerous.
- Evaluate the frisk separately from the stop. A bad stop does not automatically invalidate a good frisk (and vice-versa), though evidence from an unlawful stop can be suppressed as "fruit of the poisonous tree."
- A clear weapon-shaped bulge, by itself, is barely enough but still counts as RS.
- Hands-in-pockets plus refusal to remove them plus a context involving possible violence can supply RS.
- If the narrative is not really about a frisk (e.g., search incident to arrest, warrant execution, medical assist), output "n/a" instead of a number.

6. **Scoring scale – output must be a single integer 0-100**
   - 0–10: Clearly no RS (only prohibited or vague reasons).
   - 11–30: Very weak; maybe one element partially present.
   - 31–49: Insufficient; at least one element missing.
   - 50: Borderline; elements just barely satisfied.
   - 51–70: More likely than not that RS exists.
   - 71–89: Strong RS; all three elements well supported.
   - 90–100: Definitive RS; detailed, specific, reliable facts of a weapon or danger.
7. **Method**
   - Read only what is written. Do not invent facts or discount ones that are written, even if implausible.
   - Apply sections 1–4 to decide whether all three elements are met.
   - Choose the score from section 6 that best reflects the strength of RS.
   - Respond with only:
     - the integer score, or
     - "n/a" if no frisk analysis is required.

WHEN YOU RESPOND, PROVIDE ONLY THE INTEGER (or "n/a"). NO WORDS, NO EXPLANATION.

Appendix B: Feature Configurations for Model Evaluation

To optimize the performance of our models, we evaluated three different feature configurations with different levels of detail. This comparison allowed us to determine whether additional metadata beyond the police narrative would improve prediction performance.

Based on the performance of the configurations as described below, we ultimately decided to use police narratives alone. Using only police narratives also had the benefit of simplicity and real-world feasibility.

B1. Overview of Feature Configurations

We developed three distinct configurations designed to answer key questions about model performance and practical deployment:

Configuration 1: Narrative Only. This configuration uses only the police officer's written narrative description of the stop, representing the core information that officers must provide to justify their actions under *Terry v. Ohio*. This approach offers simplicity, interpretability, and eliminates the risk of the model learning demographic biases.

Configuration 2: Pre-Stop Information. This configuration includes the police narrative plus information readily available to officers before making a stop, enabling potential real-time guidance. This tests whether contextual information about the officer, location, and timing improves predictions while remaining practical for field deployment.

Configuration 3: All Variables. This configuration includes all available information, including post-stop data that could only be obtained after the stop was completed. While not suitable for real-time use, this configuration allowed us to test whether additional information improved the model's ability to predict legal assessments and whether demographic factors influenced coding decisions.

B2. Mathematical Formulation

We formally define each feature configuration as follows:

Configuration 1: Narrative Only

For the narrative-only configuration, we use only the police report text:

$$\mathbf{x}_{narr,i} = f(r_i) \tag{B1}$$

where $f(\cdot)$ denotes the text transformation function (bag-of-words with 5,000 vocabulary size for baseline models, or tokenization for language models), and r_i is the police stop narrative for observation i.

Configuration 2: Pre-Stop Information

The pre-stop configuration adds information available before the stop to the narrative:

$$\mathbf{x}_{pre,i} = [f(r_i); f(d_i); \mathbf{c}_i^{\text{pre}}; m_i] \tag{B2}$$

where the components are:

- $f(r_i)$: transformed police narrative
- $f(d_i)$: transformed free-text description of the suspect (if available)
- $\mathbf{c}_i^{\text{pre}} = [\mathbf{o}_i; \mathbf{p}_i; w_i; \mathbf{t}_i]$: concatenation of pre-stop categorical features
- m_i: linear time (minutes since 2000)

The pre-stop categorical features $\mathbf{c}_i^{\text{pre}}$ include:

- $\mathbf{o}_i$: officer fixed effects vector
- $\mathbf{p}_i$: police service area fixed effects vector
- w_i: binary indicator for whether officer had a partner
- $\mathbf{t}_i$: time of day fixed effects vector (daytime/evening/night)

Configuration 3: All Variables

The "all variables" configuration adds post-stop information:

$$\mathbf{x}_{all,i} = [f(r_i); f(d_i); \mathbf{c}_i^{\text{pre}}; m_i; \mathbf{c}_i^{\text{post}}; \mathbf{n}_i^{\text{post}}] \tag{B3}$$

where the additional components are:

- $\mathbf{c}_i^{\text{post}} = [\mathbf{s}_i; \mathbf{a}_i; \mathbf{l}_i; \mathbf{e}_i; \mathbf{b}_i; \mathbf{h}_i; \mathbf{k}_i; \mathbf{j}_i]$: post-stop categorical features
- $\mathbf{n}_i^{\text{post}} = [g_i; v_i]$: post-stop numerical features

The post-stop categorical features are:

- $\mathbf{s}_i$: sex indicator
- $\mathbf{a}_i$: race fixed effects vector
- $\mathbf{l}_i$: Latino/Hispanic indicator
- $\mathbf{e}_i$: eye color fixed effects vector
- $\mathbf{b}_i$: build fixed effects vector (e.g., thin, medium, heavy)
- $\mathbf{h}_i$: hair color fixed effects vector

- $\mathbf{k}_i$: complexion fixed effects vector
- $\mathbf{j}_i$: reason for stop fixed effects vector

The post-stop numerical features are:

- g_i: height
- v_i: weight

B3. Implementation Details

For baseline machine learning models (Random Forest, Logistic Regression, LPM), text features were created using count vectorization with a maximum vocabulary of 5,000 words and English stop words removed. All categorical variables were one-hot encoded with unknown categories handled by ignoring them during transformation. Numerical features were standardized to zero mean and unit variance using train set statistics.

For the Llama 3 model, the exact prompt format for stops using the narrative-only configuration was:

> "Here is a police report provided by a police officer containing details about a pedestrian stop: [REASON_DESC]. 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'."

For configurations with additional features, for Llama 3 the relevant information was incorporated into structured prompts that maintained the same basic format while including the supplementary data fields as additional text in the prompt. For the baseline machine learning models, numerical and categorical features were directly encoded as numerical variables and one-hot fixed effects, respectively.

B4. Summary of Feature Configurations

Table B1 provides a comprehensive overview of the data fields included in each configuration:

Table B1—Data Fields Included in Each Feature Configuration

Data Field	Narrative-Only	Pre-Stop	All Variables
Core Information			
Police Narrative (REASON_DESC)	✓	✓	✓
Pre-Stop Information (Available Before Stop)			
Officer ID		✓	✓
Police Service Area		✓	✓
Officer Has Partner		✓	✓
Time of Day		✓	✓
Date/Time of Incident		✓	✓
Further Description		✓	✓
Post-Stop Information (Available Only After Stop)			
Suspect Sex			✓
Suspect Race			✓
Suspect Latino/Hispanic			✓
Suspect Age			✓
Suspect Height			✓
Suspect Weight			✓
Suspect Build			✓
Eye Color			✓
Hair Color			✓
Complexion			✓
Reason for Stop Category			✓

Appendix C: Baseline Model Specifications

This appendix describes in detail the baseline machine learning models evaluated in our study. All models were implemented in Python using scikit-learn and trained on the same data splits as the fine-tuned Llama 3 model to ensure a fair comparison. The feature configurations used for these models are described in Appendix B.

C1. Model Specifications

Random Forest

The Random Forest model constructs an ensemble of decision trees, where each tree T_m is trained on a bootstrap sample of the data with random feature selection at each split:

$$\hat{p}(y=1|\mathbf{x}) = \frac{1}{M}\sum_{m=1}^{M}\mathbb{I}[T_m(\mathbf{x})=1] \tag{C1}$$

where M is the number of trees and $\mathbb{I}[\cdot]$ is the indicator function.

Hyperparameters: The model uses 100 estimators with maximum features per split set to $\sqrt{p}$, where p is the total number of features. The minimum samples per split is 2, minimum samples per leaf is 1, and bootstrap sampling is enabled.

Logistic Regression

Logistic regression models the probability of reasonable suspicion using:

$$p(y=1|\mathbf{x}) = \frac{1}{1+\exp(-(\mathbf{w}^T\mathbf{x}+b))} \tag{C2}$$

The model is trained by minimizing the regularized negative log-likelihood:

$$\min_{\mathbf{w},b} -\sum_{i=1}^{n}[y_i\log(p_i)+(1-y_i)\log(1-p_i)] + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \tag{C3}$$

where $p_i = p(y=1|\mathbf{x}_i)$ and λ is the regularization strength.

The model uses L2 regularization (ridge regression) with regularization strength $\lambda = 1.0$. The lbfgs solver is employed with a maximum of 1,000 iterations.

Linear Probability Model (LPM)

The Linear Probability Model uses ordinary least squares regression (without regularization) to directly predict the probability:

$$\hat{y}_i = \mathbf{w}^T\mathbf{x}_i + b \tag{C4}$$

The model minimizes the squared error loss:

$$\min_{\mathbf{w},b}\sum_{i=1}^{n}(y_i - \mathbf{w}^T\mathbf{x}_i - b)^2 \tag{C5}$$

Since OLS predictions are not constrained to [0,1], we clip the predictions:

$$\hat{p}(y=1|\mathbf{x}) = \max(0,\min(1,\mathbf{w}^T\mathbf{x}+b)) \tag{C6}$$

C2. Implementation Details

All models were trained using the 55%/15%/30% train/validation/test split used for all analysis in this paper. For text processing, the bag-of-words features were created using a maximum vocabulary size of 5,000 words with English stop words from NLTK removed. The token pattern followed the default configuration, capturing words of 2 or more alphanumeric characters. Text was converted to lowercase, and we used term frequencies rather than binary features.

Missing values were handled systematically across different data types. Text fields were replaced with empty strings, categorical variables were replaced with a 'missing' category, and numerical variables were replaced with 0 after standardization.

All sparse matrices from text vectorization and one-hot encoding were combined using scipy.sparse.hstack to maintain memory efficiency.

C3. Model Performance with Different Feature Configurations

Tables C1 and C2 present the complete performance metrics on the validation set for all models across all three feature configurations described in Appendix B. We used the validation set because we consider the choice of optimal feature configuration to be a pseudo-hyperparameter that should be selected during validation. We evaluated fine-tuned Llama 3, classic machine learning algorithms (logistic regression, random forest, and LPM), and OpenAI's o3 model with both short and detailed prompts. The o3 model was evaluated on a randomly selected subset of 500 observations from the validation set to reduce API token costs.

Table C1—Validation Set Performance for Stop Reasonable Suspicion Prediction

Model	**Variables**	**AUC-ROC**	**F1**	**Acc**	**Acc @ 50%**	**Max Subset w/ 95% Acc**
Llama 3	All Variables	0.900	0.928	0.880	0.975	71.2
	Pre-Stop	0.701	0.879	0.798	0.886	0.0
	Narrative	0.896	0.926	0.876	0.976	72.9
Log Reg	All Variables	0.821	0.900	0.833	0.950	50.5
	Pre-Stop	0.821	0.899	0.832	0.950	47.1
	Narrative	0.827	0.903	0.839	0.952	51.1
LPM	All Variables	0.735	0.869	0.785	0.908	0.5
	Pre-Stop	0.735	0.870	0.786	0.913	2.3
	Narrative	0.786	0.888	0.810	0.942	46.0
Rand Forest	All Variables	0.792	0.903	0.827	0.931	37.5
	Pre-Stop	0.805	0.905	0.831	0.941	45.4
	Narrative	0.815	0.910	0.841	0.942	45.3
o3 (Detailed)	All Variables	0.778	0.879	0.800	0.916	19.3
	Pre-Stop	0.741	0.888	0.816	0.912	17.6
	Narrative	0.770	0.888	0.817	0.937	17.2
o3 (Short)	All Variables	0.685	0.862	0.771	0.833	0.0
	Pre-Stop	0.742	0.863	0.775	0.816	28.9
	Narrative	0.710	0.876	0.796	0.857	13.0

Table C2—Validation Set Performance for Frisk Reasonable Suspicion Prediction

Model	**Variables**	**AUC-ROC**	**F1**	**Acc**	**Acc @ 50%**	**Max Subset w/ 95% Acc**
Llama 3	All Variables	0.753	0.861	0.778	0.897	7.9
	Pre-Stop	0.714	0.844	0.752	0.873	0.0
	Narrative	0.772	0.855	0.772	0.918	24.3
Log Reg	All Variables	0.734	0.829	0.739	0.897	19.7
	Pre-Stop	0.733	0.832	0.742	0.897	15.1
	Narrative	0.731	0.833	0.744	0.887	22.1
LPM	All Variables	0.622	0.712	0.615	0.832	0.0
	Pre-Stop	0.619	0.712	0.611	0.821	0.0
	Narrative	0.531	0.651	0.549	0.770	0.5
Rand Forest	All Variables	0.734	0.869	0.776	0.852	12.8
	Pre-Stop	0.755	0.871	0.778	0.873	26.4
	Narrative	0.753	0.871	0.782	0.883	15.1
o3 (Detailed)	All Variables	0.728	0.753	0.660	0.874	33.3
	Pre-Stop	0.665	0.724	0.630	0.838	5.7
	Narrative	0.692	0.725	0.636	0.860	16.7
o3 (Short)	All Variables	0.629	0.731	0.634	0.813	5.4
	Pre-Stop	0.631	0.722	0.636	0.882	17.8
	Narrative	0.635	0.681	0.592	0.723	0.9

C4. Additional Calibration Graphs

Figures C1 and C2 provide another illustration of model calibration for stops and frisks, respectively. They take each set of model predictions, bucket them into deciles, and plot the average probability that the model predicts of reasonable suspicion within each decile against the actual percentage of the time that the expert coder decided that reasonable suspicion was present.

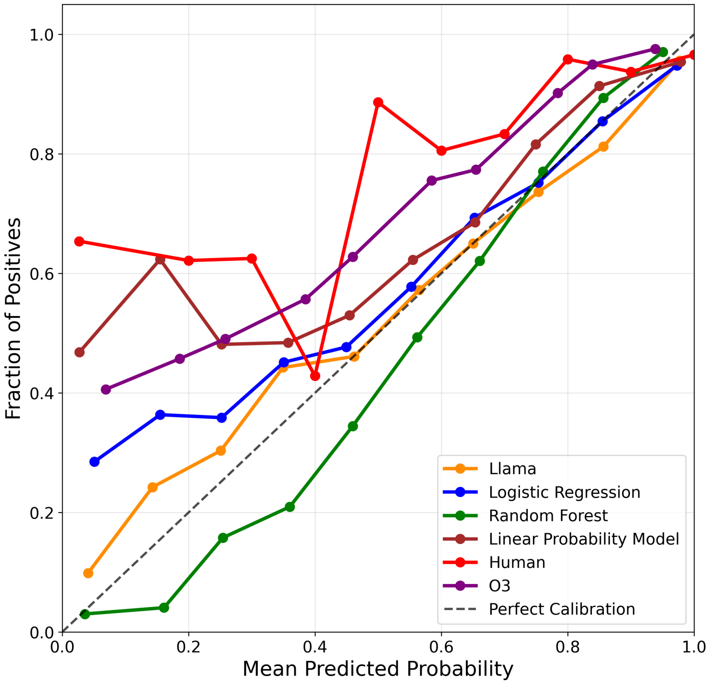

Figure C1. Model Calibration for Stop Reasonable Suspicion Predictions

Notes: This figure illustrates the calibration of different models' predictions for reasonable suspicion in police stops. Each point represents a decile of model predictions, with the x-axis showing the average predicted probability of reasonable suspicion within that decile and the y-axis showing the actual percentage of stops with reasonable suspicion. A perfectly calibrated model would follow the diagonal dashed line.

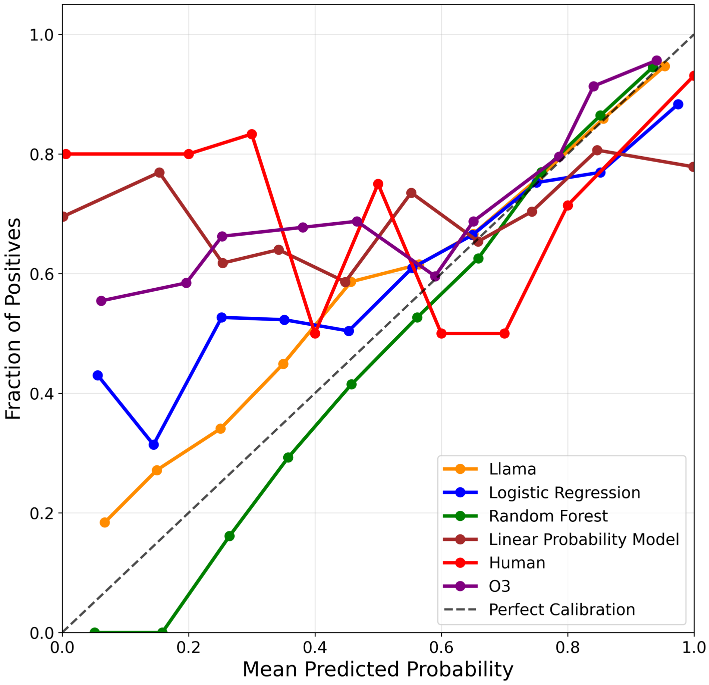

Figure C2. Model Calibration for Frisk Reasonable Suspicion Predictions

Notes: This figure illustrates the calibration of different models' predictions for reasonable suspicion in police frisks. Similar to Figure C1, each point represents a decile of model predictions, plotting average predicted probability against actual percentage of frisks with reasonable suspicion.

As the figures show, Llama and the classic ML techniques (random forest and logistic regression) were generally the best-calibrated; the LPM was generally poorly calibrated when the model deemed reasonable suspicion unlikely. o3 was poorly calibrated in that it consistently overestimated the likelihood that stops had reasonable suspicion, and the same for frisks (except at high predicted probabilities). The human codings were extremely poorly calibrated, with the human RA's predictions having only a loose relationship to empirical probabilities.

Appendix D: Ensemble Methods

We implemented a variety of ensemble methods to combine predictions from our base models (fine-tuned Llama 3, logistic regression, random forest, LPM, and OpenAI o3). The ensemble methods were evaluated on both stop and frisk reasonable suspicion predictions. In all cases, we trained initial models on the train set, determined ensemble hyperparameters using the validation set, and then tested ensemble performance using the test set. Below we describe the technical details of each ensemble approach.

Traditional Ensemble Methods

Simple Average Ensemble: This method computes the arithmetic mean of the predicted probabilities from all base models:

$$\hat{p}_{\text{avg}} = \frac{1}{M}\sum_{m=1}^{M}\hat{p}_m \tag{D1}$$

where M is the number of base models and $\hat{p}_m$ is the predicted probability from model m.

Weighted Average Ensemble: This approach assigns weights to each model based on their individual performance, specifically using the absolute correlation between each model's predictions and the target variable on the validation set:

$$\hat{p}_{\text{weighted}} = \sum_{m=1}^{M} w_m\hat{p}_m, \quad \text{where} \quad w_m = \frac{|\rho_m|}{\sum_{j=1}^{M}|\rho_j|} \tag{D2}$$

and ρ_m is the Pearson correlation coefficient between model m's predictions and the binary target.

Median Ensemble: Instead of averaging, this method takes the median of all model predictions, providing robustness against outlier predictions:

$$\hat{p}_{\text{median}} = \text{median}(\hat{p}_1, \hat{p}_2, ..., \hat{p}_M) \tag{D3}$$

Stacking: This meta-learning approach trains a logistic regression model to optimally combine base model predictions:

$$\hat{p}_{\text{stack}} = \sigma(\beta_0 + \sum_{m=1}^{M}\beta_m\hat{p}_m) \tag{D4}$$

where σ is the sigmoid function and $\beta_0, \beta_1, ..., \beta_M$ are learned parameters.

Neural Stacking with Text Integration

Our most sophisticated ensemble method is a neural stacking approach that combines both the base model predictions and the original police narrative text. This method uses a pre-trained DistilBERT model to encode the narrative text and combines these embeddings with the numerical predictions from base models.

The architecture consists of:

1) **Text Encoder:** A pre-trained DistilBERT model that converts police narratives into 768-dimensional embeddings. We use the [CLS] token representation as the text embedding.

2) **Feature Combination:** The text embedding is concatenated with the M base model predictions to create a combined feature vector of dimension $768 + M$.

3) **Classification Head:** A multi-layer perceptron (MLP) with one hidden layer processes the combined features:

$$\mathbf{h} = \text{ReLU}(\mathbf{W}_1[\mathbf{e}_{[\text{CLS}]}; \hat{\mathbf{p}}] + \mathbf{b}_1) \tag{D5}$$

$$\hat{y} = \sigma(\mathbf{w}_2^T \mathbf{h} + b_2) \tag{D6}$$

where $\mathbf{e}_{[\text{CLS}]}$ is the text embedding, $\hat{\mathbf{p}} = [\hat{p}_1, ..., \hat{p}_M]^T$ are the base model predictions, and $\mathbf{W}_1, \mathbf{b}_1, \mathbf{w}_2, b_2$ are learned parameters.

The model is trained using binary cross-entropy loss with different learning rates for the DistilBERT encoder (2e-5) and the MLP classifier (1e-3) to account for the pre-trained nature of the text encoder.

Implementation Details

All ensemble methods were implemented using the validation set for training/weight determination and evaluated on the held-out test set. For methods requiring hyperparameter tuning (stacking and neural stacking), we used 5-fold cross-validation on the validation set. The neural stacking model was trained for 3 epochs with a batch size of 16 using the AdamW optimizer.

Despite the theoretical advantages of ensemble methods, our results show that they provide only marginal improvements over the fine-tuned Llama 3 model alone. The best ensemble method (stacking) improved accuracy by only 0.2% and F1 score by 0.2% for stop predictions, and 1.4% and 1.4% respectively for frisk predictions compared to Llama 3. This suggests that the fine-tuned language model already captures most of the predictive signal in the police narratives, leaving little room for improvement through model combination.

Ensemble Method Performance Results

Tables D1 and D2 present the performance of our ensemble methods. The ensemble models were trained using all available base model predictions. As shown in Table D1, all ensemble methods achieved similar performance levels, with accuracies ranging from 85.4% to 88.0% for stop predictions and 74.8% to 78.4% for frisk predictions on the test set.

Table D1—Performance of All Ensemble Methods

	Stop Predictions			**Frisk Predictions**		
Ensemble Method	**Accuracy**	**F1**	**AUC**	**Accuracy**	**F1**	**AUC**
Test Set Results (with available models)						
Simple Average	0.872	0.925	0.894	0.777	0.854	0.806
Weighted Average	0.874	0.925	0.898	0.784	0.860	0.811
Median Voting	0.870	0.924	0.891	0.777	0.856	0.808
Stacking	0.880	0.928	0.902	0.783	0.861	0.814
Neural Stacking	0.854	0.913	0.846	0.748	0.839	0.723
Validation Set Results						
Simple Average (Val)	0.880	0.929	0.895	0.788	0.865	0.791
Weighted Average (Val)	0.883	0.931	0.898	0.792	0.869	0.795
Median Voting (Val)	0.876	0.927	0.892	0.787	0.867	0.792
Stacking (Val)	0.882	0.929	0.901	0.790	0.870	0.793
Neural Stacking (Val)	0.850	0.912	0.831	0.747	0.850	0.761

Notes: Test set results use all models available in the test data. Validation set results include O3 predictions when available. The specific models included in each ensemble depend on data availability.

Table D2 provides a direct comparison between the fine-tuned Llama 3 model and the best-performing ensemble method. The results demonstrate that while the ensemble approach achieves slightly higher performance metrics, the improvements are marginal.

TABLE D2—COMPARISON OF FINE-TUNED LLAMA 3 VS. BEST ENSEMBLE METHOD

	Accuracy		**F1 Score**		**AUC-ROC**	
Model	**Stop**	**Frisk**	**Stop**	**Frisk**	**Stop**	**Frisk**
Fine-tuned Llama 3	0.878	0.769	0.926	0.847	0.901	0.793
Best Ensemble (Stacking)	0.880	0.783	0.928	0.861	0.902	0.814
Improvement	+0.002	+0.014	+0.002	+0.014	+0.001	+0.021

Notes: Improvement row shows the absolute percentage point difference between the best ensemble and fine-tuned Llama 3.

These results suggest that the fine-tuned Llama 3 model already captures most of the predictive signal in the police narratives. The marginal improvements from ensemble methods do not justify the additional computational complexity and reduced interpretability that comes with combining multiple models. This finding aligns with work showing that large language models can achieve near-optimal performance on text classification tasks when properly fine-tuned (Hansen and Salamon, 1990; Polikar, 2006).

Appendix E: Topic Modeling Methodology

Our topic modeling approach was designed to extract semantically meaningful and generalizable topics from police narratives while handling the scale and complexity of the dataset. This approach was inspired by TopicGPT (Pham et al., 2023), which uses LLMs to identify topics in text corpora. However, our implementation differs substantively in architecture, especially in our hierarchical deduplication process. TopicGPT has been shown to outperform classic topic modeling techniques including Latent Dirichlet Allocation (LDA), BERTopic, and SeededLDA in producing coherent and interpretable topics (Pham et al., 2023). Below we describe the technical details of our implementation.

Topic Extraction Process

We employed OpenAI's o3 model to process all police narratives using the following prompt:

> You are being given the text of a police report. Your task is to identify generalizable topics within the police report that reflect the rationale or justification for the stop. Output the existing topics as identified in the police report, along with exact quotation(s) from the police report corresponding to those topics. The quotations should reflect the content in the report related to each topic, so that if the quotations were masked the topic would be removed from the report.
>
> [Examples]
>
> Example 1: Identifying "* Walking on Highway and * Known Warrant"
>
> Police Report:
>
> POLICE OBSERVED MALE WALKING ON THE HIGHWAY AT LISTED LOCATION. POLICE HAD PRIOR KNOWLEDGE THAT MALE HAD AN OPEN WARRANT FOR ASSAULT. REFER TO DC#16-02-044212. MALE WAS PLACED INTO CUSTODY AND TRANSPORTED TO SVU.
>
> Your response:
>
> - Walking on Highway: "MALE WALKING ON THE HIGHWAY"
> - Known Warrant: "POLICE HAD PRIOR KNOWLEDGE THAT MALE HAD AN OPEN WARRANT FOR ASSAULT"
>
> [Instructions]
>
> 1) Determine topics mentioned in the police report.
>   - The topic labels must be as GENERALIZABLE as possible. They must not be report-specific.
>   - Each topic label must reflect a SINGLE topic instead of a combination of topics.
>   - The new topics must be preceded by *, have a short general label, and have exact quotations that if masked would entirely remove that topic from the report. You may include multiple quotations for a single topic, separated by a comma.
> 2) Perform ONE of the following operations:
>   a) If there are topics in the police report, output the topics.

b) If the police report contains no topic, return “None”.

This prompt design ensures that topics are generalizable across reports, focused on single concepts, and accompanied by exact quotations that enable precise interpretation. To handle the large volume of reports efficiently, we used OpenAI’s Batch API, which processes multiple requests asynchronously while maintaining consistency in the extraction approach.

Topic Deduplication Algorithm

After initial extraction, we identified substantial redundancy in topic labels due to minor variations in phrasing. Our deduplication algorithm maps similar topics to canonical forms through the following process:

1) **Similarity scoring:** We compute embeddings for all topics using OpenAI’s text-embedding-3-large model (3072 dimensions) and calculate cosine similarity between topic pairs. Additionally, we employ BM25 scoring to capture lexical similarity. Topics are then processed in batches of up to 20, where each batch is constructed by selecting an anchor topic and including similar topics based on a 50/50 combination of embedding similarity and BM25 scores.

2) **Canonical selection:** Each batch of similar topics is processed by the o3 model using the following prompt:

> You are a data cleansing assistant. You will be given a list of topic labels from police reports. Your task is to analyze these labels and do one of two things:
>
> a) If multiple labels represent the same or similar concepts, map them all to a single canonical label (choose the most representative one or create a new standardized name)
>
> b) If a label is unique and doesn’t belong with others, map it to itself (the canonical label is the same as the original)
>
> IMPORTANT: For topics to be grouped together, ALL semantic elements present in one topic must also be present in the other topics being grouped with it. The canonical label must contain ALL the semantic elements that are common to all topics being mapped to it. Do not map topics together if one contains elements that the others do not share.
>
> Be conservative - only group topics that are clearly referring to the same concept, in the same direction. For example, ’Warrant for Arrest’ and ’No Warrant for Arrest’ should not be grouped together. When in doubt, keep topics separate by mapping them to themselves.
>
> Only respond in valid JSON format with one key: ’canonical_mapping’. The ’canonical_mapping’ should map each original label to its canonical label. Do not include any explanation or extra text, only provide the JSON.

The model returns a JSON mapping that assigns each topic in the batch to its canonical form. This process is repeated for all topics in the current iteration. After processing all batches, the algorithm performs transitive closure to update all mappings. That is, if topic A maps to B in one iteration and B maps to C in another, then A ultimately maps to C. After each iteration and its accompanying transitive closure, the model begins a new iteration with the resulting canonical topics. The iterative process continues until convergence is reached, defined as when at least 75% of batches produce no changes.

Main Reason Identification Methodology

After extracting and deduplicating topics from police narratives, we implemented a systematic approach to identify which topic serves as the primary justification—the "main reason"—for each stop. This analysis enables us to simulate targeted policy interventions by calculating the impact of discontinuing stops based on specific main reasons.

Main Reason Identification Prompt

We employed OpenAI's o3 model to identify the main reason for each stop using the following structured prompt:

System Prompt:

> You are analyzing a police stop report to identify which of a list of topics represents the MAIN REASON for the stop.
>
> You will be given:
>
> 1) The full police report text
> 2) A list of topics identified in this report
>
> Your task is to determine which ONE of these topics best represents the PRIMARY reason the stop was initiated. Consider:
>
> - The chronological order of events in the report
> - What triggered the initial police contact
> - The most serious or significant reason if multiple exist
>
> Output ONLY the exact topic that represents the main reason, or "NONE" if you cannot determine a clear main reason from the given topics.

Distribution of Main Reasons for Stops and Frisks

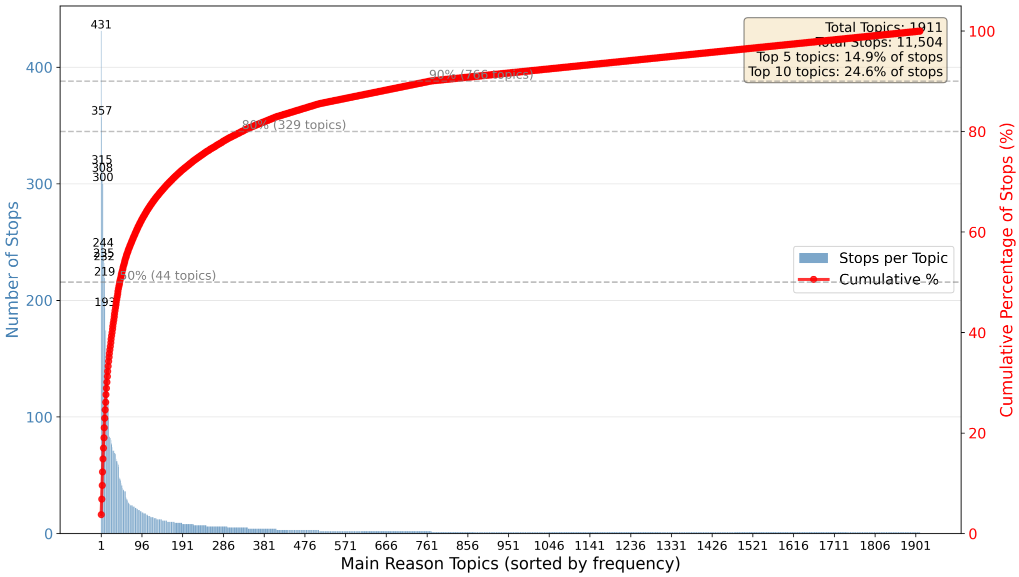

Figure E1. Distribution of Police Stops by Main Reason Topic

Notes: This figure displays the distribution of police stops in the test set across canonical topics identified as main reasons. The blue bars represent the number of stops for each topic, sorted in descending order by frequency. The red line shows the cumulative percentage of stops, demonstrating the distribution's long tail. The distribution follows an approximately exponential pattern based on its probability density function (PDF) and cumulative distribution function (CDF).

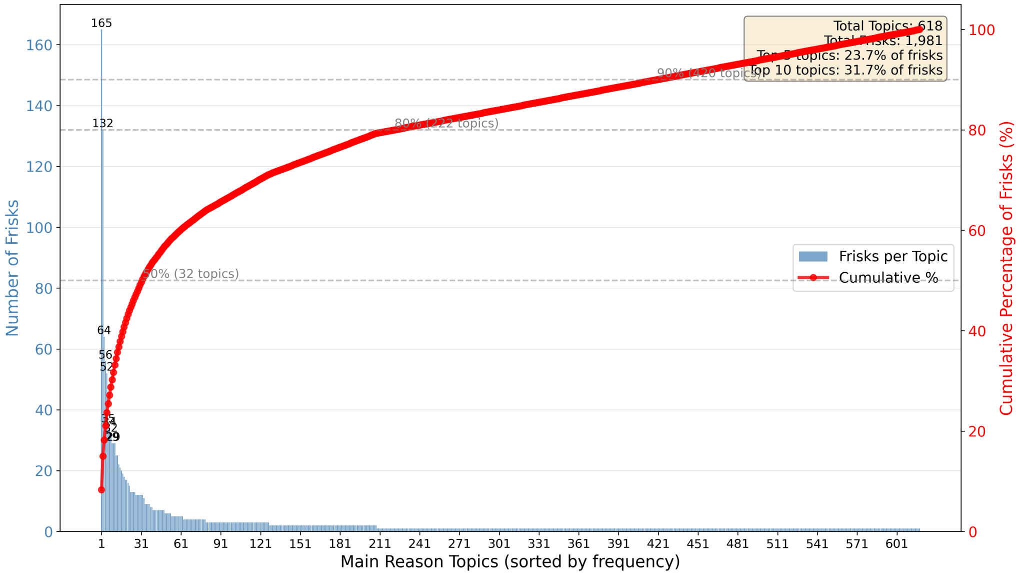

Figure E2. Distribution of Police Frisks by Main Reason Topic

Notes: This figure displays the distribution of police frisks in the test set across canonical topics identified as main reasons. The blue bars show the number of frisks for each topic (sorted by frequency), while the red line shows the cumulative percentage of frisks. Like the stops distribution, this follows an approximately exponential pattern, indicating that a small number of topics account for a disproportionate share of frisks.

E1. Decision Rule Formulas

Formally, let S denote the set of all stops, and let $S_i \subseteq S$ represent the subset of stops where topic i serves as the main reason. For each stop $s \in S$, let $y_s = 0$ if the stop lacks reasonable suspicion and $y_s = 1$ if reasonable suspicion is present. The topic-specific false positive rate for topic i is defined as:

$$\text{(E1)} \qquad \text{FPR}_i = \frac{\sum_{s \in S_i} \mathbb{I}(y_s = 0)}{|S_i|}$$

where $\mathbb{I}(\cdot)$ is the indicator function and $|S_i|$ denotes the cardinality of set S_i. The overall false positive rate across all stops is:

$$\text{(E2)} \qquad \text{FPR} = \frac{\sum_{s \in S} \mathbb{I}(y_s = 0)}{|S|}$$

To evaluate progressive policy interventions involving multiple topics, we employ a greedy optimization algorithm. This algorithm iteratively selects topics to remove based on which removal would minimize the false positive rate among the remaining stops at each step.

Let $\mathcal{T} = \{1, 2, ..., K\}$ denote the set of all canonical topics. The greedy algorithm proceeds as follows:

1) Initialize the set of available topics $\mathcal{A}_0 = \mathcal{T}$ and the set of removed topics $\mathcal{R}_0 = \emptyset$.

2) At each iteration m, identify the topic i_m^* that minimizes the FPR among remaining stops:

$$\text{(E3)} \qquad i_m^* = \arg\min_{i \in \mathcal{A}_{m-1}} \frac{\sum_{s \in S \setminus \bigcup_{j \in \mathcal{R}_{m-1} \cup \{i\}} S_j} \mathbb{I}(y_s = 0)}{|S| - |\bigcup_{j \in \mathcal{R}_{m-1} \cup \{i\}} S_j|}$$

3) Update the sets: $\mathcal{R}_m = \mathcal{R}_{m-1} \cup \{i_m^*\}$ and $\mathcal{A}_m = \mathcal{A}_{m-1} \setminus \{i_m^*\}$.

4) Continue until all topics are removed or no stops remain.

This greedy approach ensures monotonically decreasing false positive rates as topics are progressively removed. After m iterations, the cumulative false positive rate is:

$$\text{(E4)} \qquad \text{FPR}_{-\mathcal{R}_m} = \frac{\sum_{s \in S \setminus \bigcup_{j \in \mathcal{R}_m} S_j} \mathbb{I}(y_s = 0)}{|S| - |\bigcup_{j \in \mathcal{R}_m} S_j|}$$

The proportion of stops discontinued under this policy is:

$$\text{Proportion Discontinued}_m = \frac{|\bigcup_{j\in\mathcal{R}_m} S_j|}{|S|} \tag{E5}$$

This framework generates a policy frontier showing the tradeoff between false positive rate reduction ($\text{FPR} - \text{FPR}_{-\mathcal{R}_m}$) and the proportion of stops discontinued.

Cumulative Impact of Discontinuing Stops/Frisks by Main Reason Topics

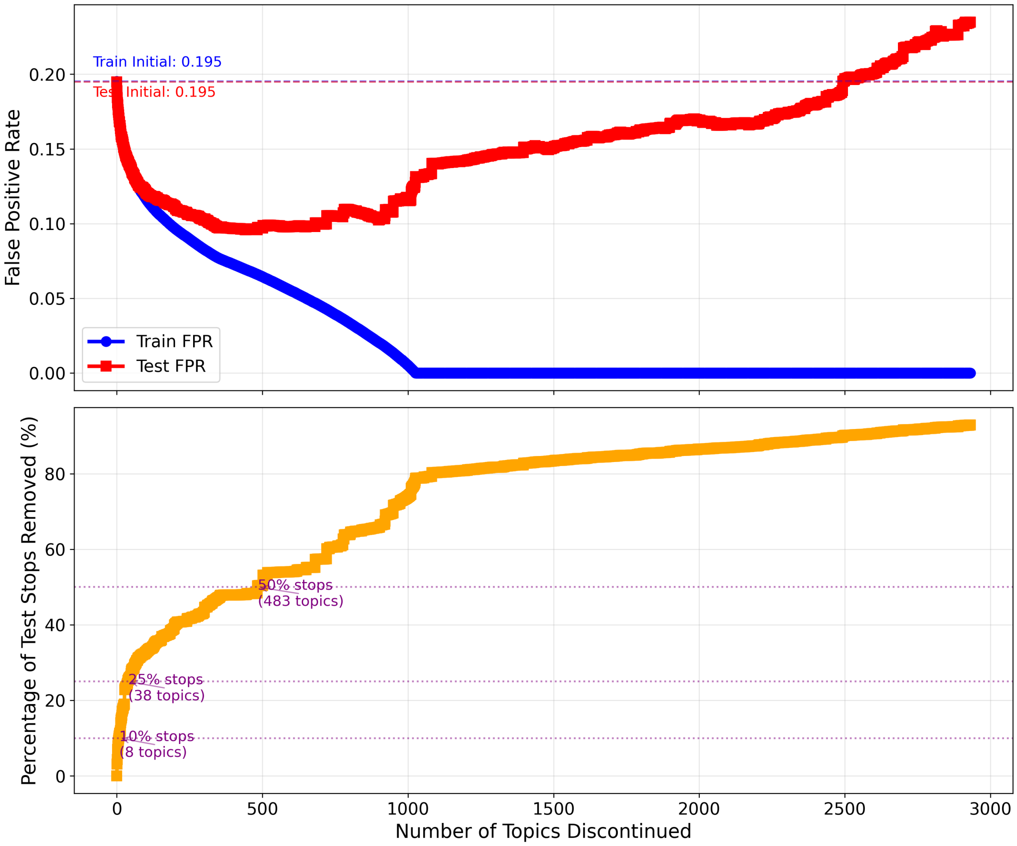

Figure E3. Cumulative Impact of Discontinuing Stops Based on Problematic Main Reasons

Notes: This figure illustrates the cumulative effect of progressively discontinuing stops based on the main reason for each stop, ranked by how much each removal would reduce the aggregate false positive rate of the remaining stops (largest reductions first) in the train set. The x-axis then shows the percentage of total stops in the test set that would be eliminated by discontinuing stops based on an increasing number of problematic topics. The y-axis shows the resulting false positive rate among the remaining stops in the test set, both for the train set (blue) and the test set (red). Because the decision rules are formulated using the train set but then evaluated on the test set, only stops with at least one main reason shared between the train set and the test set were evaluated. The figure assumes perfect officer compliance with decision rules.

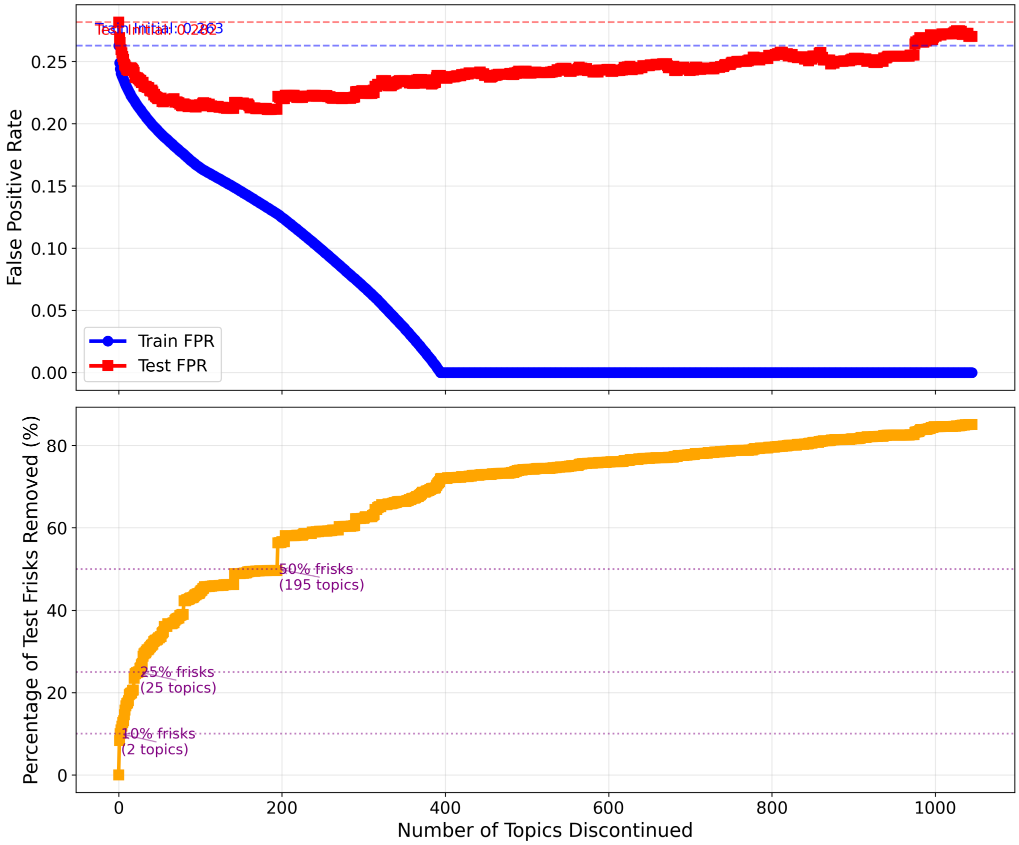

Figure E4. Cumulative Impact of Discontinuing Frisks Based on Problematic Main Reasons

Notes: This figure illustrates the cumulative effect of progressively discontinuing frisks based on the main reason for each frisk, ranked by how much each removal would reduce the aggregate false positive rate of the remaining frisks (largest reductions first) in the train set. The x-axis then shows the percentage of total frisks in the test set that would be eliminated by discontinuing frisks based on an increasing number of problematic topics. The y-axis shows the resulting false positive rate among the remaining frisks in the test set, both for the train set (blue) and the test set (red). Because the decision rules are formulated using the train set but then evaluated on the test set, only frisks with at least one main reason shared between the train set and the test set were evaluated. The figure assumes perfect officer compliance with decision rules.

Appendix F: Topic Modeling and Guidelines for Frisks

To evaluate potential policy interventions for frisks, we simulated the impact of instructing officers not to conduct frisks when certain topics constitute the main reason. We ranked topics by their efficiency in reducing false positives—that is, how many unconstitutional frisks would be prevented relative to the total number of frisks forgone. As with stops, we assume perfect officer compliance with the policy.

Table F1 presents the cumulative impact of progressively discontinuing frisks based on the most problematic main reasons. The results demonstrate that discontinuing frisks based on the single most problematic main reason topic would reduce the false positive rate from 28.3% to 26.9% while eliminating 8.3% of all frisks. When extending this policy to the top five main reasons, the false positive rate decreases to 25.4% while eliminating 12.8% of frisks.

Table F1—Cumulative Impact of Discontinuing Frisks by Main Reason Topics

Number of Topics Removed	**False Positive Rate**	**Percentage of Frisks Removed**
0	0.282	0.0%
1	0.269	8.3%
2	0.265	10.0%
3	0.260	11.0%
4	0.257	11.8%
5	0.254	12.8%
10	0.244	17.4%
15	0.245	19.8%
20	0.237	24.5%

Notes: This table shows the cumulative effect of removing frisks where progressively more topics serve as the main reason, ranked by their efficiency in reducing false positive rates. The false positive rate represents the proportion of remaining frisks that lack reasonable suspicion. The analysis assumes perfect officer compliance with policy interventions.

Figure E4 visualizes these results, showing how the false positive rate decreases as more topics are removed as valid reasons for frisks, plotted against the percentage of total frisks that would be affected by such policies.

Table F2 presents the five most effective topics to target for policy intervention—those that would reduce false positives the most.

Table F2—Top Five Topics for Reducing False Positive Rates in Police Frisks

Topic	Topic FPR	FPR Reduction	% of Frisks
Person with Gun Call	4.18%	4.40%	8.33%
Execution of Search Warrant	8.33%	1.80%	0.91%
Odor of Marijuana	47.1%	1.17%	1.72%
Suspicious Bulge	52.4%	0.92%	1.06%
Suspicious Conduct/Behavior	52.6%	0.85%	0.96%

Notes: Topics are ranked by their efficiency in reducing false positive rates—the reduction in unconstitutional frisks relative to the total proportion of frisks affected. Topic FPR represents the false positive rate within each topic category. FPR Reduction shows the relative percentage decrease in the overall false positive rate if frisks based on this topic were discontinued. The analysis assumes perfect officer compliance with policy interventions.

The distribution of main reasons for frisks reveals important patterns in police frisk justifications, as shown in Figure E2 in Appendix E. Similar to stops, the distribution follows an approximately exponential pattern.

These findings are further illustrated by comparing our decision rule approach with an alternative method using the Llama model's predicted probabilities directly. Figure F1 shows the false positive rate achieved when retaining different percentages of frisks, comparing two approaches: (1) our decision rules based on removing frisks with specific main reasons, and (2) ranking frisks by the Llama model's predicted probability of reasonable suspicion and removing those with the lowest probabilities. The blue line represents the false positive rate when implementing decision rules to discontinue frisks, where the decision rules are determined using the train set and implemented on the test set. This represents a realistic, cross-validated application of decision rules. The green line represents the false positive rate when decision rules are determined using the test set and implemented on the test set. While not realistic, this represents the theoretical optimal performance from a set of decision rules.

The comparison reveals that the Llama model consistently outperforms the decision rule approach for frisk predictions for the realistic set of decision rules (blue line), similar to the pattern observed for stops. The Llama model also dominates the theoretically optimal set of decision rules up to a point (when around 20% of frisks has been discontinued). Table F3 quantifies this difference at key retention levels.

Table F3—Performance Comparison of Decision Rules vs. Llama Model for Frisks at Key Retention Percentages

Frisks Retained (%)	**Decision Rules FPR**	**Llama Model FPR**	**Difference**
90	0.219	0.201	+0.019
80	0.220	0.156	+0.064
70	0.215	0.151	+0.065
60	0.201	0.138	+0.063
50	0.195	0.113	+0.083

Notes: This comparison was generated by ranking all frisks either by their main reason topic (for decision rules) or by their Llama-predicted probability of reasonable suspicion. The decision rule FPRs reflect results when decision rules are determined based on the train set and evaluated on the test set. The table shows false positive rates (FPR) at different retention levels. The "Difference" column shows the decision rules FPR minus the Llama model FPR; positive values indicate the Llama model achieves lower false positive rates.

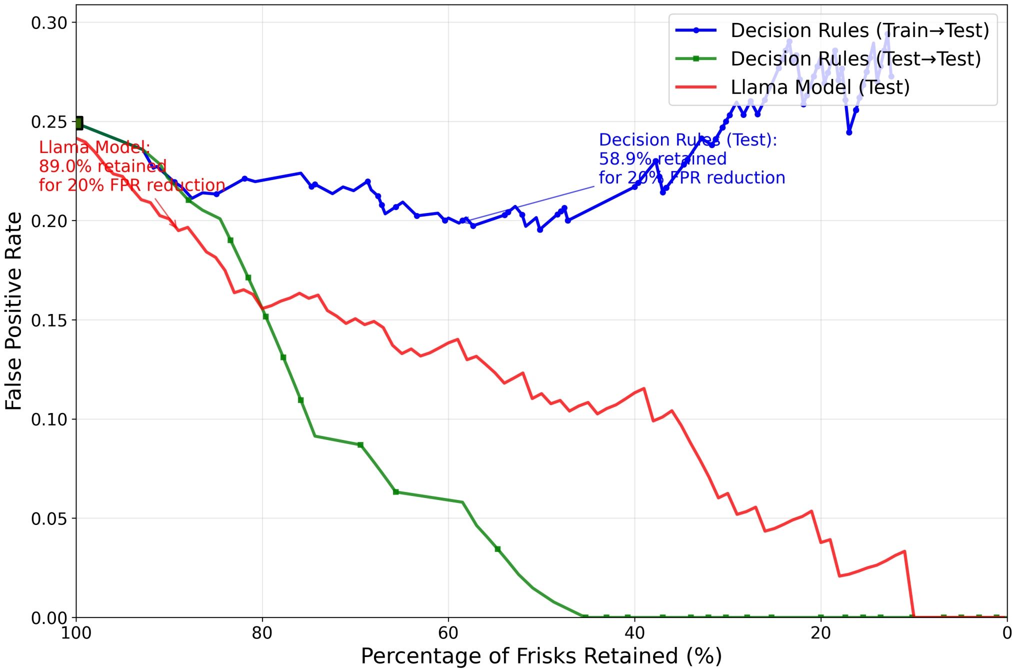

Figure F1. Comparison of False Positive Rate Reduction Strategies for Frisks

Notes: This figure compares two approaches for reducing false positive rates in police frisks. The x-axis shows the percentage of frisks retained, while the y-axis shows the resulting false positive rate. The red line shows the performance of the Llama model, if we remove frisks with the lowest predicted probabilities of reasonable suspicion (having calibrated the Llama model using the train set and then tested it using the test set). The blue and green lines represent a decision rule approach that progressively discontinues frisks based on problematic main reason topics, ranked by their ability to reduce false positives. The blue line reflects the false positive rate when decision rules are determined based on the train set, and the green line represents the false positive rate when the decision rules are determined based on the test set. The green line therefore represents the theoretical optimal performance of a decision-rule approach, whereas the blue line represents a realistic implementation. The decision rules do not monotonically decrease the false positive rate for the blue line due to overfitting; certain rare justifications for frisks have a high false positive rate in the train set and seem promising to remove but turn out to have a lower false positive rate in the test set. To ensure a fair comparison, we only include frisks in this graph where there was a main reason found in the train set as well as the test set. Because frisks had more diverse main reasons than stops as well as a much smaller starting sample size, the N for this graph was much lower, which resulted in the jaggedness in the figure. If we had included all frisks in the test set, the red line would have been smooth and monotonic.

This comparison was generated using the same methodology as for stops. For the decision rule approach, frisks were progressively removed based on the ranking of main reason topics by their efficiency in reducing false positives. For the Llama model approach, frisks were ranked by their predicted probability of having reasonable suspicion, and those with the lowest probabilities were progressively removed. At each retention level, we calculated the false positive rate among the remaining frisks.

The superior performance of the Llama model approach for frisks mirrors the results for stops, suggesting that while simple decision rules based on main reasons can effectively reduce unconstitutional frisks, the neural network captures more nuanced patterns that enable even better discrimination between lawful and unlawful frisks. The performance gap is slightly larger for frisks than for stops, which may reflect the additional complexity of frisk determinations requiring both reasonable suspicion and indications that a suspect is armed and dangerous. As with stops, however, the decision rule approach offers advantages in terms of interpretability and ease of implementation—officers can be given clear, actionable guidance about which types of frisks to avoid.