GIBSON DUNN

Gibson, Dunn & Crutcher LLP

200 Park Avenue
New York, NY 10166-0193
Tel 212.351.4000
www.gibsondunn.com

Mylan L. Denerstein
Direct: +1 212.351.3850
Fax: +1 212.351.6350
MDenerstein@gibsondunn.com

April 29, 2026

VIA ECF

Honorable Analisa Torres
United States District Judge
United States District Court
Southern District of New York
500 Pearl Street
New York, NY 10007-1312

Re: *Floyd, et al. v. City of New York*, 08-CV-1034 (AT),
Davis, et al. v. City of New York, et al., 10-CV-0699 (AT),
Stanford Study of NYPD Stop and Frisk Practices

Dear Judge Torres:

In this Court's February 12, 2021, order, *Floyd v. City of New York*, 08-CV-1034, ECF No. 817, the Court approved studies to examine compliance with applicable legal requirements in police-civilian encounters, racial disparities in officers' compliance in those encounters, and whether those encounters are appropriately documented. One study was undertaken by researchers affiliated with Stanford University. I am pleased to submit the report on the study (the "Stanford Study") conducted by those researchers.

The study examined police encounters recorded by body-worn cameras ("BWCs") in 2022, 2023 and 2024. The study used Artificial Intelligence ("AI") tools and techniques—including machine learning and natural language processing—to analyze BWC recordings and identify key indicators of constitutional compliance in what NYPD officers say and how they say it during encounters with civilians.

The study found that machine learning models can distinguish low level encounters in which a civilian is free to leave from *Terry* stops in which they are detained, on the basis of the language used in the body camera footage. It also found racial disparities in the language used, such that the language used in low level encounters with Black and Hispanic civilians was more like language used in *Terry* stops compared to the language used in low level encounters with white or other race civilians. Use of these AI models could be helpful in examining a large volume of BWC videos to screen footage for human review in identifying *Terry* stops that were undocumented.

The Stanford Study also used computational tools to analyze the language NYPD officers used to obtain consent searches during *Terry* stops. The study found that NYPD officers seldom used the word "consent" and "search" in their consent search requests. Instead,

they often used more ambiguous language, using phrases such as "[can I] check" and "[do you] mind." What officers say—and how they say it—can bear directly on whether the consent provided was voluntary, including whether they clearly communicate the nature of the request and frame it as one the person may decline.

Respectfully,

Mylan L. Denerstein
Independent Monitor

What Can AI Tell Us About NYPD Street Stops?

Using Natural Language Processing and Machine Learning to Analyze Investigative Encounters and Consent Searches from Police Body-Worn Camera Footage

April 2026

Rob Voigt University of California, Davis | Stanford SPARQ
Nicholas P. Camp University of Michigan | Stanford SPARQ
Dan Sutton Stanford Law School | Stanford Center for Racial Justice
Jennifer L. Eberhardt Stanford Graduate School of Business | Stanford SPARQ

Authors

Rob Voigt

Assistant Professor of Linguistics, University of California, Davis;
Faculty Affiliate, Stanford SPARQ

Nicholas P. Camp

Assistant Professor of Organizational Studies, University of Michigan;
Faculty Affiliate, Stanford SPARQ

Dan Sutton

Director of Justice and Safety, Stanford Center for Racial Justice, Stanford Law School

Jennifer L. Eberhardt

Professor of Organizational Behavior and Psychology, Stanford Graduate School of Business;
Faculty Co-Director, Stanford SPARQ

Acknowledgements

We would like to thank Judge Analisa Torres of the Southern District of New York who authorized this study and the late Peter Zimroth who commissioned this work as the first court appointed Monitor of the New York Police Department as a result of the *Floyd* and *Ligon* cases. We acknowledge and appreciate them both for their support and for their belief, early on, that our work could make a difference at the NYPD and to the people of New York City. We also thank the current Monitor Mylan Denerstein, the Deputy Monitor Richard Jerome, and the entire Monitor team for continuing to see the work through and for embracing the promise of machine learning and natural language processing as powerful auditing tools. We would like to thank the Department (in particular, Daniel Gorayeb, Sara Gronningsater, Sgt. Michael Benedetto, and Sgt. Xochilt Chantel), Kathleen Doherty and Annie Chen of the City University of New York's Institute for State and Local Governance, John MacDonald and James McCabe from the Monitor team, and members of Axon (Nicholas Fash, Bradford Buonasera, Griffin Cohen) for their collaboration in obtaining and understanding the data underlying the analyses we report here. Finally, we would like to thank the people of the City of New York and those who represent them for their efforts to improve police-civilian encounters now and in the future.

Contents

I. EXECUTIVE SUMMARY

In 2013, the U.S. District Court for the Southern District of New York found that the New York Police Department's ("NYPD") stop-and-frisk practices violated the Fourth Amendment, which requires stops to be based on reasonable suspicion, and the Fourteenth Amendment through a pattern of racial profiling during stops. The Court appointed an Independent Monitor (the "Monitor") to oversee reforms to the Department's policies, training, supervision, and documentation of investigative encounters. Body-worn cameras ("BWC"), required by the Court's remedial orders, now provide an objective record of what officers actually say during encounters. This information is largely invisible in stop reports and other administrative records but is essential for assessing constitutional compliance.

In 2021, the Court approved two research studies designed to leverage this footage to evaluate whether the officers of the NYPD follow the legal rules that apply during police-civilian encounters, whether patterns of compliance vary by race, and assess whether encounters are documented accurately.[1] The City University of New York's ("CUNY") Institute for State and Local Governance ("ISLG") filed a report in May 2025 showing significant rates of unconstitutional stops and underreporting of encounters, based on legal determinations by retired New York State judges reviewing BWC footage and documents related to the encounters.[2] The current study, by a Stanford-affiliated research team, uses AI tools and techniques–including machine learning and natural language processing–to computationally analyze BWC recordings and identify key indicators of constitutional compliance in what NYPD officers say and how they say it during encounters with civilians.

The Stanford study focuses on areas where AI and language-based methods provide the clearest and most meaningful insights for compliance monitoring. First, the study evaluates the NYPD's compliance with New York's four-level framework for classifying police interactions in investigative encounters established by *People v. De Bour*, with a specific emphasis on Level 3 stops and detentions that require reasonable suspicion, mirroring the Fourth Amendment's requirements.[3] Second, the study analyzes the NYPD's consent search practices for compliance with the Fourth Amendment, which requires that consent be voluntary, an inquiry that turns in part on the language officers use.[4] To assess compliance with the Fourteenth Amendment, the

[1] *Floyd v. City of New York*, No. 1:08-cv-1034 (AT) (S.D.N.Y. Feb. 12, 2021), ECF No. 817.

[2] Kathleen Doherty & Annie Chen, An Examination of NYPD Stop and Frisk Practices: Using Body-worn Camera Recordings to Determine the Constitutionality and Documentation of Street Stops (CUNY Inst. for State & Loc. Governance Apr. 2025), filed in *Floyd v. City of New York*, No. 1:08-cv-01034 (AT) (S.D.N.Y. May 1, 2025), ECF No. 956-1 ("ISLG Report").

[3] *See People v. De Bour*, 40 N.Y.2d 210 (1976); *Terry v. Ohio*, 392 U.S. 1 (1968).

[4] *See Schneckloth v. Bustamonte*, 412 U.S. 218 (1973).

study analyzes differences in officer language during encounters with civilians of different races and ethnicities, from low-level interactions to stops and searches.

The study analyzes NYPD encounters from three primary sources. For understanding *De Bour* documentation and compliance, we analyze a sample of 1,702 encounters between March 16 and May 15, 2022, drawn from data that are the focus of the ISLG Report, and 1,156 encounters assessed by the Monitor team between 2022 and 2024.[5] For our consent search analysis, we examine the full set of consent requests documented on stop reports in 2023 which we were able to match to footage, totaling 1,770 encounters and 3,695 videos.[6]

Automatic speech recognition tools convert BWC audio from these encounters into text transcripts of officer and civilian speech. Natural language processing and machine learning models then examine these transcripts to identify patterns in officer language. To evaluate *De Bour* compliance, we trained machine learning models on encounters that legal experts–retired judges and Monitor team members–had already classified, teaching the models which words and phrases in an officer's speech distinguish Level 1 and 2 encounters from Level 3 stops.[7] To evaluate consent search practices, natural language processing tools searched transcripts for specific words and linguistic patterns, measuring how often officers use explicit terms like "consent" and "search" versus ambiguous phrasing such as "you don't mind if I take a look?".

A. Evaluating the NYPD's Compliance with New York's Four-Level *De Bour* Framework for Police Interactions in Investigative Encounters

NYPD officers are required to report every stop on a stop form. The ISLG Report found that for encounters labeled by officers as stops in the BWC data, a stop report was not completed for approximately one in four persons stopped. In the sample of encounters that officers categorized as a low-level investigative encounter (i.e., did not include a stop, arrest, or summons), the report found that officers stopped at least one individual during 3% of encounters.[8] The Stanford study applied computational techniques to identify the linguistic features that distinguish low-level encounters from stops and to detect potential underreporting across the large volume of NYPD recordings. The study found that:

- Machine learning models can distinguish Level 1 and Level 2 encounters—in which a civilian is free to leave—from Level 3 stops—in which they are detained—on the basis of language appearing in the corresponding body camera footage. We establish balanced

[5] For a detailed explanation of the encounters that are the focus of the ISLG Report, *see* ISLG Report at 8-12.

[6] The consent search sample includes only encounters in which officers documented that they requested consent to search. Officers conducting a Level 3 stop based on reasonable suspicion that a subject is armed and dangerous may conduct a protective frisk without obtaining consent.

[7] The classifications used as training data reflect the Monitor team's assessments. The NYPD and Monitor team occasionally differ on individual encounter classifications, though agreement rates are high.

[8] ISLG Report at 20.

predictive tasks where random guessing would achieve 50% accuracy, and evaluate models on three tasks related to documentation and one related to constitutional compliance:

  - **Officer Documentation Classification:** separating stops from encounters in properly documented data from the ISLG study (achieved 91.0% accuracy)

  - **Undocumented Stop Classification:** separating true low-level encounters from unreported stops as assessed by the Monitor team (achieved 72.9% accuracy)

  - ***De Bour* Classification:** separating stops from encounters in all available sampled data (achieved 81.7% accuracy)

  - **Stop Compliance Classification:** separating constitutionally non-compliant from compliant stops as assessed by expert judges in the ISLG study (achieved 71.6% accuracy)

- These models produce well-calibrated probabilistic estimates that could be applied to audits. We show that such models could be used to select potentially problematic footage for human review and substantially raise identification rates of undocumented stops. For example, if the Monitor were to take the same sample of interactions to be audited but review them in order of the estimated model-based probability that a low-level encounter is a stop, they would find the problematic cases more quickly. In the Monitor's quarterly audit sample, a review of the top quarter of cases most likely to be stops under the models presented here would have yielded over half of the cases that were ultimately identified. Application of these models to a broader spectrum of BWC footage recorded as low-level encounters would likely uncover large numbers of unreported stops.

- Across both datasets, results indicate significant racial disparities: relative to White and Other-race civilians, Level 1 and Level 2 encounters involving Black and Hispanic civilians are linguistically more similar to Level 3 stops, receiving stop probability estimates 5-11 percentage points higher than comparable encounters involving White/Other-race civilians, even after controlling for a range of factors, such as local crime rate. These gaps are present as well in Level 3 and Level 4 stops of Black civilians, which include language more characteristic of stops, relative to Level 3/Level 4 stops of Hispanic, White, and Other-race civilians.[9] Applying an analogous process to estimate constitutional compliance, we find that unreported stops in particular show the largest racial disparities in language use, such that relative to properly documented stops,

[9] Level 4 encounters involve an arrest or the issuance of a summons, requiring probable cause.

unreported stops of Black and Hispanic civilians display substantially more linguistic characteristics of constitutional non-compliance.

- Analysis of the language associated with stops identifies several important features within categories of *De Bour* Level 3 stops. Properly documented stops are more likely to mention "the reason" for the stop and to explicitly state whether or not civilians are "free to go." Unreported stops are more likely to include civilian pushback and officer use of indirect or implicit search-associated language than reported stops.

- AI models like those used in this study generally improve when trained on larger amounts of higher-quality data. As additional encounters are reviewed for *De Bour* compliance—by the Monitor, legal experts, or NYPD personnel—those reviews can be used to further refine the model and improve its predictive power. The current model is also constrained by the accuracy of Axon's automatic transcription. More accurate speech-to-text processing can close the gap between human-prepared transcripts and machine-generated transcripts. Accordingly, the results described in this report should be understood as a lower bound on performance, rather than a ceiling.

B. Analyzing the NYPD's consent search practices for compliance with the Fourth Amendment

Whether a person's consent to a search is voluntary under the Fourth Amendment can depend in significant part on what officers say, including whether they clearly communicate the nature of the request and whether their words indicate that the person has a meaningful opportunity to refuse.[10] The Stanford study used computational tools to analyze the language officers use to obtain consent searches across a full year of documented searches, measuring the prevalence of explicit consent language and the clarity of officers' requests. The study found that:

- The prevalence of key words and phrases associated with the explicit language of consent is consistently low: the word "search" appears in only approximately 46.0% of consent search stops, "consent" in 12.7%, and confirmatory questions like "Do you understand?" in 20.8%. Even under highly conservative assumptions accounting for transcription error, the data suggest that at least 75% of consent search interactions do not contain the word "consent" and at least 40% do not contain the word "search" in any

[10] Under *Schneckloth*, voluntariness is assessed under the totality of the circumstances, including factors such as the number of officers present, whether a person's path is blocked, the display of weapons, and whether the person is effectively in custody. *See also Florida v. Royer,* 460 U.S. 491 (1983); *United States v. Drayton*, 536 U.S. 194 (2002). In practice, the presence of these coercive physical and situational factors tends to weigh against voluntariness, while their absence is often treated as part of the overall context supporting a finding of consent. The analysis in this study examines encounters in which NYPD officers documented requesting consent to search on the stop report. In that context, the inquiry into whether consent was voluntary can be substantially linguistic: whether officers communicated the nature of the request in a manner consistent with voluntary agreement, and whether they framed it as a request the person could refuse.

associated video. Interactions in which all three key terms appear—consistent with the phrasing set out in the NYPD Patrol Guide—occur in only 3.2% of interactions (1.0% if analysis is limited to speech predicted to be spoken by officers alone).

- Other types of indirect and implicit requests to search are common and they are often phrased in ways that create ambiguity as to the meaning of simple responses like "yes" and "no." These include, in particular, the use of requests phrased as "[can I] check" (36.7% of interactions) and "[do you] mind" (16.8% of interactions). We argue that requests involving "check" inappropriately minimize the imposition of a search and create ambiguity as to what a "check" entails. The use of "[do you] mind" to phrase requests to search directly conflicts with the stated policy intention in the NYPD Patrol Guide that officers should "ask for consent to search in a manner that elicits a clear 'yes' or 'no' response."

- Our ability to detect race-based differences is limited by the relatively small number of consent searches involving White and Other-race individuals, but we do identify two significant disparities. First, Black civilians hear "[do you] mind" requests in 20.6% of interactions, while this is true for 15.3% of interactions with Hispanic, White, and Other-race individuals. Second, we measure the use of commands in consent search interactions as a contributing contextual factor to voluntariness and find that such commands occur approximately 20% more frequently in consent search interactions with Black and Hispanic civilians.

These findings have direct implications for the Fourth and Fourteenth Amendment standards at the core of the *Floyd v. City of New York* remedial framework.[11] The finding that documented Level 1 and Level 2 encounters of Black and Hispanic individuals are linguistically more similar to Level 3 stops suggests that some encounters involved conduct more consistent with detention, even though officers' documentation characterizes them as lower-level interactions. The prevalence of implicit requests and ambiguous phrasing in consent searches raises questions about whether individuals understand they are being asked to consent, that they can refuse, and what they are consenting to. Courts regularly weigh these factors in assessing voluntariness under the Fourth Amendment.[12] The study also reveals racial disparities in officer language in consent searches including greater use of ambiguous "[do you] mind" questions in encounters with Black civilians and more commands in encounters with Black and Hispanic civilians. While not establishing discriminatory intent, these patterns are consistent with differential treatment implicating the Fourteenth Amendment's Equal Protection Clause.

This report's findings suggest practical applications across several areas of NYPD operations, including officer training, supervisory tools, Early Intervention Program ("EIP") thresholds, and

[11] *Floyd v. City of New York*, 959 F. Supp. 2d 668, 676-78 (S.D.N.Y. 2013) ("*Floyd* Remedial Order").

[12] *See Schneckloth*, 412 U.S. 218, 227; *United States v. Drayton*, 536 U.S. 194, 206–07 (2002); *Florida v. Jimeno*, 500 U.S. 248, 251 (1991).

policy development. The model's ability to assign probability scores could allow supervisors to prioritize the review of BWC videos most likely to contain compliance issues rather than relying solely on random sampling and could be integrated into ComplianceStat sessions in real time, which focus on the same issues these models identify. More broadly, this work demonstrates the potential for integrating AI-powered analytical capabilities into existing Department systems in ways that help reviewers find potential problems faster and more consistently, while preserving human oversight at every step.

This study demonstrates that key indicators of constitutional compliance can be analyzed computationally across a large number of encounters. With access to more recent and comprehensive data, these methods could support the Court, the Monitor, and the Department in evaluating the impact of ongoing reform efforts. The Stanford team has applied this approach in two large cities in California, analyzing over 1.3 million videos representing more than 300,000 hours of interactions—demonstrating the potential for computational analysis to support compliance assessment at department-wide scale.

American police departments record millions of body-worn camera videos each year, but even well-resourced review efforts can examine only a small fraction of them. Important questions about how police interact with their communities remain difficult to answer. Advances in AI, paired with appropriate safeguards and collaboration among researchers, departments, and those responsible for oversight, make it possible to analyze this footage in ways that move beyond limited audits toward more comprehensive, efficient, and systematic assessments. This creates opportunities to identify problems earlier, measure whether reforms are working in practice, and build the kind of accountability that strengthens both constitutional compliance and public trust.

II. INTRODUCTION

More than ten years after the Court issued its remedial order in *Floyd*, the Court continues to oversee the NYPD's stop-and-frisk practices because of persistent concerns related to constitutional compliance, racial disparities, and the accurate documentation of investigative encounters.[13] Although reforms have reduced the number of stops and have involved expanded training and supervisory review, significant challenges remain.[14]

Monitoring efforts over the past decade–including the Monitor's quarterly audits, stop report assessments, focused reviews of specialized units, and statistical analyses–have identified persistent issues. Encounters are inconsistently classified, consent searches raise questions about voluntariness, and much of what officers actually say during stops—the language that signals whether someone understands they are detained or is truly consenting—goes unrecorded in official reports.

BWCs were introduced in New York City to address these concerns by providing an objective record of police–civilian encounters. The Court and Monitor have emphasized the importance of leveraging NYPD's substantial investment in BWC infrastructure as a compliance and oversight tool, including the required recording of Level 1, 2, and 3 encounters.[15] Yet the sheer scale of NYPD activity makes comprehensive manual review of BWC videos infeasible. NYPD officers recorded approximately 4.7 million investigative encounter videos in 2023 alone, documenting nearly 17,000 *Terry* stops.[16]

As a result, even the most rigorous expert review can examine only a small fraction of relevant encounters. The Monitor's quarterly BWC audits review approximately 575 encounters and their BWC videos per quarter.[17] Even intensive manual studies, such as ISLG's analysis of roughly

[13] *Floyd* Remedial Order, 959 F. Supp. 2d 668.

[14] *See* Twenty-Seventh Report of the Independent Monitor (Corrected), *Floyd v. City of New York*, No. 1:08-cv-01034 (AT) (S.D.N.Y. Nov. 19, 2025), ECF No. 971-1; Twenty-Sixth Report of the Independent Monitor, *Floyd v. City of New York*, No. 1:08-cv-01034 (AT) (S.D.N.Y. Oct. 14, 2025), ECF No. 969.

[15] *See Floyd v. City of New York*, No. 1:08-cv-1034 (AT) (S.D.N.Y. Feb. 12, 2021), ECF No. 817; *Floyd* Remedial Order, 959 F. Supp. 2d 684-686; Twelfth Report of the Independent Monitor, *Floyd v. City of New York*, No. 1:08-cv-01034 (AT) (S.D.N.Y. Nov. 30, 2020), ECF No. 798-1; *see also* ISLG Report at 7.

[16] Twenty-First Report of the Independent Monitor at 9, *Floyd v. City of New York*, No. 1:08-cv-1034 (AT) (S.D.N.Y. Sept. 4, 2024), ECF No. 934-1; Twenty-Second Report of the Independent Monitor at 8, *Floyd v. City of New York*, No. 1:08-cv-1034 (AT) (S.D.N.Y. Oct. 7, 2024), ECF No. 937-1 (noting 4.7 million investigative encounter videos recorded in 2023).

[17] *See* Twenty-Seventh Report of the Independent Monitor (Corrected), *Floyd v. City of New York*, No. 1:08-cv-01034 (AT) (S.D.N.Y. Nov. 19, 2025), ECF No. 971-1 (300 reported stops per quarter and 50 stops at NYCHA properties per quarter); Twenty-Fourth Report of the Independent Monitor at 15, *Floyd v. City of New York*, No. 1:08-cv-01034 (AT) (S.D.N.Y. May 21, 2025), ECF No. 960-1 ("Monitor Quarterly Audit of Investigative Encounter BWC videos, N=225 per quarter").

2,000 encounters using retired New York State judges, represent a snapshot of NYPD operations.[18] Consequently, several critical questions remain challenging to answer at Department-wide levels:

- How often do Level 3 encounters—where a person is detained—go unreported or misclassified?
- Do officers request consent to search in the clear, explicit manner required by NYPD policy and constitutional law?
- Are there systematic racial differences in officer language or behavior during investigative encounters?
- Is compliance improving over time—or not?

The Monitor's audits and prior studies have been essential for identifying patterns, establishing baselines, and detecting underreporting.[19] But sampling-based methods, however rigorous, can capture only a small subset of encounters, making it difficult to answer the questions above at scale.

Recent advances in Artificial Intelligence (AI) and, in particular, natural language processing and machine learning, now offer a way to examine these questions Department-wide. These methods make it possible to analyze thousands of encounters at once, identify patterns in officer language that approximate expert legal judgments, and identify linguistic and behavioral patterns that would be infeasible to detect through manual review alone. The use of such methods aligns with the Court's expectation that BWC footage serve as an analyzable compliance tool, not merely a passive record.[20]

The Stanford team used these analytic tools to examine two central questions within the Court's remedial framework:

***De Bour* Compliance.** New York law establishes a four-level framework for police-civilian encounters based on the degree of suspicion and restraint. Under *People v. De Bour*, courts evaluate whether an officer's conduct–and the resulting constraint on a person's liberty–matches the level of suspicion required at each stage, from a request for information up through an arrest. The constitutional concerns in *Floyd* center on Level 3: the stop and detention of a person, which requires reasonable suspicion that the person engaged in criminality. Can

[18] *See* ISLG Report.

[19] *See* ISLG Report; Twenty-Seventh Report of the Independent Monitor (Corrected); Twenty-Third Report of the Independent Monitor, *Floyd v. City of New York*, No. 1:08-cv-01034 (AT) (S.D.N.Y. Feb. 3, 2025), ECF No. 952-1; Twenty-Second Report of the Independent Monitor at 8; Thirteenth Report of the Independent Monitor at 8, *Floyd v. City of New York*, No. 1:08-cv-01034 (AT) (S.D.N.Y. Sep. 1, 2021), ECF No. 853.

[20] *See Floyd* Remedial Order, 959 F. Supp. 2d 684-686.

modern language-based computational models help identify encounters where a person was effectively detained even though the paperwork does not reflect that, and can these tools highlight patterns in officer language, across the city and across racial groups, that relate to these constitutional standards?

Consent Search Compliance. The Fourth Amendment permits searches based on consent, but that consent must be voluntary. In New York City, the Right to Know Act also requires officers to explain a person's right to refuse and confirm their understanding of what is about to happen. When officers document that a search was based on consent, does their spoken language reflect these requirements? Do officers use explicit consent language, or do they rely on indirect or implicit phrasing that may create ambiguity about the nature of the request? And are there racial disparities in how these encounters unfold–not just in how often they occur, but in the language officers use during them?

These questions speak directly to the Fourth and Fourteenth Amendment concerns at the core of the Monitor's mandate and to specific tasks identified in the Court's remedial framework, including the proper documentation of stops and frisks and the assessment of the lawful conduct of investigative encounters and potential racial disparities therein.[21] The analyses that follow address these questions in detail, along with their implications for constitutional compliance and the future of scalable monitoring.

---

[21] *Id*. at 676-678.

III. BACKGROUND AND CONTEXT

A. The *Floyd* Remedial Framework

In August 2013, after a nine-week trial, the U.S. District Court for the Southern District of New York issued a Liability Opinion finding that NYPD's stop-and-frisk practices violated the constitutional rights of the plaintiff class.[22] The Court found that the NYPD violated the Fourth Amendment, which requires *Terry* stops to be based on a reasonable suspicion that a person has committed, is committing, or plans to commit a felony or Penal Law misdemeanor.[23] The Court also found New York City liable for a pattern and practice of racial profiling during *Terry* stops, holding that the City "adopted a policy of indirect racial profiling by targeting racially defined groups for stops based on local crime and suspect data," in violation of the Equal Protection Clause of the Fourteenth Amendment.[24]

The Court held that the City "acted with deliberate indifference toward the NYPD's practice of making unconstitutional stops and conducting unconstitutional frisks," and that, "the NYPD's unconstitutional practices were sufficiently widespread as to have the force of law."[25] Alongside the Liability Opinion, the Court issued a separate Remedies Opinion requiring "immediate reforms" to NYPD's training, documentation, supervision, and monitoring and appointing an Independent Monitor to oversee implementation and ensure constitutional compliance.[26]

The Monitor team's work is guided by a set of foundational legal standards that govern when and how NYPD officers may initiate and escalate encounters. In *De Bour*, New York's highest court, the Court of Appeals, established a four-level framework for police interactions in investigatory situations.[27] The standards set forth the amount of information police need to take action at each of the four levels.[28] Level 1 encounters involve a request for information, which requires an "objective credible reason," to approach a person. Level 2 interactions involve the common law right of inquiry (allowing accusatory questions, but the person is still free to leave), which requires a "founded suspicion" of criminality. Level 3 encounters involve the stop and detention of a person, which requires "reasonable suspicion that a particular person has

[22] *Floyd v. City of New York*, 959 F. Supp. 2d 540 (S.D.N.Y. 2013) ("*Floyd* Liability Opinion").

[23] *Id.* at 560-562; *see also Terry*, 392 U.S. 1.

[24] *Id.* at 562.

[25] *Id.* at 562.

[26] *Floyd* Remedial Order, 959 F. Supp. 2d 676-78.

[27] *De Bour*, 40 N.Y.2d 223.

[28] *Id.*

committed, is committing, or is about to commit a felony or [Penal Law] misdemeanor." This is the level at which the NYPD's "stop, question, and frisk" practices–and the constitutional violations found in *Floyd*–occur. At Level 4, officers effect an arrest or issue a summons, which requires probable cause.

Under the Fourth Amendment, an officer may conduct a brief investigatory stop only when the officer has reasonable suspicion, based on articulable facts, that the person is involved in criminal activity.[29] A frisk, or pat down of a person, is only permitted if the officer reasonably believes that the person is armed and dangerous.[30] Searches beyond a frisk require either probable cause, consent, or another recognized exception to the warrant requirement.[31] *De Bour* further established that New York police officers generally must have at least "founded suspicion" of criminal activity, a standard requiring more than a hunch but less than reasonable suspicion, based on observable conduct or reliable information, before requesting consent to search a person or vehicle.[32]

The *Floyd* Liability Opinion also found violations of the Fourteenth Amendment's Equal Protection Clause, which prohibits racial profiling in police encounters.[33] Officers may not target people for stops based on race or ethnicity, even when other factors may contribute to suspicion.[34]

The Monitor is charged with assessing the Department's compliance with both constitutional requirements: the Fourth Amendment's prohibition against unreasonable searches and seizures, and the Fourteenth Amendment's guarantee of equal protection.[35] To advance this compliance assessment, the Court approved this study by a Stanford-affiliated research team to analyze multiple dimensions of constitutional compliance at scale: whether NYPD officers correctly classify encounters under the *De Bour* framework, whether the Department's consent searches meet Fourth Amendment requirements, and whether there are racial disparities in officer language that implicate the Fourteenth Amendment.[36]

---

[29] *See Terry*, 392 U.S. 1.

[30] *Id.* at 27.

[31] *See Schneckloth*, 412 U.S. 218.

[32] *De Bour*, 40 N.Y.2d at 223.

[33] *Floyd* Liability Opinion at 562, 667.

[34] *Id.*

[35] *See Floyd* Remedial Order.

[36] *Floyd v. City of New York*, No. 1:08-cv-1034 (AT) (S.D.N.Y. Feb. 12, 2021), ECF No. 817.

B. Evolution of Stop-and-Frisk Practices

Since the Court approved this study in 2021, reported *Terry* stops have increased significantly. Officers stopped approximately 8,900 individuals in 2021, rising to more than 15,000 in 2022 and nearly 17,000 in 2023.[37] The composition of *Terry* stops has also shifted over that time. "Self-initiated" stops, where officers make stops based on their own observations rather than responding to 911 or 311 calls, rose from 19% of reported *Terry* stops in 2020, during the height of the COVID-19 pandemic, to 46% by 2023.[38]

These shifts are significant because self-initiated stops have much lower constitutional compliance rates than dispatched stops. The Monitor team found that 96% of reviewed *Terry* stops related to 911 calls in the first half of 2023 were lawful.[39] Self-initiated stops by specialized units, on the other hand, were assessed to be lawful at rates as low as 65%.[40]

Between 2013 and 2022, the total number of stops of Black and Hispanic individuals decreased substantially.[41] But despite the lower stop numbers, the overall racial composition of stops remained largely consistent, with Black and Hispanic individuals comprising approximately 88% of people stopped by the NYPD.[42]

C. New Compliance Systems

In January 2024, the NYPD implemented ComplianceStat, a new accountability system modeled after the Department's long-running CompStat crime-reduction approach, but focused on stop, frisk, search, and reporting compliance.[43] Before ComplianceStat meetings, the Patrol Services Bureau reviews BWC footage and stop-related documentation to identify potential undocumented stops, and improper stops, frisks, and searches.[44] During the sessions, NYPD leadership questions patrol borough and command executives on their units' compliance performance. The Monitor team has described ComplianceStat as a significant accountability innovation but also emphasized that its impact on improving compliance depends on

[37] Twenty-First Report of the Independent Monitor at 9.

[38] *Id.*

[39] Twenty-First Report of the Independent Monitor at 3-4.

[40] Twenty-Third Report of the Independent Monitor at 15.

[41] 2023 End of Year Monitor Update at 5, *Floyd v. City of New York*, No. 1:08-cv-1034 (AT) (S.D.N.Y. Feb. 22, 2024), ECF No. 923.

[42] Twentieth Report of the Independent Monitor at 10.

[43] Twenty-Sixth Report of the Independent Monitor at 16.

[44] *Id.* at 16-17.

consistent implementation and whether the approach is ultimately embraced by all NYPD commands at all levels.[45]

The Department has taken further accountability steps to address its compliance issues. In February 2025, Police Commissioner Jessica Tisch issued memorandum #2025-01-20 titled "Discipline in Connection with Level 3 Stops."[46] The memorandum discussed the lack of discipline on officers who repeatedly engage in unconstitutional practices during *Terry* stops, acknowledging that the NYPD "cannot allow officers to repeatedly make the same errors without consequence" and that the Department's "failure to impose proper discipline for repeated or intentional violations of law and policy in connection with Level 3 stops undermines the integrity of our discipline system."[47] While the Commissioner's memorandum represents an important directive toward serious disciplinary consequences for constitutional violations, its implementation and impact are not yet clear.

D. Ongoing Compliance Challenges

Significant compliance challenges remain that have been discussed in the Monitor team's recent reports. Despite the Department's Quality Assurance Section ("QAS") identifying more than 4,000 improper stops, frisks, or searches in 2024, only 8 incidents (less than 0.2% of identified violations) resulted in a Command Discipline–a supervisor-imposed penalty for officer misconduct.[48] In another 110 instances, an officer was issued a negative Cop's Rapid Assessment Feedback Tool ("CRAFT") comment, the lowest form of misconduct sanction.[49]

NYPD supervisors continue to miss unlawful stops, frisks, and searches. In the third quarter of 2024, sergeants and lieutenants reviewing stop reports in the field flagged just 3% of stops, 1% of frisks, and 7% of searches as unlawful.[50] In the Monitor's sample of reviewed stops for the same quarter, 11% of stops, 26% of frisks, and 26% of searches were deemed improper.[51] More recent data show a persistent gap. In the second quarter of 2025, supervisors identified just 2% of stops, 4% of frisks, and 4% of searches as unlawful, while the Monitor found rates of 9%, 21%, and 23%.[52]

---

[45] *Id.* at 5-6.

[46] *Id.* at 25.

[47] *Id.*

[48] *Id.* at 3.

[49] *Id.*

[50] Twenty-Fourth Report of the Independent Monitor at 10.

[51] *Id.*

[52] Twenty-Ninth Report of the Independent Monitor at 12-13, *Floyd v. City of New York*, No. 1:08-cv-01034-AT (S.D.N.Y. Feb. 26, 2026), ECF No. 979-1.

The Monitor's audit of the Neighborhood Safety Teams (the "NST Report") documented an even greater disparity. During the second quarter of 2022, reviewing supervisors found 100% of the stops legally sufficient while the Monitor determined only 63% were lawful–a 37-percentage point gap illustrating the failure of front-line supervisors to reliably surface constitutional problems.[53]

Officers also under-document *Terry* stops at unacceptable rates. The Monitor team's 2024 BWC audit found that approximately 41% of confirmed stops lacked the required documentation.[54] More recent quarterly audits suggest some improvement, with underreporting rates of 31% in the first quarter of 2025 and 27% in the second quarter of 2025.[55]

Specialized units appear to underreport at even higher rates. In the Monitor's 2023 BWC sample, Neighborhood Safety Team ("NST") officers documented only 46.7% of their stops, and Public Safety Team ("PST") officers documented only 47.1%, compared to 65.8% for patrol officers.[56]

These specialized units that engage in proactive stops, frisks, and searches continue to have substantially lower constitutional compliance rates than patrol officers. The Monitor's 2023 audit determined that 25% of NST stops and 36% of PST stops were unlawful, compared with 8% of patrol stops.[57] For frisks, the gap is wider. In the same audit, NST frisks were legally insufficient 42% of the time, compared with 11% for patrol, while 84% of PST frisks did not meet legal standards.[58] The ISLG study confirmed these patterns, finding that stops conducted by NST officers were unconstitutional 37% of the time, compared to 15% for stops with no NST or PST officers present.[59]

Together, these persistent compliance challenges point to the need for analysis that can systematically assess compliance across units, time periods, and encounter types. Computational methods are particularly well-suited to this and, when paired with the large volumes of BWC data collected by the NYPD, create an opportunity not only to produce insights that strengthen constitutional compliance, but also to improve policing practices.

---

[53] Nineteenth Report of the Independent Monitor at 17-18, *Floyd v. City of New York*, No. 1:08-cv-01034-AT (S.D.N.Y. June 5, 2023), ECF No. 915-1 ("NST Report").

[54] 2024 End of Year Monitor Update at 5, *Floyd v. City of New York*, No. 1:08-cv-01034-AT (S.D.N.Y. Feb. 26, 2025), ECF No. 953.

[55] Twenty-Eighth Report of the Independent Monitor at 13, *Floyd v. City of New York*, No. 1:08-cv-01034-AT (S.D.N.Y. Jan. 20, 2026), ECF No. 974-1; Twenty-Ninth Report of the Independent Monitor at 14.

[56] *Id.* at 9.

[57] Twenty-Third Report of the Independent Monitor at 15.

[58] *Id.*

[59] ISLG Report at 33.

E. Development of Computational Methods for Analyzing BWC Footage

Despite the widespread adoption of BWCs, law enforcement agencies have largely treated the resulting footage as case-specific evidence.[60] Our research suggests the footage can support much more.[61] A 2017 study applied computational linguistics methods to nearly 1,000 traffic stops and found that officers spoke less respectfully to Black community members even after controlling for multiple contextual factors.[62] The study demonstrated that BWC data could reveal dimensions of police-community interactions–like respect and procedural justice–that standard administrative data rarely records.[63]

Further research has extended beyond respect to other aspects of officer communication. Computational methods can detect whether officers state reasons for stops, ask consent for searches, or offer reassurance to drivers.[64] Studies have found that racial disparities in officers' tone of voice can undermine institutional trust, and that linguistic patterns in an officer's first 45 words (roughly, the first 27 seconds of a stop) can predict whether those stops will escalate to searches, handcuffing, or arrests.[65] Parallel technical work demonstrates that building automated pipelines for transcribing and processing BWC audio is possible.[66]

Recent research has shown that these methods can measure the effectiveness of reforms on policing practices. In a 2024 study comparing body-worn camera footage before and after a police department's procedural justice training, officers employed more of the techniques recommended in the training such as expressing concern for drivers' safety, offering

---

[60] *See generally* Carolyn Naoroz, Body-Worn Cameras Then and Now: A 10-Year Retrospective and Future Directions, *Police Chief Magazine* (Oct. 2025); Cynthia Lum, Megan Stoltz, Christopher S. Koper & J. Amber Scherer, Research on Body-Worn Cameras: What We Know, What We Need to Know, 18 *Criminology & Pub. Pol'y* 93 (2019).

[61] Nicholas P. Camp & Rob Voigt, Body camera footage as data: Using natural language processing to monitor policing at scale & in depth 10(2) *Behavioral Science & Pol.*, (2024).

[62] Rob Voigt et al., Language from Police Body Camera Footage Shows Racial Disparities in Officer Respect, 114 Proc. *Nat'l Acad. Sci. U.S.A.* 6521 (2017).

[63] *Id*.

[64] Vinodkumar Prabhakaran et al., Detecting Institutional Dialog Acts in Police Traffic Stops, 6 *Transactions Ass'n for Computational Linguistics* 467 (2018).

[65] Nicholas P. Camp et al., The Thin Blue Waveform: Racial Disparities in Officer Prosody Undermine Institutional Trust in the Police, 121 *J. Personality & Soc. Psychol.* 1157 (2021); Eugenia H. Rho et al., Escalated Police Stops of Black Men Are Linguistically and Psychologically Distinct in Their Earliest Moments, 120 *Proc. Nat'l Acad. Sci. U.S.A.* e2216162120 (2023).

[66] *See* Anjalie Field et al., Developing Speech Processing Pipelines for Police Accountability, in *Proc. Interspeech* 2023 1229, 1229–33 (2023).

reassurance, and providing explicit reasons for stops.[67] This line of research by the Stanford team provides the foundation for the methods applied in this report.

F. Computational Methods in the *Floyd* Monitoring Context

Since the Court directed the NYPD to implement BWCs in 2013, the footage has figured centrally in the Monitor's compliance assessments. The NST Report relied on BWC video review to assess specialized unit compliance; underreporting studies have used footage to identify undocumented stops; and the Housing Bureau evaluation examined the impact of the cameras on officers in NYCHA developments.[68] At the core of these efforts, the Monitor's quarterly audits involve detailed human review of camera footage.

This work is crucial for assessing constitutional compliance but is resource-intensive given the NYPD's scale. The Department conducts thousands of reported *Terry* stops annually while officers record millions of investigative encounter videos each year.[69] Computational approaches offer the potential to analyze these large volumes of data in ways that would be impossible through manual review.

These methods can systematically assess aspects of police-community interactions that are captured in officer and civilian language but challenging to evaluate encounter by encounter. The voluntariness of a consent search, for example, can depend on how officers frame their requests and these language patterns are well suited to analysis by machine learning models. Similarly, whether an encounter constitutes a Level 2 investigative encounter or a Level 3 *Terry* stop can turn in significant part on what officers say, how the civilian responds, and the physical circumstances of the encounter. These computational methods complement rather than replace human review, offering an AI-powered way to analyze many of the constitutional compliance issues that are the focus of the Monitor's work.

The Court approved this study alongside the ISLG study in February 2021, with each designed to contribute distinct but complementary analyses of the NYPD's compliance with Fourth and Fourteenth Amendment principles.[70] The ISLG Report used retired New York State judges to measure constitutional compliance, providing reliable legal determinations.[71] The Stanford

---

[67] Nicholas P. Camp et al., Leveraging Body-Worn Camera Footage to Assess the Effects of Training on Officer Communication During Traffic Stops, 3 *PNAS Nexus* pgae359 (2024).

[68] *See* NST Report; Twenty-Second Report of the Independent Monitor; Seventeenth Report of the Independent Monitor, *Floyd v. City of New York*, No. 1:08-cv-1034 (AT) (S.D.N.Y. Oct. 17, 2022), ECF No. 894.

[69] *See* Twenty-First Report of the Independent Monitor at 9; Twenty-Second Report of the Independent Monitor at 8.

[70] *Floyd v. City of New York*, No. 1:08-cv-1034 (AT) (S.D.N.Y. Feb. 12, 2021), ECF No. 817.

[71] *See* ISLG Report.

study, by contrast, uses machine learning techniques to analyze BWC recordings, focusing on what officers say and how they say it during encounters.

Consistent with the Court's order, this report presents the results of a large-scale computational analysis of officer language in recorded encounters. Within this scope, we prioritized two areas where AI tools and language-based methods provide the clearest and most meaningful insights for compliance monitoring: (1) *De Bour* encounter classification and documentation, and (2) consent-search requests. These topics arise in many encounters, are captured in the routine audio officers record, can be measured in a way that supports validation and systematic auditing, and are well-suited to identifying rare, hard-to-find noncompliance that human review can miss. Our initial research proposal included analyses of complaints. Although this topic is not separately analyzed here, the measures and findings we present, such as the use of commands, ambiguity in requests, and interactional markers (i.e., signals in how a conversation unfolds) are relevant to understanding how encounters may deteriorate over time and whether they generate complaints.

G. BWC Data Preparation and Processing

The manner in which we analyze body camera recordings differs across our aims–predicting *De Bour* level classification and examining linguistic variability among consent searches. However, the general procedure required to prepare this footage for analysis is similar. Recordings are *matched* to administrative records, they are *transcribed* (manually and automatically), *diarized* to note what is said in interactions and who says it, and *preprocessed* to convert those transcriptions into a computer-readable structured format. Table 1 provides an overview of this preparation process and its application to the Stanford team's aims.

Table 1. Data Used in the Report.

Sample Name	Administrative Records	Matching Provenance	Usable Encounters and Transcripts	Application in Report
ISLG Sample	Stratified sample of 2,133 encounters from 3/16/2022 to 5/15/2022	ISLG	1,702 encounters (1,702 auto-transcripts, 341 manual transcripts)	*De Bour* Documentation and Compliance Consent Searches Model Evaluation
Monitor-Assessed Sample	1,200 Monitor-assessed Level 2 encounters 2022-2024	Monitor, NYPD	1,156 encounters (1,156 auto-transcripts)	*De Bour* Documentation and Compliance Model Evaluation
Consent Search Sample	Full population of 2,005 consent search requests documented in stop reports taking place in 2023	NYPD	1,770 encounters (3,695 auto-transcripts)	Consent Searches

We note that the number of usable encounters in each case is fewer than the number of associated administrative records. For the Monitor-Assessed and Consent Search Samples, a small number of encounters are lost due to noise in the data, such as repeated failed Evidence.com Application Program Interface (API) requests, extremely short transcripts, and matching errors. For the Consent Search Sample, since we are particularly interested in understanding requests to search, our analyses are restricted to the subset of encounters in which officers documented that they requested consent to search. The final usable number of encounters represents those which we are able to match to BWC footage provided by the Department, with transcripts containing at least 50 words. For the ISLG Sample, research assistants manually coded for each encounter a "verbal recording," that is, the piece of body camera footage out of the potentially multiple associated with an encounter that best captured the verbal exchanges occurring between the officers and civilians. Targeting high data quality, we analyze the subset of encounters for which research assistants in their study identified a "verbal recording" that was also identified as being conducted by the "lead officer" in the encounter.

The first step of the BWC data pipeline is identifying the set (or subset) of encounters to be analyzed, and matching them to relevant metadata. One approach is to associate administrative records to body camera recordings, such as identifying consent searches from NYPD stop, question, and frisk ("SQF") records and pulling the associated BWC recording. A second path is to

conduct a stratified random sample of BWC recordings and then search for corresponding recordings. For example, ISLG randomly sampled body camera recordings tagged as low-level encounters (those labeled as Level 1/Level 2 in the BWC database), stops (Level 3-labeled recordings), and arrests/summonses (Level 4-labeled recordings), and gathered administrative records associated with each encounter. Since there are typically multiple recordings per encounter, and no mandatory common identifier between a BWC recording and administrative data (e.g., an SQF form), recordings are grouped and then matched based on the timestamp associated with each video and the ID of the recording officer.

Next, matched BWC recordings are transcribed, producing a text script of each video that logs who is speaking, what they are saying, and when each line is spoken. All BWC video files were processed through Axon's auto-transcription service, and professional transcribers manually corrected a subset of these transcripts. Each approach carries strengths and limitations: auto-transcripts can be generated at scale, but tend to be less accurate, particularly in identifying *who* said what (a speech processing task called *diarization*). Manual transcription, in which a person views and transcribes the contents of each recording, is more accurate, but also more time-consuming to produce.

We evaluate the quality of automatic transcripts provided by the Axon API (Appendix VIII.C.), finding a usable word-level error rate of approximately 31 errors in 100 words that aligns with prior research in the area, but a poor diarization error rate of 0.706 out of 1.0 where higher is worse. To partially mitigate this problem, we separately train high-precision officer identification models. Nevertheless, the quality of automatic transcription–and particularly diarization–represents fundamental limitations on how fine-grained our analyses can be.

To comply with NYPD data restrictions, transcribers were required to edit auto-generated transcripts within Axon's Evidence.com platform. This requirement further limited our access to manual transcription; professional transcribers had difficulty working in the Evidence.com interface, and several opted not to work on the project. Transcripts–machine or human-generated–were then preprocessed into suitable formats for computational analyses, including linkage with all available administrative data.

We note that in the real world these data have a complex, interwoven structure. Encounters are fluid multi-person interactions in which multiple officers could interact with multiple civilian contacts, which are imperfectly captured by pieces of BWC footage. Analysis can therefore take place at several levels: expert judgments of constitutionality in the ISLG study, for example, targeted the level of individual civilian contacts. Particularly due to the challenges of accurate automatic diarization (identification of who is speaking), we are limited in our ability to evaluate behavior at the level of individual contacts. To do so accurately would likely further require systems for automated addressee detection to identify which specific individuals the officers are speaking to in each utterance. We expect such a goal is relatively far off, however: more

research in this nascent technical field has focused on detecting when users are speaking to devices than on distinguishing human addressees from each other.[72]

Here, we treat an encounter with the public as documented in administrative records as our primary unit of analysis, which may be associated with multiple BWC videos and thus multiple transcripts. For *De Bour* Compliance (Section IV) we use a single transcript from the encounter as primary, as described below. For analyses of Consent Searches (Section V) we use all available transcripts, targeting whether specific behaviors were recorded in the language of any associated transcript.

The end result of this process is a structured dataset, wherein all transcripts associated with an encounter are logged, allowing us to link the content of specific utterances to higher-level features of the encounter, such as the recorded *De Bour* level, as well as expert judgments where available, such as evaluations of compliance for each civilian contact in the ISLG Sample. These linguistic data can then be used for predictive purposes, or to observe heterogeneity in officers' language based on factors such as the type of encounter or the civilian's race.

H. BWC Data Sampling

We employ three sets of BWC recordings to predict *De Bour* classification (Section IV) and analyze the language of consent search requests (Section V):

ISLG Sample

In their report to the Monitor, ISLG collected a stratified random sample of BWC recordings collected over a two-month period (3/16/2022-5/15/2022). After screening out recordings that were beyond the scope of the study, in a language other than English/Spanish, or which provided an incomplete record of the officer-civilian interaction, the final sample consisted of 2,133 recordings labeled as *De Bour* Level 1 or Level 2 (N=622), Level 3 (N=1,453), or Level 4 (N=58).[73]

The ISLG team then proceeded to match BWC recordings to administrative records of each incident (e.g., stop reports for Level 3-labeled encounters) and officer (e.g., personnel records). As part of this process, ISLG researchers also grouped together multiple recordings associated with each encounter in which more than one officer was recording. From a panel of retired New York State judges, two judges reviewed the recordings associated with each encounter, as well as any available police reports, to assess the constitutionality of encounters labeled as stops, and to determine whether encounters labeled as low-level contacts rose to the level of a stop.[74]

---

[72] *See* Ingo Siegert, Norman Weißkirchen, and Andreas Wendemuth, Acoustic-Based Automatic Addressee Detection for Technical Systems: A Review, 4 *Frontiers in Computer Science* (2022).

[73] *See* ISLG Report at 8-12.

[74] *See* ISLG Report at 12-14. In cases where the two judges disagreed in their assessment of an encounter, a third judge reviewed the footage to establish a consensus judgment. Approximately 8% of low-level encounters required a third judge to adjudicate rater disagreement.

Underreported stops are defined as encounters that were labeled as a *De Bour* Level 1 or Level 2 by the recording NYPD officer but that reviewing experts determined to be *De Bour* Level 3 or Level 4 encounters.

We added officers' language to this corpus by applying Axon's auto-transcription, provided by Rev.ai, on all 1,702 videos in the ISLG study (the "ISLG Sample").[75] As auto-transcription tends to be less accurate than human transcribers, we sampled a subset of these recordings (N=341) to be transcribed by expert transcriptionists, who received an auto-transcription of the recording, which they then corrected in the Evidence.com interface. In our analysis of *De Bour* classification, we use human-corrected transcripts wherever possible. Further, we draw on these transcripts to refine and evaluate auto-transcriptions: to compare the accuracy of Evidence.com auto-transcription in identifying words and speakers, and to examine fine-grained instances of language associated with frisks and searches.

As mentioned above, the ISLG study identified a "verbal recording" that best captured the verbal exchanges occurring between the officers and civilians. We analyze only the subset of cases where this recording is captured by the officer identified as a "lead officer," and for the predictive modeling described in Section IV we take this verbal transcript as our primary record of the encounter and use it for all analysis.

Monitor-Assessed Sample

In addition to the encounters sampled by ISLG, we drew on prior audits of Level 2-labeled recordings conducted by the Monitor team in 2022, 2023, and 2024. Two experts on the Monitor team reviewed 400 BWC recordings per year which were labeled as a Level 2 investigative encounter by the conducting officer. These experts categorized each recording as a low-level encounter (below Level 3) or a stop (Level 3 or above). We compiled Evidence.com auto-transcriptions of all BWC recordings associated with 1,156 of these encounters successfully extracted using the API (the "Monitor-Assessed Sample").

Consent Search Sample

Alongside these data, which had already been assembled by the Monitor team, we compiled a set of 3,695 usable BWC recordings of 1,770 NYPD-documented stops in which officers sought consent to search civilians to analyze the language of consent search requests. The NYPD provided the study team with records of all encounters in which an NYPD officer documented that they requested consent to search, as logged in 2023 NYPD SQF data; the study team then requested all BWC recordings associated with each of these stops. The NYPD matched these recordings, which were then auto-transcribed in Evidence.com.

---

[75] *See* Axon, Rev.ai, https://www.axon.com/partners/rev-ai (last visited Dec. 23, 2025).

IV. *DE BOUR* DOCUMENTATION AND COMPLIANCE

The Court's Liability Opinion found that the NYPD's stop, question, and frisk practices were unconstitutional.[76] As part of the Remedial Order, the Court required NYPD officers to properly and fully document stops, consistent with *De Bour*.[77] In response to the plaintiffs' recommendations to the Court, the City proposed that officers record a wider range of encounters, including those at *De Bour* Level 1 and Level 2, and document these contacts by logging the level of the stop in the Axon metadata and, for Level 2 encounters, recording in the BWC metadata the race and gender of the civilian with whom the officer interacted in the Level 2 encounter report.[78]

In 2021, the Court approved two studies to assess NYPD compliance with Fourth Amendment requirements, as well as the documentation and proper classification of encounters based on their *De Bour* level.[79] The ISLG Report, filed in May 2025, sampled BWC recordings labeled as low-level encounters (*De Bour* Level 1 or Level 2), stops (Level 3), and arrests (Level 4), matched these recordings to incident-level data, and had retired New York State judges determine whether a stop had indeed occurred.[80]

A core finding from the ISLG Report is that underreporting takes different forms depending on encounter type. Among encounters already documented as stops (Level 3), approximately 23% to 24% of contacts–roughly one in four individuals stopped–were not documented with a stop report, even though at least one officer had labeled a recording as a stop (e.g., either a stop report was prepared, but not for all of the persons stopped in the encounter, or at least one officer recorded the encounter as a stop in the BWC metadata, but a stop report was not prepared). Among encounters documented as low-level (Level 1 or Level 2), the rate of unreported stops was lower–approximately 3% of encounters–but this rate should be interpreted against the large number of low-level encounters to which it is applied. As described above in Section III.G., we are limited in our ability to look at behavior towards individual civilian contacts, so we focus our analyses at the level of the entire encounter. In the ISLG Sample, if any contact in an interaction is documented as a stop (*De Bour* Level 3 or 4), for our purposes we classify the entire interaction as a stop. The Monitor-Assessed Sample consists of audits of individual pieces of body camera footage, so we take each piece of footage as the full available record of the relevant encounter.

[76] *Floyd* Liability Opinion at 561-563.

[77] *Floyd* Remedial Order at 681-683.

[78] *See Floyd v. City of New York*, No. 1:08-cv-1034 (AT) (S.D.N.Y. Feb. 12, 2021), ECF No. 817.

[79] *Id.*

[80] *See* ISLG Report.

In both cases, many unreported stops are unlikely to be detected through existing oversight. While stops that end in arrest become "visible" through arrest reports, over half of the unreported stops identified by ISLG did not end in an arrest, leaving no such documentation trail.

The Stanford team was charged with applying computational techniques to identify underreported and non-compliant encounters in the massive pool of NYPD recordings. To do so, we use experts' *De Bour* categorization–retired judges in the ISLG Report as well as the Monitor team's audits–to identify linguistic features that distinguish low-level (i.e., Level 1 and Level 2) encounters from Level 3 stops. We can then use these features to identify BWC recordings in which the *De Bour* level is most likely to be underreported: BWC recordings labeled as Level 1 or Level 2, but whose linguistic content more closely resembles those of Level 3 stops.

In identifying the linguistic signatures of different *De Bour* levels, we are also able to shed light on the content of these encounters, and how they differ across race, findings relevant to the Monitor's ongoing assessment of Fourteenth Amendment compliance.

A. Classifying Documentation of Stops vs. Low-Level Encounters

Key Questions:

- How linguistically distinct are stops (*De Bour* Level 3 and Level 4) from low-level encounters (*De Bour* Level 1 and Level 2)?
- Can computational models augment human judgment to identify cases of under-documentation?
- What linguistic features are predictive of stops vs. low-level encounters?

We first test whether low-level encounters (those with a *De Bour* Level of 1 or 2) can be distinguished from stops (those with a *De Bour* Level 3 or 4), based on the language used in the encounter. If officers' words are a meaningful signal to the *De Bour* classification of an interaction, then it would be possible to compare officers' classifications of their encounters to prediction based on their words, identifying potential under-documentation. On the other hand, legal judgments of *De Bour* level could require additional visual and contextual information that cannot be gleaned from language alone. Accordingly, we first derive linguistic models that classify whether NYPD encounters meet the threshold for a stop, then compare these models to expert human judgment from retired judges in the ISLG study and members of the Monitor team to assess models' ability to identify under-documented encounters.

1. Tasks and Methodological Approach

In order to explore the above questions, we train machine learning classification models to predict the *De Bour* encounter level from the language in transcripts alone. The fundamental pipeline employed is one in which an input piece of BWC footage is associated with a

categorical label, in this case *low-level encounter* (Level 1 and Level 2) vs. *stop* (Level 3 and Level 4). We train machine learning models to predict this label from the language in the associated transcript and evaluate their performance. We decompose the problem into three related tasks, each of which aims to distinguish encounters from stops, but using different sources of judgment and samples of data.

Task 1: Officer Documentation Classification. Given a correctly documented interaction, predict whether it was recorded as an encounter or a stop. This is a fundamental task that we evaluate to understand baseline linguistic separability. If properly documented stops are possible to separate from true low-level encounters with high performance, this would provide evidence that the linguistic information to do this type of task is present in the transcript. If not, this would be suggestive that much of the relevant information requires the audio or video. We evaluate this task using the ISLG Sample, removing improperly documented stops as evaluated by expert judges, leaving only encounters and stops properly documented by officers as training data.

Task 2: Unreported Stop Classification. Given a transcript from an investigative encounter documented as low-level in BWC metadata, predict whether expert judgment would classify it as a stop. This is a more challenging task which seeks to evaluate fine-grained differences among investigative encounters documented as low-level to assess whether they are in fact undocumented stops. We evaluate this task using the Monitor-Assessed Sample, where primary labels are assessed judgments made by members of the Monitor team.

Task 3: *De Bour* Classification. Given a transcript from any interaction with the public, predict whether it was a stop (*De Bour* Levels 3 and 4) or low-level encounter (*De Bour* Levels 1 and 2). For this task we use all available data from both the ISLG and Monitor-Assessed Samples and regard the "true label" of each interaction to be its final expert-corrected classification as an encounter or stop.

Classification tasks such as these encounter challenges in interpretation if the underlying data is imbalanced. For example, in a classification task between outcomes A and B where 99% of the outcomes are A, a model can achieve near-perfect numerical accuracy by always predicting A, even if this is not useful because B has been completely disregarded. The imbalance is not quite as severe in this data–the available data in the ISLG Sample has 75% Level 3 and Level 4 stops, and the Monitor-Assessed Sample has 80% Level 1 and Level 2 encounters. For clarity and comparability across tasks when evaluating predictive accuracy, for each task we randomly undersample the data to balance equal amounts of stops and encounters.

This means predictions can be interpreted across outcomes relative to a 50% baseline that would be achieved by random guessing. Undersampling functionally results in increasing the weight of samples in the under-represented class, and is often a preferred approach for related

problems (such as medical diagnostics[81]) where the cost of false negatives is higher than the cost of false positives.[82] The union of both samples, used for Task 3, by happenstance has almost equal proportions of true encounters and stops, so requires only minimal undersampling and therefore makes best use of the available data for training.

We explore two modeling approaches; detail on each is provided in the Appendix (Section VIII.E.). The first is statistical machine learning, in which we extract word and phrase features ("n-grams") from each transcript, and the learning process fits weights from scratch, associating these features with the labeled outcome using relatively low parameter count models to identify a classification boundary. We report primary results with gradient boosting classifiers as implemented in the machine learning software library XGBoost[83] and linear support vector machines ("SVM"). The relative benefits of these more traditional models are *interpretability*–we can identify direct associations between the presence of a given feature and a predicted outcome–and *calibration*–we can fit these models in such a way that they produce meaningful probabilities, as will be described further in the findings below.

The second modeling approach is LLM fine-tuning, in which we adjust the weights of pre-trained large language models towards the classification task at hand by training on labeled examples. We report results for full fine-tuning on DistilBERT[84] and ModernBERT,[85] two encoder-only models well-suited for classification tasks (contrasting with the more well-known generative models better suited for producing text). The relative benefits of this contemporary approach are the potential for greater *robustness*–since they rely less directly on the presence of particular words or phrases due to exposure to vast amounts of pre-training data–and in some cases resulting in stronger *performance*. The drawbacks include that this class of models is less inherently interpretable and calibratable, and requires heavier computational resources to train than traditional statistical models. We ultimately find that these models fail to perform better than traditional statistical machine learning models in the current context and discuss the implications of this towards the end of this Section.

---

[81] *See* Mabrouka Salmi et al., Handling Imbalanced Medical Datasets: Review of a Decade of Research, 57 *Artificial Intelligence Review* 273 (2024).

[82] *See* Yanmin Sun, Andrew K. C. Wong & Mohamed S. Kamel, Classification of Imbalanced Data: A Review, 23 *International Journal of Pattern Recognition and Artificial Intelligence* 687 (2009); Oded Maimon & Lior Rokach, eds., Data Mining and Knowledge Discovery Handbook (2010).

[83] XGBoost ("eXtreme Gradient Boosting") is an open-source software library used to train gradient-boosting machine learning models. *See* XGBoost Documentation, https://xgboost.readthedocs.io/en/stable/.

[84] *See* Victor Sanh, Lysandre Debut, Julien Chaumond & Thomas Wolf, DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter, *arXiv* preprint arXiv:1910.01108.

[85] *See* Benjamin Warner et al., Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference, *In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (2025).

2. Evaluation of Predictive Performance

Our core findings regarding predictive performance from the best-performing model for each task are given in Table 2. We measure generalization performance, meaning the ability of the model to correctly predict new samples it has not previously been trained on. This is done using 5-fold cross-validation: each model is trained five times independently on different 80% subsets of the data, holding out 20% of the data each time for evaluation. Reported accuracies represent the average performance across these five folds.

Table 2. Generalization accuracy for encounter vs. stop predictive tasks across four model types and three tasks.

	Model			
Task	*XGBoost*	*SVM*	*DistilBERT*	*ModernBERT*
1. Officer Documentation Classification *(ISLG Sample)*	89.8%	91.0%	86.2%	88.9%
2. Unreported Stop Classification *(Monitor-Assessed Sample)*	68.2%	72.9%	68.3%	72.2%
3. *De Bour* Classification (*ISLG and Monitor-Assessed Samples*)	81.8%	81.7%	79.0%	81.2%

We broadly find high predictive accuracy across all model classes: up to 91% for Officer Documentation Classification, above 70% for Unreported Stop Classification, and above 80% for full *De Bour* Classification on the joint sample, again relative to a 50% random guessing baseline. These findings are significant for several reasons. Most directly, they provide strong evidence that the language captured in BWC footage contains a signal that can be used to automatically identify key classes of stops at accuracies substantially better than chance.

From a behavioral perspective, high performance on the Officer Documentation Classification task suggests that encounters and stops are substantively linguistically separable, that is, these interactions are distinct in their linguistic behavioral profile. If models struggled to distinguish these correctly documented classes of stops it could be suggestive that the distinction was linguistically subtle. By contrast, here we find evidence for clear differences even with traditional word- and phrase-based statistical models.

Even on the challenging task of Unreported Stop Classification, which seeks to estimate the appropriate *De Bour* level within only encounters documented by officers as low-level, we find models can achieve substantially better than chance performance. Classification performance on the joint sample which includes both properly and improperly documented stops is also strong; the use of this joint sample is potentially most analogous to the real-world context in which we may want to incorporate information both from ongoing officer documentation as well as judgments from experts and audits, the possibility of which we discuss below.

3. Auditing Potential and Model Calibration

Given the strong performance identified above, important potential applications of such models moving forward could include computational monitoring or computationally assisted auditing, in which models survey larger amounts of data than is practical for humans to manually review and flag cases which seem particularly problematic for further human assessment.

The usability of these models for auditing in this way depends, in part, on calibration. Models such as those trained here generally inherently produce probabilistic estimates of the likelihood of each outcome. A well-calibrated model is one in which these estimates align with the actual probability that each prediction is correct. For example, in a well-calibrated model, if we observe ten predictions which each have a probability near 60%, we should expect that roughly 6 out of 10 of those predictions should be correct. If models are poorly calibrated, the probabilities they return are much less directly useful. A model could have similar accuracy to a well-calibrated one, but if it always returns probabilities that are close to 100% for the predicted class then we have less information on when to trust or distrust any given prediction.

Indeed, this type of "over-confidence" is a well-documented issue for contemporary large language models, though some methods exist for inducing better calibration.[86] For traditional statistical machine learning, more clear and established methods are available; since SVMs represent our strongest performing model, we perform cross-validated calibration using the sigmoid method on our SVM models.[87] Applying this method to both models, we find relatively high capacity for calibration. Calibration curves for each task are shown in Figures 1, 2, and 3.

[86] *See e.g.,* Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, & Iryna Gurevych, A Survey of Confidence Estimation and Calibration in Large Language Models, In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2024); Yangyi Chen, Lifan Yuan, Ganqu Cui, Zhiyuan Liu, & Heng Ji, A Close Look into the Calibration of Pre-trained Language Models, In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (2023).

[87] *See* John C. Platt, Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods (1999); Alexandra Niculescu-Mizil & Rich Caruana, Predicting Good Probabilities with Supervised Learning, In *Proceedings of the 22nd International Conference on Machine Learning* 625-632 (2005).

Figure 1. Calibration plot for Task 1, Officer Documentation Classification.

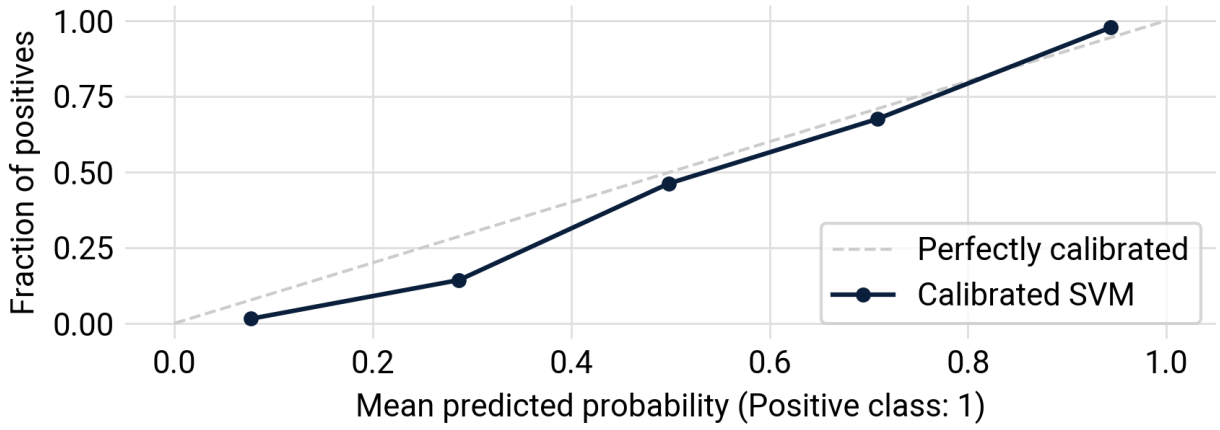


Figure 2. Calibration plot for Task 2, Unreported Stop Classification.

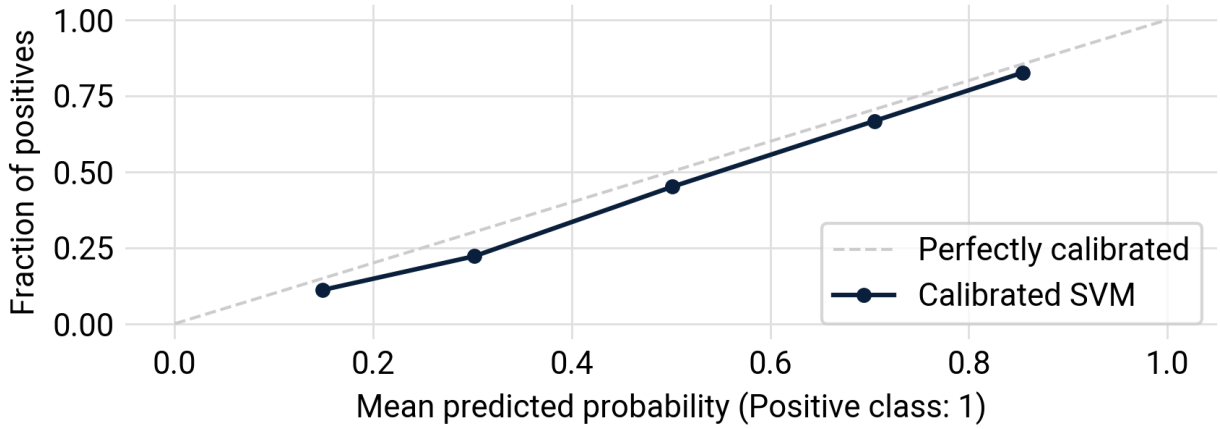
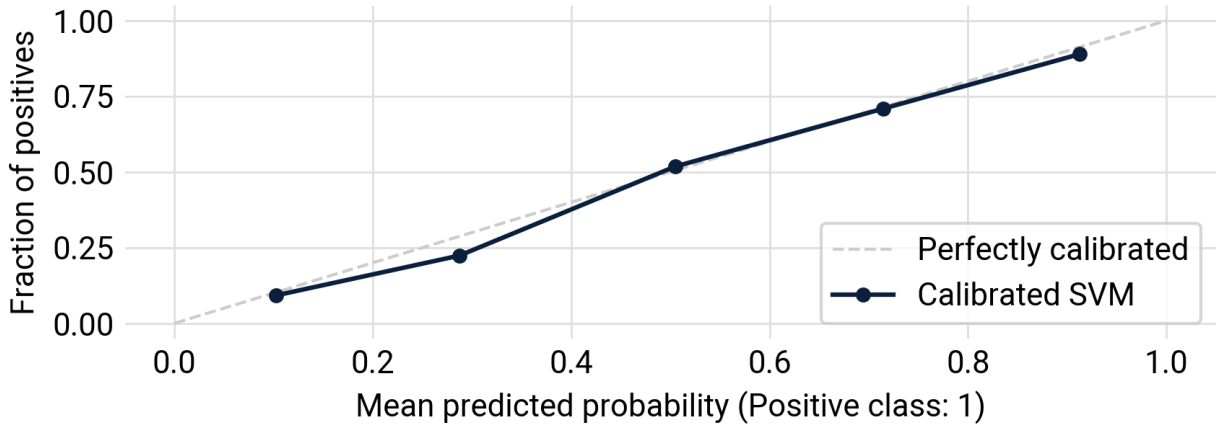


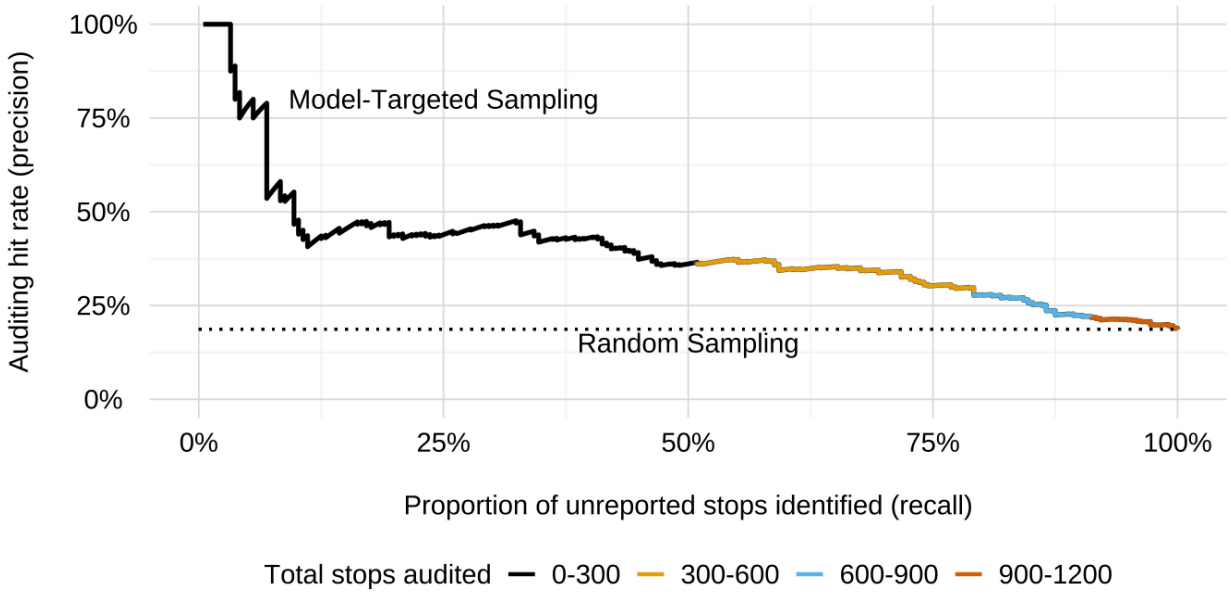
Figure 3. Calibration plot for Task 3, *De Bour* Classification.



Broadly we find a high capacity for calibration; theoretical perfect calibration is shown in each plot as a dotted diagonal line. We find that our models tend to be slightly over-confident in their predictions, as illustrated by measurements below that line, but that overall models exhibit sufficient calibration that we expect the generated probabilities to be useful for auditing.

We further test this possibility by simulating a scenario in which the *De Bour* classification model was applied to the full sample of Level 2-labeled recordings selected for quarterly audits by the Monitor, i.e., the Monitor-Assessed Sample. For each instance we calculate the model's stop probability and compare it to the true labels using a precision-recall curve, shown in Figure 4. The probability for each sample is derived within cross-validation, and therefore represents a generalizable estimate made without the model having been trained on that sample.

Figure 4. Precision-recall curve for Monitor-Assessed Sample using *De Bour* Classification model



This figure compares an auditing strategy we will call model-targeted sampling, in which the Monitor team would audit stops in decreasing order of their stop probability as predicted by the model to a random sampling process currently used by the Monitor team to choose which *De Bour* Level 2-labeled videos to review. On the figure, random sampling is represented as a horizontal line at 18.7%, indicating the proportion of audited L2 encounters in the sample which, following review, were determined to be improperly documented stops.

We can visualize the auditing process as movement along the jagged line from the top left to the bottom right as additional recordings are audited. On the x-axis is the *recall*, or the proportion of the total unreported stops as identified by the Monitor team audit. On the y-axis is the *precision*, equivalent to the cumulative auditing hit rate at that point. That is, the y-axis identifies the proportion of audited encounters found to be undocumented stops, out of the total number audited at a given level of recall on the x-axis. The colored segments help visualize the auditing sequence as additional sets of recordings (depicted in bins of 300 videos each) of lower and lower stop probability are selected for review.

Several key observations can be taken from this figure. First, the model-based sampling line never falls below the horizontal random sampling line: model-targeted sampling consistently provides a higher hit rate than random audits. A substantial subset of unreported stops–approximately 30% of the total–can be identified with an auditing hit rate near 50%. If the Monitor team audited only the 300 encounters in the sample with the highest stop probability (the dark blue segment of the line at the left), they would have had a hit rate of 36.3% (the precision on the y-axis) and found half of the total unreported stops the Monitor ultimately identified in the sample (the corresponding recall on the x-axis). They would have done so with only one-quarter of the effort expended on random sampling to find those stops.

Second, there is a substantial long tail of unreported stops that are more challenging for the model to identify. Each successive segment of the model-targeted line covers a shorter distance on both x and y-axes: there are smaller gains in recall and precision as the model selects recordings with lower probabilities of being a stop to audit. This reflects the increasing challenge of classifying recordings assigned lower probabilities, but also suggests that the bases on which Monitor team audits identified some encounters as unreported stops emerge from aspects of BWC recordings beyond transcripts (for instance, the visual presence of multiple officers surrounding a civilian in such a way that they would no longer reasonably feel free to leave). Future modeling approaches may aim to incorporate these factors to improve this performance in such situations, but, for the time being, this is suggestive that the type of model-targeted sampling proposed here should not be the only approach applied.[88]

[88] The Monitor's current sampling procedure for L2 encounters serves an important purpose: it produces unbiased prevalence estimates that can be compared over time to track the Department's efforts in addressing underreporting. Model-directed sampling, by contrast, focuses on encounters with the highest predicted probability of being underreported stops. Because it is intentionally designed to oversample the cases most likely to be non-compliant, it cannot by itself provide the same kind of prevalence estimates (although statistical techniques such as post-stratification could be used to approximate them). Separate

Finally, it is important to note that the above estimates are being made only *within* the small sub-sample of footage already audited by the Monitor. Therefore, they are likely substantial under-estimates of the hit rates we could expect in the real world across the full span of NYPD BWC footage by targeting audits toward encounters with high model probabilities.

To illustrate this, consider that there is a small proportion of encounters at the high end of predicted probability (upper left) for which the model estimates are, in our sample, completely correct. This suggests that roughly the top 3% of samples can be audited with a hit rate near 100%. This estimate is based on the relatively small Monitor-Assessed Sample, where this 3% represents only 7 out of 216 unreported stops identified, so this must be taken as a rough estimate. Whether this trend holds at full scale across a broader spectrum of footage is an empirical question; nevertheless, if it does hold, we could expect that the subset of footage with the highest model probabilities is likely to contain nearly 100% unreported stops. In 2025, there were approximately 183,000 recordings per month categorized as low-level encounters; audits of the top 3% by model probability would therefore yield 5,490 recordings per month with a very high likelihood of being unreported stops based on language alone.

4. Language Associated with *De Bour* Levels 3 and 4

The calibrated statistical machine learning models used above have the additional benefit of interpretability: as a part of the learning process, the models learn weights which represent the relative association of any of a set of terms with the predicted output class, so we can then directly inspect these trained models to understand the impact of these features that form the basis of their predictions. The core models above use features representing sequences of words or phrases ("n-grams") of lengths 1, 2, or 3 – also known as unigrams, bigrams, and trigrams. Here, for each task we separately train models using only unigram, bigram, and trigram features for descriptive purposes to better understand the language associated with *De Bour*-elevated stops in Level 3 and Level 4. Table 3 shows the top 20 features with the highest learned weights from each model and phrase length.

Qualitatively observing the language associated with stops under each model we make several observations. Across tasks, stops are broadly associated with several behaviors we might expect to occur in such interactions: explicit mentions of stops ("we stopped [you]"), offering of business cards, and increased use of commands ("come over here", "put your hands").

Comparing language associated with Task 1 and Task 2 offers additional insights. Recall that stops in Task 1 were all properly documented stops while stops in Task 2 were unreported (i.e., labeled as low-level encounters). While properly documented stops are more likely to mention "the reason" for the stop and to explicitly state whether civilians are "free to go," these features

from prevalence estimation, identifying particular non-compliant encounters has independent value for purposes such as officer feedback, supervisor review, and discipline, and model-targeted sampling is well suited to support that work. These two approaches could also be combined, using a random sampling component for prevalence estimation while supplementing it with a model-directed approach aimed at identifying non-compliant encounters.

are absent as top predictors for unreported stops. Explicit mentions of a potential "search" are predictive for documented stops in Task 1; indirect or implicit search-associated language such as "check" and "you don't mind" (discussed in detail in Section V.B.) are predictive for unreported stops in Task 2.[89] Words which were likely spoken by civilians seem predictive in both cases ("don't/ain't got nothing"), but occur much more prominently among predictive features for unreported stops ("nothing on me," "what the fuck," "you can check").

[89] Here and elsewhere, these features are predictive in that they are associated with a particular category or outcome. For an overview of statistical versus causal prediction, *see* Galit Shmueli, *To Explain or to Predict?*, 25 Statistical Science 289-310 (2010).

Table 3. Top predictive words and phrases for stops relative to encounters from models trained on one-, two-, and three-word feature sets across each task.

Task	Top Stop Associated Terms
1. Officer Documentation Classification (ISLG Sample)	**unigram:** got, id, business, card, have, stopped, knife, that, ticket, my, why, search, pocket, your, turn, put, gun, information, nothing, take **bigram:** you have, you got, business card, what's your, the reason, your hands, the phone, why you, hold on, on you, turn around, on me, nothing on, you good, got nothing, public safety, right here, over here, free to, your id **trigram:** come over here, do you have, put your hands, what's your name, your last name, you have anything, on hold on, back to the, we stopped you, anything on you, don't got nothing, hold on hold, your first name, the reason for, lemme see your, do me favor, free to go, take your hands, you coming from, you have id
2. Unreported Stop Classification (Monitor-Assessed Sample)	**unigram:** check, what, yo, saw, stopped, search, trying, man, business, who, well, apartment, walking, stop, card, into, gun, hands, appreciate, coming **bigram:** business card, hold on, trying to, appreciate you, you saw, that's all, at the, the park, waiting for, your hands, on me, you understand, what the, we stopped, got nothing, check you, to talk, came from, right have, me bro **trigram:** put your hands, nothing on me, what the fuck, but you can't, we stopped you, you have any, got the wrong, you can check, want business card, you don't mind, you know me, you got no, lemme see your, business card you, in my car, ain't got nothing, you guys okay, supposed to have, you want business, you got nothing
3. *De Bour* Classification (ISLG and Monitor-Assessed Samples)	**unigram:** stopped, hold, cv, put, supposed, receipt, pocket, check, figure, somebody, description, knife, from, female, positive, site, warrants, cold, shoes, him **bigram:** hold on, your pocket, we stopped, on me, the cv, stopped you, put him, business card, what's your, the receipt, public safety, pay it, waiting for, got nothing, him in, in your, my bag, that's not, the description, we saw **trigram:** put him in, we stopped you, in your pocket, on hold on, can you just, put your hands, but you can't, what's your name, hanging out with, or are you, bro why you, thank you man, you have any, one male stop, do you have, gonna go back, it all right, get out here, what is this, you can call

B. Classifying Compliance within Stops

Key Questions:

- How linguistically distinct are constitutionally compliant stops from non-compliant stops?
- Can computational models augment human judgment to identify cases of constitutional non-compliance?
- What linguistic features are predictive of constitutional non-compliance?

Though the initial scope of this report was focused primarily on the issue of underreporting of stops, the same methods used in Section IV.A. for predictive modeling of documentation can be applied to questions of constitutionality, so we briefly present relevant findings in this regard here.

1. Tasks and Methodological Approach

In the ISLG study, expert judges reviewed interactions either initially documented as or ultimately found to be stops and assessed them for constitutionality along various axes including whether the officer had reasonable suspicion for the stop and whether frisks were conducted constitutionally. We leverage these labels as a source of ground truth, marking a stop as constitutionally non-compliant if any person officers engaged with in the interaction experienced a non-compliant stop. We implement computational models in an analogous fashion to Section IV.A. for the following additional predictive task.

Task 4: Stop Compliance Classification. Given a transcript from a stop encounter, predict whether expert judgment would classify it as containing any unconstitutional actions from officers. We evaluate this task using the ISLG sample, subset to only encounters which were ultimately identified as stops and use as primary labels the consensus of expert judges in the ISLG study regarding constitutionality. As in Tasks 1-3, since non-compliant stops only constitute 18.4% of the total we undersample compliant stops to produce a balanced dataset for comparable evaluation.

2. Evaluation of Predictive Performance

Predictive performance findings of the models, presented as generalization accuracies via 5-fold cross-validation, are given in Table 4. We find performance around 70%, meaning that the model can predict whether the stop was compliant or not 70% of the time, compared to random guessing which would make such predictions with 50% accuracy. This is similar to but slightly underperforming models for Task 2 (Unreported Stop Classification).

Table 4. Generalization accuracy for stop constitutionality prediction across four model types.

	Model			
Task	XGBoost	SVM	DistilBERT	ModernBERT
4. Stop Compliance Classification (ISLG Sample)	70.0%	71.6%	70.2%	65.8%

3. Auditing Potential and Model Calibration

We evaluate calibration for the purposes of auditing in the same manner as presented in Section IV.A.3., fitting calibrated probability estimates to an SVM model using the sigmoid method. A calibration curve for our Task 4 model is presented in Figure 5, and a precision-recall curve is presented in Figure 6.

Figure 5. Calibration plot for Task 4, Stop Compliance Detection.

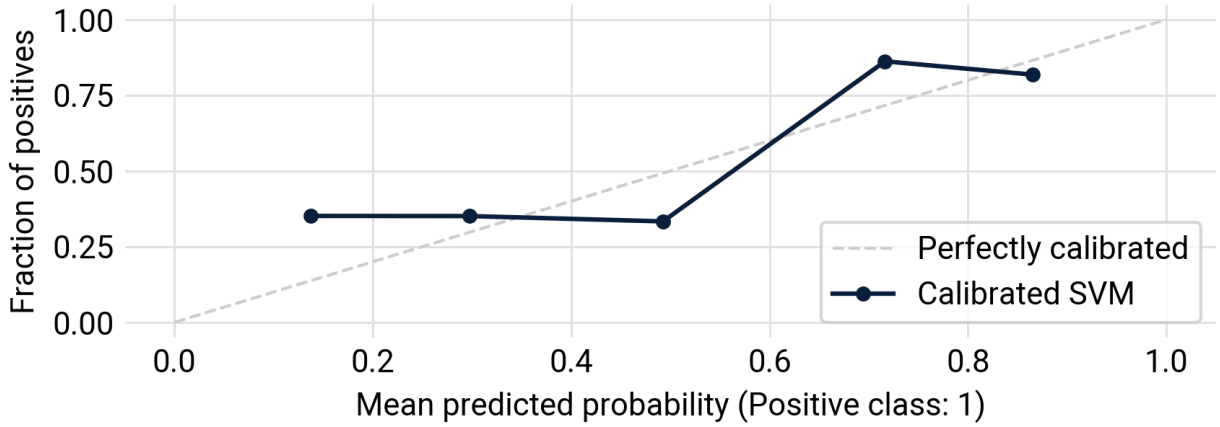
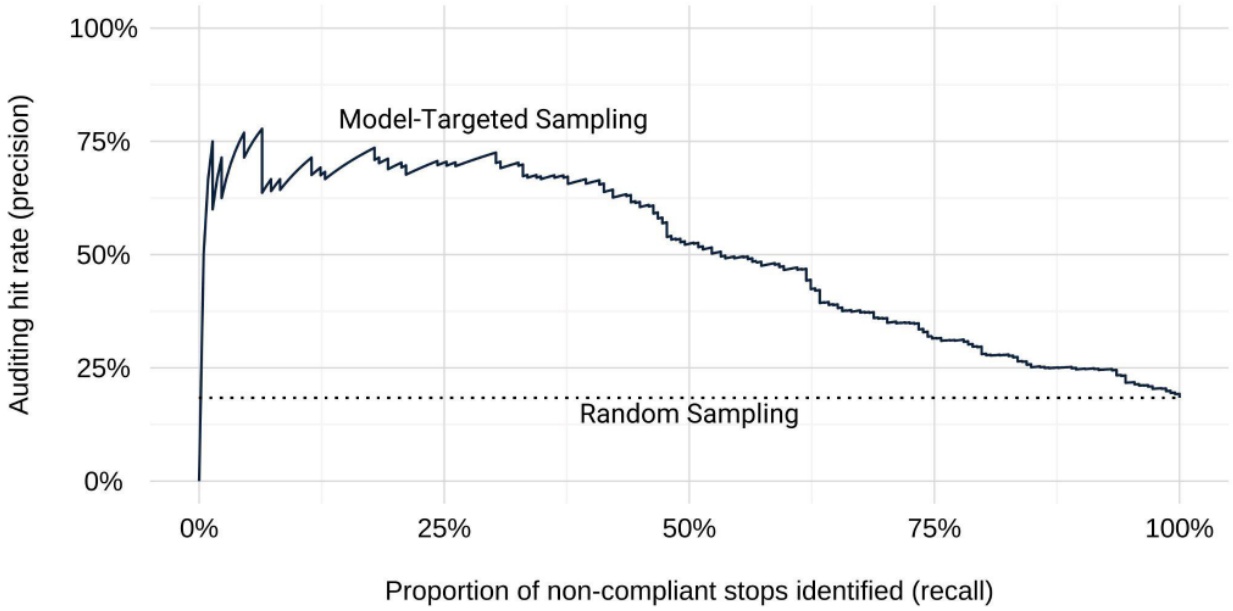


Figure 6. Precision-recall curve for ISLG Sample using Stop Compliance Detection model



We find these models to be somewhat less amenable to calibration than documentation models, showing some propensity for overconfidence in the middle bucket (predictions near 50% confidence) and underconfidence in the second-highest bucket (predictions near 70% confidence). Nevertheless, these models produce potentially very useful estimates; the precision-recall curve is suggestive that, similar to the case of *De Bour* classification, model-targeted sampling could substantially raise hit rates for auditing non-compliant stops.

4. Language Associated with Non-Compliant Stops

Using the same methods as in the documentation case, we can identify words and phrases learned by the models to be most predictive of non-compliance, given in Table 5.

Table 5. Top predictive words and phrases for stops relative to encounters from models trained on one-, two-, and three-word feature sets across each task.

Task	Top Non-Compliance Associated Terms
4. Stop Compliance Classification (ISLG Sample)	**unigram:** nothing, east, right, stop, wrong, listen, male, contact, video, running, everything's, weapon, touching, good, keys, jacket, radio, here, up, those **bigram:** you running, got nothing, over here, nothing on, stop stop, you're good, on you, the video, hang out, contact card, had to, what you, you need, go go, do not, book bag, don't reach, you just, the wrong, reason for **trigram:** why you running, what you got, nothing on me, supposed to have, what's your name, want contact card, you got nothing, go go go, the reason for, that's why that's, why that's why, in the building, stop stop stop, for no reason, explain to you, want business card, good to go, don't got nothing, you go to, are you running

We can observe that the top predictive features for non-compliant stops involve, on one hand, officers asking questions like “why you running,” “what you got,” and “[got nothing/anything] on you” as well as potential commands such as “stop,” “hang out,” and “don’t reach.” On the other hand, utterances likely spoken by civilians are also prominent predictors, including “nothing on me,” “for no reason,” and mentions of “touching.” Mentions of “[got] the wrong [person]” are predictive of non-compliant stops, which also appear as a predictive feature for Task 2 (Unreported Stop Classification). Several features that on their own seem potentially indicative of a properly conducted stop–including explanations (“explain to you”) and offering contact/business cards–are predictive of non-compliant stops.

C. Racial Disparity Across Levels and Interaction Types

Key Questions:

- Do we observe racial disparities when using computational models to compare Level 1/Level 2 encounters to Level 3/Level 4 stops? Are racial disparities present when we compare compliant stops to non-compliant stops?
- What do model predictions of *De Bour* level tell us about the similarity between low-level encounters and stops?
- Are interactions with racial minorities more “stop-like,” independent of their *De Bour* level?

Our model assigns a level of confidence that a given BWC recording captures a low-level (*De Bour* Level 1 or Level 2) versus a stop-level (*De Bour* Level 3 or Level 4) encounter. One way of interpreting these values is as a measure of probability, or the likelihood that a recording contains a stop. Recordings can be categorized as falling above or below a threshold to be manually reviewed or flagged for underreporting.

This confidence metric also helps us to determine how "stop-like" an encounter appears, based on the language it contains. As such, it speaks to concerns raised in *Floyd* that officers may "inadvertently treat [civilians] in such a way that a reasonable person would not feel free to leave."[90] Given the Court's finding that the NYPD's stop practices violated the Fourteenth Amendment through the "disproportionate and discriminatory stopping of Blacks and Hispanics," it is important to examine whether these linguistic patterns vary by race.[91]

With these concerns in mind, we compare the similarity between language in BWC recordings and known Level 3/Level 4 interactions across Black, White, and Hispanic civilian demographic groups at each labeled *De Bour* level. This approach lets us ask, for example, whether a Level 2 interaction between an NYPD officer and a Black civilian may more closely resemble a Level 3 interaction than a Level 2 interaction with a White civilian. If so, it would suggest racial disparities in how free civilians are (or would feel) to leave in Level 2 interactions.

1. Estimated Stop Probability by Race

The calibrated models for each task described above produce probabilities representing how likely a given interaction is to be a stop under the model, providing an estimated judgment of how "stop-like" the recording appears based on the language it contains. We compare this metric across civilian race at each *De Bour* level in two ways: observationally with reference to density histograms of stop probability faceted by race, and quantitatively with linear regressions predicting stop probability from race.

In the ISLG Sample, the race of multiple individuals who officers interacted with ("contacts") in an interaction is recorded. Since our analyses operate at the encounter level, to capture the Monitor's interest in disparate treatment toward Black and Hispanic civilians in particular, we operationalize civilian race in the following way: if any contact is recorded as Black, we record the interaction as Black; if not, and any contact is marked as Hispanic, we record the interaction as Hispanic; if neither, we record the interaction as White/Other.[92] Monitor assessments identify a race of the primary person stopped, which we code analogously.

---

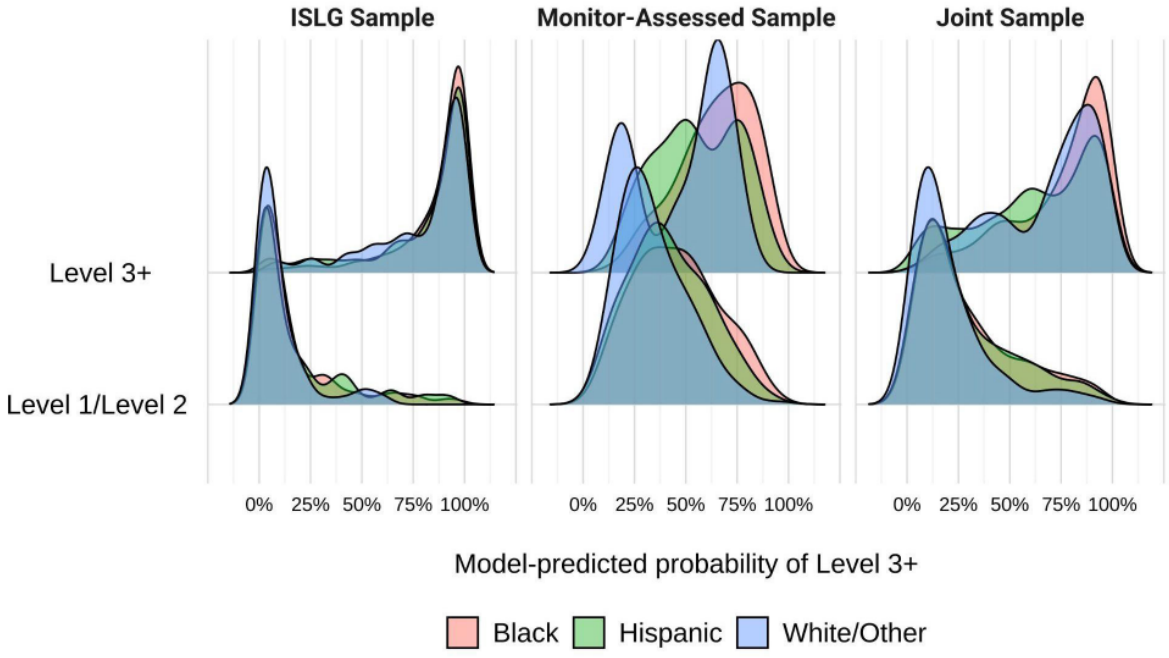
[90] *Floyd* Remedial Order, 959 F. Supp. 2d 679 n.38.

[91] *Floyd* Liability Opinion at 562.

[92] These categories match those used in previous Monitor analyses. *See* Twentieth Report of the Independent Monitor at 8.

Observing Figure 7, we find evidence for a subtle but consistent racial disparity, such that independent of their true *De Bour* level, interactions with Black and Hispanic civilians more closely resemble *De Bour* Level 3 and above in the language used, as judged by the model. This figure communicates two further points of note. First, there is substantial increased density towards the left of the distribution for stops of White/Other civilians, particularly within the Monitor-Assessed Sample. This is indicative of a cluster of stops which, though legally classified as *De Bour* Level 3 or Level 4, linguistically appear more similar to encounters at *De Bour* Level 1 or Level 2. Second, is the small but significant long tail of Level 1 and Level 2 encounters, with Black civilians in particular receiving high probability estimates (i.e., above 0.5). This is indicative of a cluster of interactions that, though legally classified as encounters rather than stops, share substantial linguistic characteristics with stops.

Figure 7. Model-estimated stop probability by race in the ISLG Sample (among properly documented stops, left panel), the Monitor-Assessed sample (among encounters documented as low-level, center panel), and the combination of these samples (among all interactions, right panel).



We test these differences quantitatively by fitting independent linear regressions predicting the model-estimated stop probability from civilian race for each level and sample. As previous analyses of racial disparities conducted by the Monitor have done, we also fit regressions adjusting for a suite of control variables: gender (whether any contact was identified as female), borough, and the local crime rate (as measured by the count of arrests, shootings, and criminal

summons in the preceding 30 days in the census block in which the encounter took place).[93] For Task 1 (Officer Documentation Classification) these controls further include whether an NST officer was present, the suspected crime (Weapons, Violence, Other), whether the encounter was self-initiated or NYCHA-associated, and the location (indoor, outdoor, transit). These variables are only available for the ISLG sample, so they are not included for Tasks 2 and 3.

These findings are shown in Table 6, presented as the change in probability of an L3 or higher classification relative to interactions with White/Other civilians.

Table 6. Difference in estimated stop probability for Hispanic and Black civilians relative to White/Other reference group for each task/sample. P-values are given in parentheses. Each cell has both a raw (R) and covariate-adjusted (C) estimate.

		Civilian Race	
Model	Level	Hispanic	Black
1. Officer Documentation Classification (ISLG Sample)	Level 1/2 Encounters	R: 7.5%** C: 7.5%**	R: 6.1%** C: 7.4%**
2. Unreported Stop Classification (Monitor-Assessed Sample)		R: 7.8%*** C: 5.5%**	R: 11.2%*** C: 8.9%***
3. *De Bour* Classification (ISLG and Monitor-Assessed Samples)		R: 8.2%*** C: 9.0%***	R: 9.0%*** C: 10.3%***
1. Officer Documentation Classification (ISLG Sample)	Level 3/4 Stops	R: 0.1% C: 0.0%	R: 6.4%** C: 5.3%*
2. Unreported Stop Classification (Monitor-Assessed Sample)		R: 6.2% C: 2.4%	R: 16.9%* C: 10.4%
3. *De Bour* Classification (ISLG and Monitor-Assessed Samples)		R: -4.7% C: -2.9%	R: 6.1%* C: 7.4%**

**p≤.05, **p≤.01, ***p≤.001*

[93] *See* Twentieth Report of the Independent Monitor; Thirteenth Report of the Independent Monitor; Fifth Report of the Independent Monitor, *Floyd v. City of New York*, No. 1:08-cv-01034 (AT) (S.D.N.Y. May 30, 2017), ECF No. 554.

We find that Level 1 and Level 2 encounters with Hispanic and Black civilians receive stop probability estimates 5-11% higher than White civilians across tasks, even when accounting for relevant control variables like local crime rate. These findings provide robust evidence that under the models presented here these low-level encounters are significantly more linguistically "stop-like" than encounters with White and Other-race civilians. These differences largely hold in Level 3 and Level 4 stops for Black (5-17% higher), but not Hispanic civilians, where the differences are not significant. This suggests that in terms of the language that appears in these interactions, stops with Black civilians in particular display relatively more language characteristic of stops than those of other civilians, and in turn stops of Hispanic, White, and Other-race civilians display relatively fewer.

One possible additional explanation for these patterns could be that in the original datasets civilian race is unequally distributed across encounters vs. stops: across the data roughly 57% of interactions with Black civilians are stops while only 30% of interactions with White and Other-race civilians are stops. If models are picking up on race-specific language, these estimates could be generated mechanically, since stops are more likely to occur with Black civilians. To control for this, we fit a separate Calibrated SVM on the joint sample using only interactions with Black civilians and use this model to predict generalization stop probabilities for all interactions. Doing so establishes Black interactions as a baseline against which to judge stop probability for all races. We find that, using this model, the above trends hold. *De Bour* Level 1 and Level 2 encounters receive higher stop probabilities by 6.8% ($p=<0.001$) for Black and 6.9% ($p=<0.001$) for Hispanic civilians, while *De Bour* Level 3 and 4 encounters receive higher stop probabilities by 6.5% ($p=<0.001$) for Black civilians.

2. Estimated Compliance Probability by Race

We apply similar analyses as those in the preceding section to constitutional compliance using the model from Section IV.B. for Task 4: Stop Compliance Classification. To do so, we subset the joint sample to only stops, and use the calibrated Task 4 model to estimate the probability of non-compliance for each case, splitting analyses between reported and unreported stops. We show a density histogram of these probabilities in Figure 8, and estimated change in non-compliance in Table 7. As with Tasks 1, 2, and 3, we adjust these estimates for potential confounds including controls for gender, borough, and the local crime rate.

Figure 8. Model-estimated non-compliance probability by race, across reported and unreported stops.

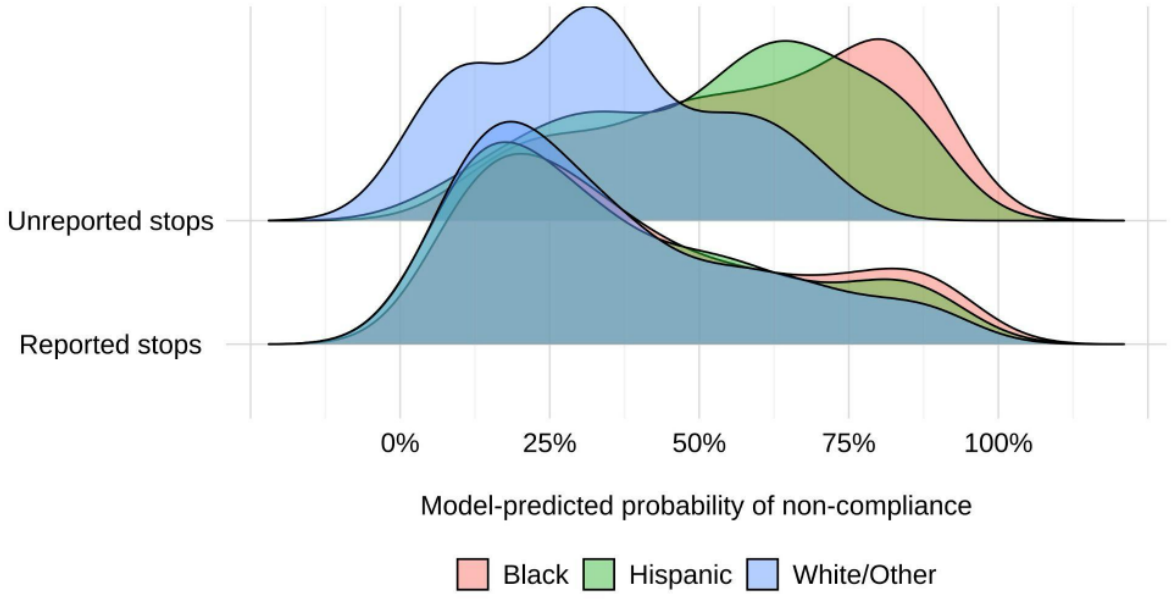


Table 7. Change in estimated non-compliance probability for Hispanic and Black civilians relative to White/Other reference group. P-values are given in parentheses. Each cell has both a raw estimate (R) and one accounting for control variables (C).

		Civilian Race	
Model	Level	Hispanic	Black
4. Stop Compliance Classification (ISLG and Monitor-Assessed Samples, L3/L4 Stops Only)	Reported Stops	R: 2.6% C: 1.3%	R: 5.7%* C: 3.7%
	Unreported Stops	R: 22.1%*** C: 16.6%**	R: 26.6%*** C: 22.2%***

**p≤.05, **p≤.01, ***p≤.001*

We find that properly documented stops show no significant racial disparity for Hispanic civilians and a 5.7% increased estimate of non-compliance for Black civilians (which is not significant under controls); however, within unreported stops we find highly statistically significant increases ranging from 16-27% higher non-compliance estimates for both Hispanic

and Black civilians compared to White and other-race individuals, even accounting for controls. This suggests that while racial disparities in non-compliance estimates for reported stops are more modest or non-existent, amongst unreported stops the interactions officers have with Hispanic and Black civilians display substantially more linguistic characteristics of constitutionally non-compliant stops.

As in the case of under-documentation, non-compliant stops are more frequent with Black civilians (who represent the largest group of individuals stopped) than with civilians of other races. Analogously to the previous section, we control for the possibility of the observed differences being caused by race-associated language by fitting equivalent models only for interactions with Black civilians, and find all effects are still significant for reported stops, with an increased non-compliance probability for Black civilians of 4.3% (p=0.038), and for unreported stops, with increased probabilities for Hispanic civilians of 12.5% (p=0.013) and Black civilians of 14.9% (p=0.002).

D. Summary

Previous audits of the NYPD's *De Bour* classification of investigative encounters have identified the challenges in both assigning the proper *De Bour* level to encounters and in monitoring compliance with the requirements of the *Floyd* Remedial Order. The sheer number of recordings captured by NYPD personnel creates a formidable "haystack" in which to search for the "needles" of improperly documented and non-compliant stops. Random audits of small numbers of recordings may not be sufficient to find the needles, and a thorough search of the haystack is costly. Is there a way of identifying the recordings which are most likely to contain an undocumented stop? As we show here, automated analyses of officer language hold promise as a way of, if not finding all needles, greatly reducing the size of the haystack.

AI models like those used in this study generally become more accurate as they are trained on larger amounts of higher-quality data. As additional encounters are reviewed for *De Bour* compliance–by the Monitor, legal experts, or NYPD personnel–those reviews can be used to refine the model further, increasing its predictive power. This type of approach is often referred to as "human-in-the-loop" machine learning.[94] The area of "active learning" within this subfield provides a set of methods such as uncertainty sampling, which allow practitioners to mathematically identify samples about which the model is least certain, and these can be foregrounded for human auditing and annotation, potentially increasing the value of each individual sample as training data for future iterations of the model.[95] Ultimately the long-term promise held by models such as those presented here for auditing is that the value of a given audit is not isolated to only the particular interaction being evaluated, but rather it also

[94] *See* Robert (Munro) Monarch, Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-centered AI, *Simon & Schuster* (2021).

[95] *See* David D. Lewis & Jason Catlett, Heterogeneous Uncertainty Sampling for Supervised Learning, In *Machine Learning Proceedings* 148-156 (1994).

contributes to the ongoing improvement of computational tools that improve accuracy and reduce effort.

We found better performance from statistical machine learning models compared to fine-tuning large language models (see Section IV.A.2 above); similar findings have been documented for text classification tasks with smaller-scale data in specialized domains.[96] We also expect it is possible that with increased amounts of training data, newer models may be able to overtake the more traditional approaches. The models presented here are also limited by the quality of Axon auto-transcription, which is poor. More accurate speech-to-text processing (such as models tuned to the policing context) can close the gap between human transcriptions and machine-generated ones.[97] Thus, the results we describe above can be considered as a "floor" for this task–a minimum level of performance that we expect can be improved with additional labels and more, higher-quality training data.

One way the Monitor or NYPD could integrate our approach into ongoing audits would be to identify the BWC recordings most in need of review. Consonant with ISLG's observations from their manual review of encounters, we uncover disparities in police-civilian interactions which could not be discerned from NYPD administrative records. We find that the language used in interactions with Black and Hispanic civilians more closely resembles that in stops, particularly for Black civilians. This suggests that, even in low-level police encounters, NYPD interactions with Black and Hispanic civilians take on a more "stop-like" tenor, raising questions about whether a reasonable person in such encounters would understand they were free to leave. Moreover, looking at constitutional compliance within stops, our model estimates provide evidence that unreported stops in particular display the most substantial racial disparities in officer language, which is particularly concerning since such unreported stops are far less likely to be subject to audit and review.

Variation in officer language is important not just for capturing the nuances of police-community relations, but also for determining whether officer conduct falls within Department policies or constitutional requirements. Many of these guidelines explicitly apply to officers' communication when interacting with the public, words which carry the weight of law. The information officers provide–or withhold–from civilians can distinguish a legal stop from an unconstitutional one.

---

[96] *See* Yasmen Wahba, Nazim Madhavji & John Steinbacher, Attention Is not Always What You Need: Towards Efficient Classification of Domain-Specific Text: Case-Study: IT Support Tickets, In *Science and Information Conference* 1159-1166 (2023); Scott Barnett, Zac Brannelly, Stefanus Kurniawan & Sheng Wong, Fine-Tuning or Fine-Failing? Debunking Performance Myths in Large Language Models, *arXiv* preprint arXiv:2406.11201 (2024).

[97] *See* Field et al., supra note 66.

V. CONSENT SEARCH COMPLIANCE

Consent is a recognized exception to the Fourth Amendment's warrant requirement: a police officer may search a member of the public without probable cause or a warrant if the person freely and voluntarily consents.[98] Upwards of 90 percent of warrantless police searches have historically been conducted by means of this consent exception, but the conditions under which such searches are truly "voluntary" are a complex subject of legal inquiry that is shaped by the circumstances of the encounter, including in significant part how officers communicate during the interaction.[99]

Under the totality-of-the-circumstances approach established in *Schneckloth*, courts assess whether consent was voluntary rather than a mere submission to a claim of lawful authority.[100] What officers say–and how they say it–can bear directly on this inquiry, including whether they clearly communicate the nature of the request and frame it as one the person may decline. Under New York's *De Bour* framework, officers must also have at least a "founded suspicion" of criminality before requesting consent to search, a standard requiring more than a hunch but less than reasonable suspicion, based on observable conduct or reliable information.[101] Consent may also be ineffective if obtained during an unlawful detention.[102]

In New York City, the Right to Know Act–and the NYPD's Patrol Guide implementing it–further operationalizes these constitutional requirements by specifying the manner in which officers must seek consent.[103] Procedure No. 212-11 of the NYPD Patrol Guide notes that if an officer is seeking consent to search, the request and the person's response must be video-recorded if the officer has a BWC, and further describes a clear sequence of questions–and responses–necessary to conduct a consent search:

> *You may seek consent to search. Consent must be voluntarily given.*
>
> > *(a) Ask for consent to search in a manner that elicits a clear 'yes' or 'no' response. When seeking consent, make clear that the search will not occur if the person does not consent. For example, in a non-threatening manner and without making promises, you may ask the*

[98] *See Schneckloth*, 412 U.S. 218.

[99] *See* Ric Simmons, Not Voluntary but Still Reasonable: A New Paradigm for Understanding the Consent Searches Doctrine, 80 *Ind. L.J.* 773 (2005); *See* note 10 on the role linguistic factors can play in determining the voluntariness of the consent obtained during a consent search.

[100] *See Schneckloth*, 412 U.S. 218, 227; *Bumper v. North Carolina*, 391 U. S. 543, 548–549 (1968).

[101] *De Bour*, 40 N.Y.2d at 223.

[102] *Florida v. Royer*, 460 U.S. 491, 501 (1983).

[103] N.Y.C. Admin. Code § 14-173(a)(1), (3).

> *following: "I can only search you, if you consent. Do you understand? May I search you?"*[104]

However, reviews of investigative encounters have raised questions about NYPD officers' compliance with these requirements. A recent Monitor audit of recordings from NYPD's Community Response Team traffic stops found that 54% of consent searches in the audit appeared not to meet the "founded suspicion standard" established in *People v. De Bour*, a finding consistent with–and even higher than–rates of noncompliance in earlier audits.[105]

These findings raise broader questions about the voluntariness of consent searches under the Fourth Amendment. Here, we ask how NYPD officers request consent to search: whether their language clearly communicates the nature of the request and that consent may be refused–factors courts consider under *Schneckloth's* totality-of-the-circumstances approach–or whether officers rely on implicit or ambiguous phrasing that may leave civilians uncertain about what they are being asked to consent to. New York City's Right to Know Act and Patrol Guide requirements provide concrete benchmarks for assessing these practices.

To this end, we employ transcripts of BWC footage to observe variations in language with which officers conduct consent searches, variation that bears directly on Fourth Amendment voluntariness but is not observable in SQF records. Specifically, we first calculate the prevalence of explicit consent search requests–those which use the NYPD's recommended language and make the reason for the search clear–versus other kinds of requests. When officers deviate from these guidelines and request consent otherwise or implicitly, how do they do so? We identify prevalent forms of requests as well as other linguistic aspects of the stop that bear on voluntariness such as the presence of commands. We also examine whether there are racial disparities in how officers request consent, implicating the Fourteenth Amendment's equal protection guarantee. As with the predictive tasks above regarding *De Bour* documentation, the goal of this analysis is to offer both a set of results and a suite of tools to observe these features of NYPD interactions at scale and on an ongoing basis.

A. Explicit Requests for Consent

Key Questions:

- What is the prevalence of explicit consent language in consent searches?
- Does the consent language officers use differ by the race of the civilian?

[104] NYPD Public Patrol Guide, Procedure 212-11, Revision R.O. 53, https://www.nyc.gov/assets/nypd/downloads/pdf/public_information/public-pguide2.pdf.

[105] Twenty-Fifth Report of the Independent Monitor at 12; *See* Nineteenth Report of the Independent Monitor.

Our first question of these data is a simple one: do NYPD officers request consent to search members of the public consistent with the Fourth Amendment and the NYPD's internal policies? How often are such requests made explicitly? And, if officers do not make such requests explicitly, how do they elicit consent?

While this question is straightforward, it is also challenging to answer at scale. Where our first aim of *De Bour* classification drew on language throughout BWC recordings to estimate a feature of the encounter (i.e., whether the *De Bour* level crossed the threshold for a stop), consent search requests require a more fine-grained approach: search requests are particular events that occur within an encounter, and the sequence of officer and civilian language is critical.

In light of the limitations of auto-transcriptions under the current data pipeline, we triangulate between identification of key words pertaining to consent, content coding of language during consent searches, and, where possible, supplement auto-transcripts with gold-standard human transcriptions. These analyses reveal important large-scale trends in the language of NYPD consent searches, even as they are limited by the quality of the transcripts.

1. Quantifying Explicit Consent Search Language

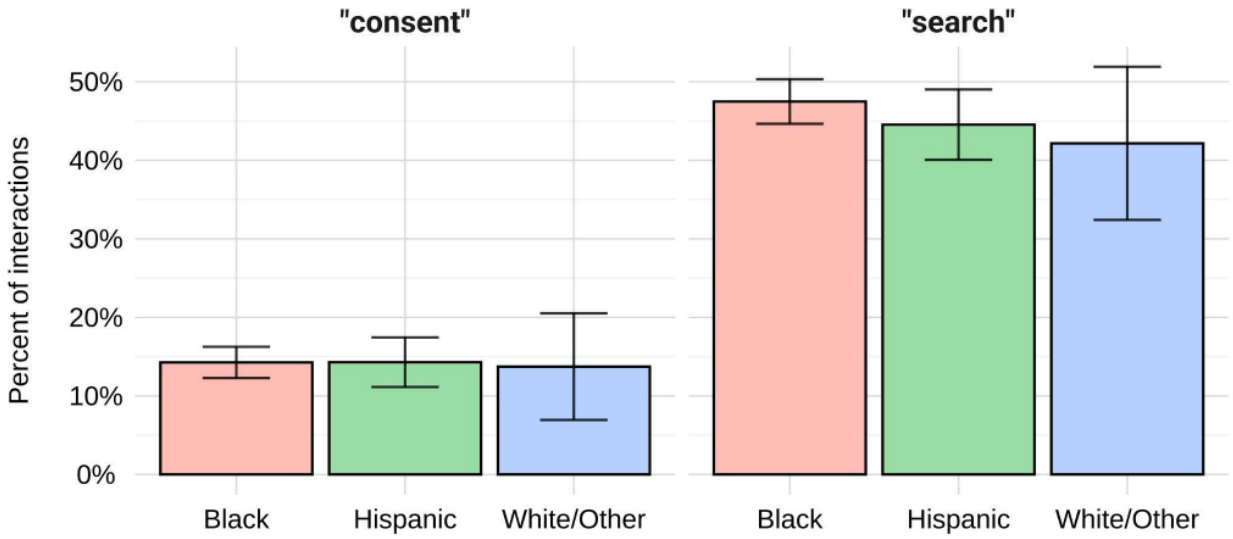
While we have discussed the limitations of auto-transcriptions, these limitations are particularly salient in analyzing the language of consent searches, in which law and policy considerations depend on both what was said and who said it. As noted, the speech-to-text transcription from Evidence.com, while reasonable at identifying words spoken in an interaction, performs very poorly at attributing those words to the proper speaker (the task of *diarization*).

Therefore, our first goal is to observe large-scale trends, where we aim to make the best possible use of the available data in the following ways. First, we operationalize measurement of explicit requests to search at a high level in a simple, transparent, and maximally conservative way: whether the explicit words "consent" or "search" appear in the transcript for any video associated with a consent search encounter. This measure does not require the words to be spoken by an officer, only that they appear somewhere in the transcript. We further estimate maximally conservative upper bounds for their occurrence by comparing the interaction-level auto-transcription hit rates for these particular words.

Officers documented 2,005 consent search requests on SQF forms during reported stops in 2023 and recorded civilian consent in 82.2% of those requests. In our sample of 1,770 stops with consent search requests, this proportion is even higher, with civilian consent recorded in 89.5% of available cases. However, explicit language related to consent searches is surprisingly rare in the data: "consent" is spoken in only 12.7% of interactions, and "search" is spoken in 46.0% of interactions. In Figure 9 we show these means broken down by race. The overall mean is dominated by consent searches with Black civilians, which represent two-thirds of the total interactions.

We test for race-based differences using logistic regression models predicting binary interaction-level occurrence of these explicit terms, incorporating control variables for gender of the civilian, borough, location (inside, outside, transit), whether the encounter ended in arrest, whether the background circumstances leading to the encounter involved violent crime, and the local crime rate, as measured by the count of arrests, shootings, and criminal summons in the preceding 30 days in the census block in which the encounter took place. When fitting regressions with a three-way distinction (Black, Hispanic, White/Other) we find no significant differences in the prevalence of these terms by race. However, since these encounters occur primarily with Black civilians, we also fit regressions comparing Black civilians to those of all other groups and find an estimated increase in the interaction-level probability of the term "search" by 5.1% (p=0.048).

Figure 9. Explicit mentions of "consent" and "search" in any recording associated with a consent search interaction, by civilian race.



A reasonable possible objection is that these findings may be substantially affected by the quality of transcription. To estimate the magnitude of this impact, we use the available human-transcribed data to estimate the word-level hit rate for these key words. The word "consent" appears in six interactions in the 341 human-transcribed encounters from the ISLG Sample; the corresponding automatic transcripts capture its occurrence in four, representing a hit rate of 66%. The word "search" appears in 93 human-transcribed interactions, and is captured by automatic transcripts in 80, representing a hit rate of 86%. These hit rates are substantively similar to what might be expected from the overall word error rate we identify of approximately 31 errors per 100 words, suggesting these words are not particularly more or less likely to be accurately transcribed than the average.

The number of occurrences is small, particularly for “consent,” which limits our ability to use these measures to place strict statistical bounds on the outcomes. Nevertheless, if we use these hit rates as a preliminary estimate for missed cases to scale the upper bound of the confidence interval for our measures, we find the maximally conservative estimates provided in Table 8. Even under these extremely conservative conditions–looking for mere mentions rather than full requests to search, allowing for any speaker to use the relevant word, in any associated video, and accounting for transcription errors–we find that out of all encounters containing consent search requests as documented by the NYPD, it is likely that at least 75% do not contain the word “consent” and at least 40% do not contain the word “search.”

Table 8. Maximal estimates for the upper bound of the proportion of consent search interactions containing mentions of key words related to explicit consent.

	Civilian Race		
Key word	Black (N = 1,192)	Hispanic (N = 476)	White/Other (N = 102)
“consent”	24.3%	26.3%	30.8%
“search”	58.9%	57.2%	60.3%

2. Findings by Predicted Officer vs. Civilian Speaker

We use human-transcribed encounters from the ISLG Sample to train models to classify officer speech in a procedure explained in detail in the Appendix (Section VIII.D.) and use these to estimate which portions of the speech signal in a transcript are spoken by officers as opposed to non-officers. Given the relatively poor diarization quality of the automatic transcripts, we are limited in our ability to make strong estimates by speaker, so we tune these models to have high precision at the cost of recall: when we predict officer speech we can be confident it is in fact spoken by an officer, but we may miss some instances. We then calculate estimates for the prevalence of an expanded set of relevant explicit key words broken down by the predicted speaker.

We augment the above findings with detailed information about an expanded set of phrases related to explicit consent and separated by the predicted speaker, including mentions of “frisk,” “permission,” confirmatory questions in the form of “Do you understand?”, given in Table 9. We find that mentions of frisks and permission are rare, but that a substantial minority of interactions do include confirmatory questions. Our estimate of interactions containing at least one of these five key terms remains below 60%. We also measure mentions of “don’t/not/no consent” to estimate the prevalence of the subset of consent mentions representing explicitly rejected consent (by civilians) or explicit mentions of the possibility to not consent (by officers).

Again, taking predicted speakers as preliminary, we do find that mentions of searches, consent, and permission are more commonly spoken by civilians compared to officers. We identify approximately 2.7% of interactions in which the phrases “don’t consent,” “not consent,” or “no consent” appear, virtually always spoken by predicted civilians. Looking in detail at this subset of interactions in which it appears the civilian did not consent, we find that they are documented by officers in the NYPD’s SQF data as “consent given” 51% of the time.

The NYPD Patrol Guide stipulates officers must clearly communicate that a civilian can refuse a consent search request but does not mandate specific words. In a consent search using the suggested phrasing from the NYPD Patrol Guide (*“I can only search you, if you consent. Do you understand? May I search you?”*) the interaction should contain the terms “consent,” “search,” and “you understand.” The bottom row of the table examines the prevalence of interactions in which all three key terms appear. We find that this occurs in only 3.2% of interactions agnostic to speaker, and 1.0% of interactions if we subset to utterances predicted to be spoken by officers. This estimate is still conservative for the prevalence of explicit requests since it does not place restrictions on the usage of these key phrases, only their occurrence. Ultimately these findings provide strong evidence that consent searches in which officers clearly and explicitly request consent in the manner suggested by the NYPD Patrol Guide are exceedingly rare.

Table 9. Interaction-level probability of occurrence of an expanded set of key terms related to explicit consent, broken out by predicted speaker.

	Predicted Speaker		
Key Term	Civilian	Officer	Any Speaker
“consent”	8.4%	7.4%	12.7%
“search”	33.6%	30.6%	46.0%
“frisk”	1.6%	3.7%	4.7%
“permission”	7.2%	4.4%	9.2%
“you understand”	11.6%	14.9%	20.8%
any of the above	38.1%	44.0%	58.8%
“don’t/not/no consent”	2.3%	0.0%	2.7%
“consent”, “search”, and “you understand” all occur	1.3%	1.0%	3.2%

3. The Context of "Consent" in BWC Transcripts

The above analyses do not account for the meanings in context of these key words, only their mere presence, in the interest of making conservative large-scale estimates given noisy transcripts. To supplement this analysis, we conduct a manual small-scale analysis of mentions of these terms. To do so we randomly sample 50 encounters in which the word "consent" appears and 50 in which the word "search" appears, identify the first instance of each in the transcript, and categorize how it is used. We make four key categorizations. First, we identify the actual speaker and the speech act the speaker intends to accomplish. Then, for requests by officers, we identify the key verb used and whether the request is followed by a confirmatory statement like "do you understand?" Our core findings for this analysis are given in Tables 10 and 11.

Table 10. Types and prevalence of in-context uses of explicit mentions of "consent" in 50 hand-labeled examples.

Speaker	Category	"Consent" is used…	Count
Officer 76% of cases	Request to Search	to request permission to search.	14
	Confirm Search	to confirm permission to search for a request initially asked in another way.	10
	Intra-Officer	to discuss the circumstances of a consent search with dispatch or other officers.	8
	Already Conducted	to reference a search that has already been conducted.	4
	Non-Search	regarding something other than the consent search in the interaction.	2
Civilian 24% of cases	Refusal	to explicitly refuse permission to be searched, either proactively or in response to an officer's request.	10
	Proactive Consent	to provide proactive consent to search before it is asked.	2

Among the 14 requests to search identified by use of the word "consent," we find associated verbs to be "search" (N=12), "check" (N=5), "give" (as in "give consent," N=4) and "mind" (N=1).

Table 11. Types and prevalence of in-context uses of explicit mentions of “search” in 50 hand-labeled examples.

Speaker	Category	“Search” is used…	Count
Officer 62% of cases	Intra-Officer	to discuss the circumstances of a consent search with dispatch or other officers.	12
	Request to Search	to request permission to search.	9
	Already Conducted	to reference a search that has already been conducted.	5
	Statement	to state that a search is occurring or will occur.	4
	Confirm Search	to confirm permission to search for a request initially asked in another way.	1
Civilian 38% of cases	Proactive Consent	to provide proactive consent to search before it is asked.	11
	Already Conducted	to reference a search that has already been conducted.	5
	Refusal	to explicitly refuse permission to be searched, either proactively or in response to an officer’s request.	2
	Transcription Error	due to an incorrect auto-transcription.	1

Among the 9 requests to search identified by use of the word “search,” we find associated verbs to be “mind” (N=4, one of which is the negated “[you] don’t mind”) and “check” (N=1), and four cases where “search” is the key verb framed with “permission” (N=1), “okay with” (N=1), and “consent” (N=2).

We find 3 out of the 14 direct requests to search using the word “consent” are followed by a confirmatory question like “do you understand?”, and none of the 9 requests to search using “search” are followed by such questions.

Taken together, these findings provide further detail for our large-scale observations. Namely, they suggest that roughly only half of the mentions of “consent” and less than half of the mentions of “search” identified in our large-scale estimates in the previous sub-section are likely to be genuine instances in which the key word is used to explicitly request or confirm consent to search.

Table 12. Examples of in-context uses of explicit mentions of “consent” by officers; automatic transcription with manually cleaned diarization (who is speaking).

Officer “Consent” Category	Example
Request to Search	OFFICER: How about you give us consent to search? CIVILIAN: I'm giving you consent to search. OFFICER: You consent to search?
	OFFICER: Do you have, do you gimme consent to search you? CIVILIAN: To search me? OFFICER: Yeah. Anything on you that, that can cut me? CIVILIAN: No. Why would you need to search me?
Confirm Search	OFFICER: You got anything on you? CIVILIAN: No. You can search me. OFFICER: I can search you. CIVILIAN: Yeah. OFFICER: So you giving me consent to search you?
	CIVILIAN: Please search my bag and not can get my bag so I can be on my way. OFFICER: You're giving me consent. Search your bag on? CIVILIAN: Yeah, you can search my bag.
Intra-Officer	OFFICER: Zero one. We have it. We haven't checked the trunk yet. He give us consent.
	OFFICER: Um, so we saw him, he match the description. So jumped out was just talking to him. He gave us consent to search him.
Already Conducted	OFFICER: No, that's fine. Just let me explain to you why I asked you to come out, okay? When I was looking at you from the other side, I saw something in your pocket. It looked like it could have been a weapon, okay? CIVILIAN: No ma'am. OFFICER: All right. So that's why I asked you to step out and that's why I asked for your consent to check.
Non-Search	OFFICER: This requires consent to prevent myself from outstanding information to knowingly misrepresent your name, date of birth, or address. All right? If you lie to me, it's an extra charge. That's all that means. All right? CIVILIAN: Mm-Hmm. OFFICER: Alright, so that being said, what's your first name?

Table 13. Examples of in-context uses of explicit mentions of “consent” by civilians; automatic transcription with manually cleaned diarization (who is speaking).

Community “Consent” Category	Examples
Refusal	CIVILIAN: What is the reason for your search officer? Officer? What is the reason for your search? OFFICER: You got your hands in your pocket and you got a box. CIVILIAN: All right. I have. All right. I have I can I explain? I don't. OFFICER: Okay. CIVILIAN: First of all, I do not consent to our search. Alright? I do not consent to you. You wanna search? OFFICER: Everything's on camera. CIVILIAN: You searched me without my consent.
	OFFICER: Okay, let me just check you real quick. Okay? Don't you don't reach in your pockets, yes or no. Yeah. Can I? CIVILIAN: No, I'm not. I give you no consent to search. No consent search. OFFICER: Okay. So what is that in there? Because you, you have something in here?
Proactive Consent	CIVILIAN: I'm giving you permission to search the vehicle. I'm, I'm consented to that. You gonna search the vehicle? I give you consent to search the vehicle. I, because y'all saying that somebody said they had a firearm. I – OFFICER: No, no, no. I understand.

B. Other Requests and Linguistic Context

Key Questions:

- How do officers request consent to search when they do not ask explicitly?
- How might such indirect or implicit requests be understood by civilians?
- What other linguistic behaviors occur during consent searches that are relevant to legal considerations of whether the civilian is free to decline?

The extremely low rates of consent requests that are consistent with NYPD guidelines raise the question of how officers actually ask for consent to search. Given the challenges in automatically detecting the precise moment and nature of a consent search request, we use a three-pronged strategy to answer this question. First, we perform manual content coding on searches in BWC recordings and the language officers use to make them. Second, we use this

information to construct pattern-matching heuristics to identify key types of other search requests at scale and measure their occurrence within the auto-transcribed 2023 Consent Search Sample to quantify their distribution at scale and qualitatively analyze them with reference to relevant linguistic literature. Third, we examine the case of commands as an element of linguistic context that potentially influences perceptions of whether civilians are free to decline a search, quantifying the presence of general commands and their distribution by race.

1. Categorization of Frisk- and Search-Associated Behavior in Human Transcripts

As our ability to analyze fine-grained linguistic behavior is limited by transcript quality, we examine a smaller, human-transcribed subset of interactions from the ISLG Sample. In their transcription guidelines, human transcribers were asked to annotate utterances which represented explicit consent to search or in which officers were visually conducting a search or frisk. Trained professional transcribers annotated 27 encounters with 33 instances of language surrounding consent searches and frisks out of the total 341 human-transcribed encounters. Of these, 1 was an arrest and 26 were stops, with 23 documented frisks and 25 documented searches. We note that these transcribers are not legal experts, so these annotations correspond to a layperson's perception of whether an officer is touching or physically appearing to conduct a frisk or search. Nevertheless, they provide a useful perspective on the linguistic contexts of searches; notably, none of the examples identified by trained transcribers contained the word "consent," and only one contains the word "search."

We manually examined and annotated each case for the manner in which consent was asked in order to understand the distribution of such requests. Table 14 presents the proportions of different categories of linguistic characteristics surrounding these events. For each of these characteristics we provide an illustrative example in Table 15 from the data to understand the linguistic context and real-world instantiation of these requests.

Table 14. Characteristics of implicit or indirect requests to search and their proportions in 33 instances of human-annotated data in the ISLG Sample.

Category	Description	Count	%
“anything”	Questions as to whether the civilian “has anything,” most commonly in the context of frisks and asking about sharp objects.	17	52%
statement	Direct statements that a search or frisk is occurring or will occur.	9	27%
tag question	Confirmatory questions such as “right?” and “yeah?”, often appearing following statements.	7	21%
“check”	Questions asking to “check” a person or item.	5	15%
“mind”	Questions asking if the person “minds” a search or a particular related behavior such as opening a bag.	3	9%
command	Verbal commands.	3	9%
nonverbal	Language co-occurring with a search or frisk yet does not reference a search or frisk.	2	6%

Table 15. Examples from human-transcribed data of language co-occurring with physical searches and frisks.

Request Characteristics	Examples
“anything,” tag question, “check”	OFFICER: Hold on. Do you have anything on you that we need to know about? CIVILIAN: Nah. OFFICER: Yes or no? So I could check you, right? Yes? CIVILIAN: Check me. OFFICER: Okay.
“anything,” command	OFFICER: All right. You got anything on you? CIVILIAN: No. OFFICER: No? CIVILIAN: (unintelligible). I told you, I don't have nothing on me. OFFICER: All right. Turn around. Turn around.
“anything,” statement	OFFICER: All right. We're gonna toss you one more time. Hold on. This is your last chance. You got anything I should know about? CIVILIAN: I said no.
“check,” tag question	OFFICER: Oh, yeah, actually, sorry, sorry. Just step out for one sec. Sorry. That's my fault. That's my fault. Lemme just check your pockets and shit. All right? CIVILIAN: I don't got no- nothing (unintelligible). OFFICER: No, no, I understand. You got a back pocket?
“check”	CIVILIAN: It's in the fanny pack. Check. OFFICER: I can check it? CIVILIAN: Mm-hmm. I give you permission. (unintelligible)
“mind”	OFFICER: You have the stuff in your bag? CIVILIAN: Yeah. OFFICER: You mind if I open your bag? CIVILIAN: I don't mind. OFFICER: All right.
“check,” “mind”	OFFICER: My man, mind if I just check you out real quick?
command	OFFICER: Go back against the wall -- go back against the wall. All right. Zip up your pants. Zip up your pants. Um --

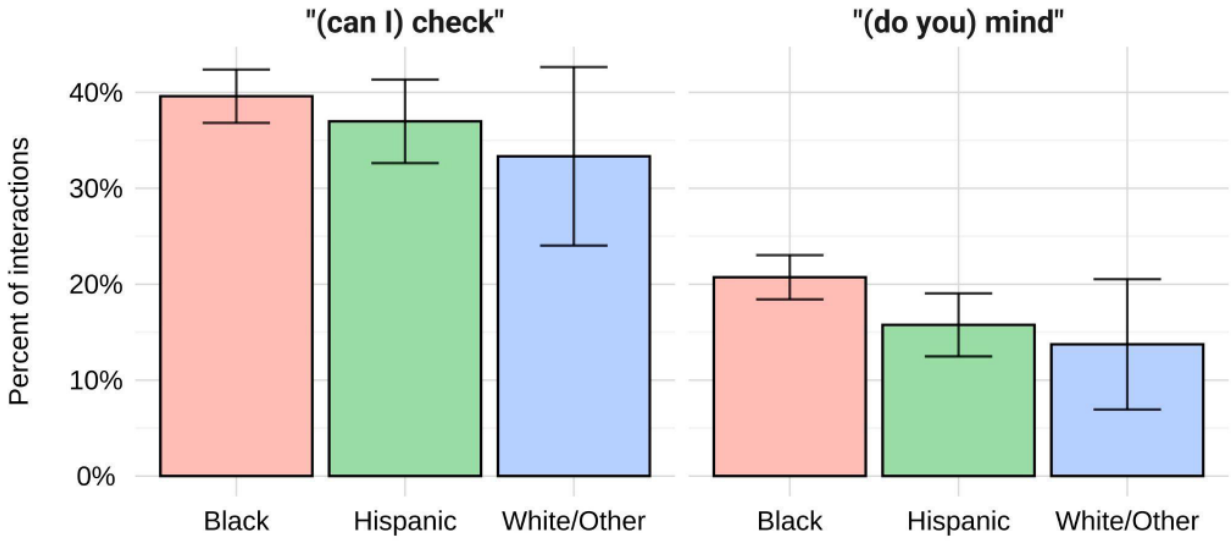
2. The Cases of “Mind” and “Check”

In our manual coding of requests, both associated with explicit language of consent (Section V.A.3.) and in human-transcribed data from the ISLG sample above, we identified “check” and “mind” as prevalent less-explicit ways that officers request consent to search. We therefore explore the particular cases of “(do you) mind” and “(can I) check” at scale further in the Consent Search Sample. Both words are multi-use and appear frequently in the data in other instances beyond requests to search. Therefore, we develop high-precision pattern-matching heuristics (detailed in Appendix VIII.F.) to identify instances in which these words are likely to be used to request consent to search and measure their prevalence across all 1,770 recordings in the Consent Search Sample.

Of all documented consent searches in 2023, we find “mind” requests appear in 16.8%, “check” requests appear in 36.7%, and one or the other appear in 41.8% of interactions. 16.9% of all interactions contain one or the other of these request types, but no mentions of “consent” or “search.” Figure 10 shows these estimates broken down across the sample by race of the civilian.

As with the explicit terms, we test for race-based differences using logistic regression models predicting binary interaction-level occurrence of “check” and “mind,” incorporating control variables for gender of the civilian, borough, location (inside, outside, transit), whether the encounter ended in arrest, whether the background circumstances leading to the encounter involved violent crime, and the local crime rate as measured by the count of arrests, shootings, and criminal summons in the preceding 30 days in the census block in which the encounter took place. Similar to the case of explicit terms, when fitting regressions with a three-way distinction (Black, Hispanic, White/Other) we find no significant differences in the prevalence of these terms by race, but in regressions comparing Black to all Other-race civilians we find an estimated increase in the interaction-level probability of “mind” requests by 5.4% ($p=<0.001$).

Figure 10. Estimated use of “check” and “mind” as requests to search in any recording associated with a consent search interaction, by civilian race.



Across interactions with civilians of all races, the substantial prevalence of each of these terms is concerning with regards to the clarity of the request to search, for independent reasons.

The use of “check” as a verb to describe a request to search introduces two related linguistic challenges for the listener, particularly in comparison to a more explicit request using the word “search.” First is that “check” minimizes the magnitude of the request. Whereas a “search” implies some degree of depth and intrusion, in common parlance a “check” is associated with a type of examination that is more cursory, brief, or from a distance. Second, “check” is ambiguous as to what specific behavior the consent is being requested for. Whereas police (and other public authorities with whom the public have frequent contact such as TSA officials) are known to “conduct searches” as a part of their job, and many laypersons are familiar with the concept of “being searched” and what that may entail, there is no corresponding clarity as to what specifically it might mean for a police officer to “check” something.

“Mind,” on the other hand, is problematic because it introduces inherent ambiguity as to the meaning of a response. From a linguistic perspective, we can note that “mind” itself includes some degree of semantic negation: if a person actively “minds” something in the sense of the question “yes, I do mind,” they dislike or disprefer that thing. To “mind” is to ask that something not take place, so one could say “yes” to mean a rejection of the search. On the other hand, a listener might commonly respond “yes” to illustrate their acceptance of the search.

The ambiguity in this case arises between what linguists call the “locutionary act” (the surface forms of the words spoken) and the “illocutionary act” (the action performed with those words,

such as making a request).[106] In phrasing a request for consent to search as, for example, "Do you mind if I search you?", the officer establishes conflicting interpretations of the meanings of "yes" and "no" for the listener.[107] The listener can respond to the locutionary, literal meaning of the words, in which case "yes" entails a rejection of the search; or they can respond to the underlying illocutionary action of a request to search, in which case "yes" entails an acceptance of the search. We illustrate these possible conflicting interpretations in Table 16; notice that the verbal polarity of the response and the concordant refusal or acceptance of the search are opposite under the alternative interpretations.

Table 16. Schematic of possible listener understandings of the meanings of a request to search phrased as "Do you mind if I search you?"

Proposition	Polarity of Response	Interpretation of Yes/No	Consent to Search
locutionary, surface "Do you mind if I search you?"	affirmative	Yes, I do mind.	Refused
	negative	No, I do not mind.	Accepted
illocutionary, underlying request Can I search you?	affirmative	Yes, you can search me.	Accepted
	negative	No, you can not search me.	Refused

[106] *See generally* J.L. Austin, How To Do Things With Words, *Harvard University Press* (1962); John R. Searle, A Taxonomy of Illocutionary Acts, in K. Gunderson, Language, Mind, and Knowledge, *University of Minnesota Press* 344-369 (1975).

[107] *See* Peter M. Tiersma & Lawrence M. Solan, Cops and Robbers: Selective Literalism in American Criminal Law, 38(2) *Law & society review* 229-266 (2004).

Table 17. Instances of civilian refusals to consent to search in response to requests phrased with “(do you) mind.”

Polarity of Verbal Response	Relevant Proposition	Example
negative	Responsive to Illocutionary Request to Search	OFFICER: You mind if I check you real quick? CIVILIAN: No, I don't got no fucking gun.
		OFFICER: Do you mind if I take a look? CIVILIAN: No, I don't want you to take a look.
		OFFICER: So you mind if I check because I let you guys a firearm in there? CIVILIAN: No, I'm good. OFFICER: You can back up my check. I'm gonna check open you bro. CIVILIAN: I just said don't check me.
mixed	Responsive to Both	OFFICER: Do you mind if I check that? CIVILIAN: No, no, no. I do mind. I do mind.
affirmative	Responsive to Locutionary Surface Form (“do you mind”)	OFFICER: Alright, so then you mind if we check? CIVILIAN: I mind. OFFICER: You do mind? CIVILIAN: Yes. OFFICER: Why? CIVILIAN: Yes. OFFICER: Why? CIVILIAN: Because I, I don't want to check. I don't want,
		OFFICER: Bro. I'm saying you don't have nothing in the vehicle? CIVILIAN: No, there's nothing in the vehicle. OFFICER: You don't mind if I search right? CIVILIAN: Yeah, I mind if you search 'cause I didn't do nothing.

This ambiguity is easy to confirm in the real world: looking at the data, we find many instances of civilians responding to both possible interpretations of the proposition presented by the officer. Examples of these conflicting interpretations are given in Table 17 above. In all examples, the civilian’s clear intent is to refuse the request to search; nevertheless, we see that when requests to search are phrased with “mind,” civilians can express that intent with negative, mixed, and affirmative responses depending on which proposition they interpreted as relevant in the original question. At times, this causes notable confusion between the officer and civilian

who, due to this ambiguity, may have differing interpretations as to whether consent to search has been given.

A substantial portion of requests involving "mind" are also negated (e.g., "you don't mind if I search you?") and even at times followed by a tag question like "right?" or "yeah?" (e.g., "you don't mind if I take a look, right?"). These types of negated and tagged questions introduce further layers of potential for confusion in the context of a request to search, and are known to linguists who study meaning (called semanticists) to introduce systematic ambiguities.[108] Moreover, experimental evidence[109] confirms the intuitive interpretation that these types of questions *presuppose* the outcome in which the civilian "does not mind" if the search takes place, and therefore may be unlikely to be understood as a genuine request.

A final angle is worth discussing in light of the prevalence of these types of less explicit requests in the data. Under a generous interpretation, from a social and behavioral perspective it is natural that officers might minimize requests to search using words like "check" and "mind." Politeness theory in linguistics focuses heavily on requests, and argues that people use politeness strategies in language to soften perceived threats to the "face"–social esteem or freedom of action–of the person of whom they are making a request.[110] These strategies can include indirectness, minimizing the imposition, and hedging, all of which reasonably characterize the types of requests made using "check" and "mind." Officers may understand that the request to search is a substantial one and avoid using direct and explicit language in order to attempt, naturally and with good-natured intent, to minimize the impact of this request on the listener. However, given the legal and policy implications that accompany requests for consent to search, officers should be made aware that softening such requests with indirectness adds unnecessary ambiguity and potential for misunderstanding that could ultimately lead to less policy-compliant and more challenging interactions with the public.[111]

3. Measuring the Co-Presence of Commands

To conclude this section, we briefly consider a last possible element of the linguistic context surrounding requests for consent to search, namely that of commands. While the presence of commands alone does not necessarily impact the voluntariness of a search, we argue that

---

[108] *See* D. Robert Ladd, A First Look at the Semantics and Pragmatics of Negative Questions and Tag Questions, In *Papers from the Regional Meeting of the Chicago Linguistic Society* 164-171 (1981).

[109] *See* Floris Roelofsen, Noortje Venhuizen & Galit Weidman Sassoon, Positive and Negative Polar Questions in Discourse, In *Proceedings of Sinn und Bedeutung* 17, 455-472 (2013).

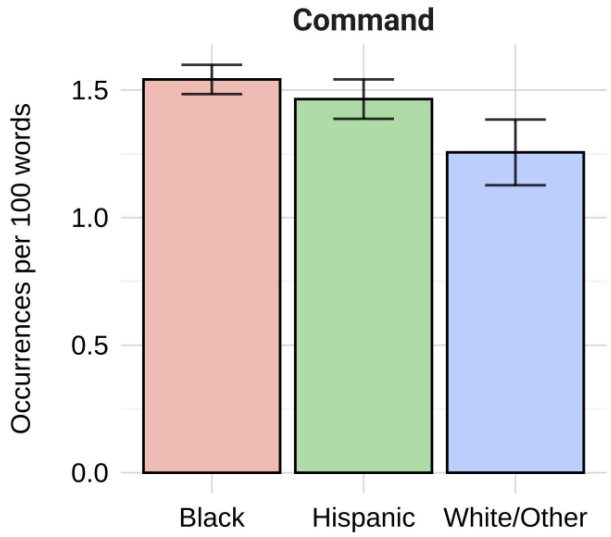
[110] *See e.g.,* Penelope Brown & Stephen C. Levinson, Politeness: Some Universals in Language Usage, *Cambridge University Press* (1987).

[111] Officers may also use indirect language to manage safety risks during an encounter. While voluntariness is assessed under a totality-of-the-circumstances framework rather than on a continuum, the more an officer's language obscures the nature of the consent request, the more it weighs against a finding of voluntary consent under the Fourth Amendment.

interactions in which commands are more prevalent may holistically create a linguistic context in which a civilian feels they have less agency and therefore are less free to decline.

We operationalize commands by identifying utterances starting with a bare verb like "wait," "stop," "stay," "hold [on]", and so on, and measure the prevalence of these commands as a proportion of all words spoken in the interaction. Figure 11 provides estimates of the occurrence of such commands in all consent searches broken down by civilian race. We test for race-based differences using a linear regression controlling for the same variables used to test explicit and implicit mentions (gender, borough, location, background circumstances of violent crime, and local crime rate), and find that Black and Hispanic civilians in these interactions hear more commands from officers, an additional 0.28 instances per 100 words for Black civilians (p=0.005) and an additional 0.21 instances for Hispanic civilians (p=0.042).

Figure 11. Prevalence of commands in documented consent search interactions by civilian race.



C. Summary

The Monitor's Twenty-Fifth Report expressed concern over NYPD consent search practices in traffic stops and noted Fourth and Fourteenth Amendment implications.[112] Such observations raise the question of how explicitly officers are requesting to search members of the public, and whether those members are free to grant or refuse that consent. Given the current performance of Evidence.com diarization, automatic analyses are necessarily tentative. However, even the most conservative detection of "consent" and "search"–one which identifies these words as spoken by anyone, in any associated video–finds that these terms appear very rarely in

[112] Twenty-Fifth Report of the Independent Monitor, *Floyd v. City of New York* at 19-20, No. 1:08-cv-01034-AT (S.D.N.Y. May 3, 2025), ECF No. 963-1.

documented consent search interactions, even accounting for automatic transcription error. Explicit requests to search appear to be the exception rather than the rule. A close manual reading of officers' language surrounding "consent" indicates they may be rarer still, as the word is not typically used during the request itself.

If NYPD officers are not posing search requests in a clear and explicit manner, how are they seeking consent? From the manually transcribed ISLG study recordings in which a consent search occurred, we identified the most frequent ways in which NYPD officers asked for consent, then searched for similar language in the auto-transcribed consent search recordings. These analyses identify two linguistic patterns which are directly tied to search voluntariness: minimizing the apparent imposition of the search by framing it as a "check," and using phrasing where a refusal may require an affirmative response ("do you mind?").

We propose that the use of "mind" in particular to phrase requests to search directly conflicts with the stated policy intention in the NYPD Patrol Guide that officers should "ask for consent to search in a manner that elicits a clear 'yes' or 'no' response." By asking a straightforward question such as "Can I search you?" or "Do I have your consent to search you?", this ambiguity can be entirely avoided. In these cases, "yes" both answers the surface form of the question and assents to a search, while "no" both negatively responds to the surface form and declines.

In addition to its implications for Fourth Amendment concerns around the voluntariness of civilian consent, our findings also raise potential Fourteenth Amendment issues of racial disparities in how NYPD officers elicit consent to search. Implicit search requests–and confusing framing–are most likely to occur in consent search requests of Black and Hispanic civilians and are markedly less common in requests of White civilians. Alongside our findings on "stop-like" interactions in the first aim, they suggest a more investigatory tenor to NYPD encounters with communities of color.

To be sure, there is much more work to be done analyzing consent search requests. The current gap between human and auto-transcripts demonstrates the need for more human-transcribed recordings in our Consent Search sample and alternatives to the lackluster diarization within Evidence.com, such as open-source models that can be fine-tuned for this specific domain and potentially achieve better performance. Another limitation is that our current sample only includes consent searches properly documented in stop reports, and not potential consent searches which were not recorded as such.

Our methods also demonstrate, however, how close readings of a small number of transcripts can be applied to searches of large sets of BWC recordings. They prove the feasibility of observing the nuances of consent searches, whereby differences in language have a profound effect on civilians' voluntariness, at the scale with which they occur. They further raise the prospect of using predictive models to identify consent requests, documented or otherwise, in a similar fashion as we have used predictive models to identify stops.

VI. DISCUSSION

A. Interpreting the Findings

The findings reported here should be understood and interpreted in the context of the Monitor's broader, continuing effort to assess the NYPD's compliance with constitutional principles. In the ISLG Report, retired judges who manually reviewed encounters found that a meaningful proportion of stops were improperly documented or unconstitutional.[113] The use of expert human review, however, is inherently difficult to scale. The pool of qualified reviewers is limited, and each encounter can demand significant time and judgment.[114] This study's machine learning models, which achieve 81.7% accuracy on broad-scale *De Bour* classification, offer a path to detecting the kinds of documentation and constitutional issues ISLG found across the full universe of NYPD police-civilian encounters, not just in samples.

These models also build upon the Monitor's work on underreporting. The Monitor's underreporting studies found that a significant number of encounters labeled as Level 2 were in fact Level 3 *Terry* stops, with compliance rates as low as 17% for stops labeled as Level 2 encounters in 2023.[115] Our finding that documented Level 1 and Level 2 encounters of Black and Hispanic individuals are linguistically more similar to Level 3 stops suggests a potential mechanism to detect this mislabeling. These encounters may be de facto detentions documented as lower-level encounters, and computational tools could help identify them.

Our findings also reinforce concerns identified in the Monitor's work on racial disparities. The MacDonald racial disparities analyses examined disparities in *who* was stopped using stop report data.[116] Our research extends that work by analyzing disparities in *how* those stops unfold in practice, including differences in officer language, rather than focusing on stop frequency alone. Additionally, a recent traffic stop audit by the Monitor found that officers rarely used explicit consent language when seeking to search a person.[117] Our analysis of a full year of consent search requests confirms the same pattern: only 46.0% mention "search," 12.7% mention "consent," and 20.8% include confirmatory questions.

[113] *See* ISLG Report.

[114] *See Id.* at 13 (explaining that the complexity of the law at issue in the ISLG study led to frequent disagreements among the initial two-judge panels, requiring review by a third, tie-breaking judge for a substantial number of encounters).

[115] Twenty-Second Report of the Independent Monitor at 7.

[116] *See* Thirteenth Report of the Independent Monitor; Fifth Report of the Independent Monitor, *Floyd v. City of New York*, No. 1:08-cv-01034 (AT) (S.D.N.Y. May 30, 2017), ECF No. 554.

[117] *See* Twenty-Fifth Report of the Independent Monitor at 16-21.

Together, the findings show that indicators of compliance with constitutional standards can be analyzed computationally and at scale. This study demonstrates that machine learning models can reliably distinguish *De Bour* encounter levels at rates substantially better than chance, that linguistic analysis can evaluate consent search practices, and that both approaches can detect racial disparities in officer conduct. The patterns identified in this analysis have direct implications for the Fourth and Fourteenth Amendment standards at the core of the *Floyd* remedial framework.

B. Fourth and Fourteenth Amendment Implications

Central to the *Floyd* liability findings is the Court's determination that the NYPD engaged in unconstitutional stops, violating the Fourth Amendment.[118] The Monitor team has closely examined the differences between Level 1/2 encounters and Level 3 stops because, unlike the lower-level interactions, Level 3 stops require reasonable suspicion, mirroring the Fourth Amendment's reasonable suspicion requirement for an investigatory seizure under *Terry*.[119]

Here, we demonstrated that machine learning models can distinguish between *De Bour* levels with strong accuracy, but we also found that under model estimates of stop probability, the NYPD's Level 1 and Level 2 encounters with Black and Hispanic individuals are linguistically more similar to Level 3 stops than encounters with White and Other-race individuals. This suggests that some documented Level 1 and Level 2 encounters involved conduct more consistent with detention, even though officers' documentation characterizes them as lower-level interactions in which detention was neither intended nor acknowledged. While these linguistic similarities do not, by themselves, establish that a particular encounter involved a Fourth Amendment seizure requiring reasonable suspicion, the pattern warrants attention.

Fourth Amendment analysis also bears directly on the lawfulness of consent searches. When an officer relies on consent as the recognized exception to the Fourth Amendment's warrant requirement, that consent must be voluntary, with courts looking to the totality of the circumstances to determine whether it was freely given.[120] New York's *De Bour* framework generally requires that police requesting consent have at least a "founded suspicion" of criminality, a standard that requires more than a hunch but less than reasonable suspicion, based on observable conduct or reliable information.[121] In New York City, the Right to Know Act—and the Department's guidance implementing it—further requires officers to explain a

[118] *Floyd* Liability Opinion at 561-563.

[119] *See De Bour*, 40 N.Y.2d 210; *Terry*, 392 U.S. 1.

[120] *See Schneckloth*, 412 U.S. 218, 227.

[121] *See De Bour*, 40 N.Y.2d 210.

person's right to refuse a search and affirm that the person understands they are being asked to give permission.[122]

Analyzing a full year of consent requests documented on stop reports, we found that NYPD officers inconsistently use key words and phrases associated with the explicit language of consent. The word "search" appears in less than half of consent search encounters (46%), "consent" in less than 13%. Indirect or implicit requests are more common than explicit requests, and "yes" or "no" questions phrased in the negative—"You don't mind if I take a look?"—are common as well. This kind of phrasing can create ambiguity about whether individuals understand that they are being asked to consent, that they can refuse, and what they are consenting to. Those questions map onto factors courts weigh in assessing whether consent was voluntary under the Fourth Amendment and *Schneckloth's* totality of the circumstances approach.[123]

The Monitor is also charged with assessing whether there are racial disparities in the NYPD's stop, question, and frisk practices, addressing the Department's "policy of indirect racial profiling" the Court found in violation of the Equal Protection Clause of the Fourteenth Amendment.[124] This study reveals racial disparities, in that the language used by officers both in lower-level encounters as well as stops of Black civilians is more "stop-like" and suggestive of detention. Moreover, in the consent search context, even where rates of explicit consent language are similarly low across groups, we find implicit requests with "mind" and the use of commands are more prevalent in consent search encounters with Black and Hispanic civilians. Together, these findings extend the Monitor's prior analyses, which showed racial disparities in how often stops occur, by demonstrating disparities in how encounters unfold in practice. These patterns point to differential treatment reflected in officer language and, while they do not establish discriminatory intent, they are consistent with the concerns about racially disparate policing that animate the Court's remedial orders.

C. Applications for NYPD Operations

This study's findings suggest practical applications across several areas of NYPD operations relevant to the Court's remedial orders. The computational methods and linguistic patterns identified here could be used to enhance officer training, strengthen supervisory tools and BWC review processes, inform EIP threshold considerations, and guide policy development around consent search protocols and documentation requirements. More broadly, this work demonstrates the potential for integrating AI-powered analytical capabilities into existing Department systems—from ComplianceStat audits to Quality Assurance Section reviews—in

[122] N.Y.C. Admin. Code § 14-173(a)(1), (3).

[123] *See Schneckloth*, 412 U.S. 218, 227.

[124] *Floyd* Liability Opinion at 562.

ways that help reviewers find potential problems faster and more consistently, while preserving human oversight at every step.

Officers' ability to conduct consent searches and accurately document investigative encounters in compliance with the Fourth Amendment, the *De Bour* framework, and the Right to Know Act depends in no small part on how they are trained. Our findings suggest the NYPD's consent search curriculum should re-emphasize its instruction to officers to use explicit language when requesting consent, avoid ambiguous sentence constructions, and ask confirmatory questions to ensure the voluntariness of any consent obtained. The encounter classification findings could similarly inform training aimed at ensuring alignment between the documented level of an encounter and actual officer conduct, with potential for integration into the Court-approved training videos developed collaboratively with the Monitor team.[125]

This study's findings could also be used to strengthen the NYPD's supervisory tools and review processes. According to the Department, the Patrol Services Bureau already samples BWC videos for review from categories such as "Level 1 encounter" and "Level 2 encounter," and our model's ability to assign probability scores indicating a higher likelihood that certain videos are actually Level 3 encounters, based on linguistic features in officer speech, could allow supervisors to prioritize reviewing those videos rather than relying solely on random sampling.[126] ComplianceStat offers particularly promising opportunities for integrating this research. We have demonstrated how our models identify potential undocumented *Terry* stops and consent search deficiencies, issues that are a focus of the Patrol Services Bureau's pre-meeting audits and ComplianceStat sessions themselves.[127]

Recent reforms to the NYPD's EIP emphasize supervisor accountability and tracking interventions to assess whether they succeeded.[128] Yet as the Monitor has observed, reliably measuring whether interventions actually change officer behavior remains a significant challenge.[129] Our models identify patterns in officer language that correlate with problematic encounters: command-oriented rather than request-oriented framing, absence of explicit consent language, and markers associated with escalation. These linguistic patterns could serve as leading indicators for the EIP, identifying concerning trends in an officer's BWC footage before those patterns result in a suppression decision following an unlawful stop or a Civilian Complaint Review Board ("CCRB") complaint. The May 2025 EIP reforms introduced a six-month look-back period to evaluate intervention outcomes, and computational analysis could provide

[125] *See* Training Videos, NYPD Monitor, https://www.nypdmonitor.org/training-videos/ (last visited Dec. 23, 2025); Twenty-First Report of the Independent Monitor at 30-31.

[126] *See* Twenty-First Report of the Independent Monitor at 20.

[127] *Id.* at 35.

[128] Twenty-Sixth Report of the Independent Monitor at 15-16.

[129] *See Id.* at 4, 15-16.

objective before-and-after measures such as whether escalation patterns decreased in an officer's speech or whether consent search language became more explicit.[130]

Accurate documentation is essential for the NYPD's supervisory and accountability system—from Patrol Guide encounter review requirements to Quality Assurance Section audits, ComplianceStat monitoring, and EIP thresholds. Our findings reveal a gap between legal requirements, such as the Fourth Amendment and Right to Know Act's expectations for how officers communicate and confirm consent to search, and officers' practices in the field. Targeted updates to Patrol Guide guidance could operationalize what compliant consent search requests sound like. Similarly, policy guidance could help officers and supervisors recognize language indicating whether a Level 3 encounter has occurred, potentially narrowing the gap between how officers label encounters and how they function in practice. Rather than impose new obligations, these revisions could provide officers with clearer guidance on how to implement standards they are already required to meet.

D. Study Limitations

While this study identifies several significant patterns in the NYPD's handling of investigatory encounters that are relevant to assessing compliance with Fourth and Fourteenth Amendment standards, it is subject to inherent limitations. The character of a police-civilian interaction can turn on precise language and sequencing.[131] Encounter classification and voluntariness also depend on non-linguistic factors such as officers' tactical positioning, the number of officers present, and whether a person has a clear path to leave, none of which are captured in a text-based analysis.

Analyzing these encounters demands a high degree of accuracy.[132] For this study, analyses are limited to what is captured on the recording. To the extent that BWC audio is activated after an encounter has begun, the earliest moments would not be reflected in the transcripts. BWC recordings were processed through Axon's auto-transcription service and manual editing of these transcripts was conducted solely within Axon's platform, constraints that affected transcript quality, particularly in identifying *who* said *what*. In addition, while the linguistic patterns identified in the transcripts can document differences in how encounters unfold across groups, they do not, on their own, establish officers' intent. For example, explicit language that clearly signals discriminatory intent is rarely captured on video.

Research using BWCs comes with inherent limitations. Our analyses for consent searches use all available footage to estimate whether critical language occurred in any associated video. For the ISLG data, on the other hand, we use the single "verbal recording" taken by the lead officer

[130] *See Id.* at 15-16.

[131] *See generally* Rho et al., *supra* note 65; Prabhakaran et al., *supra* note 64.

[132] *See generally* Voigt et al., *supra* note 62.

as the single source of information, because integrating information from multiple videos in a predictive task is inherently challenging. We note that this means that reported predictive performance on the ISLG sample is in some sense a conservative estimate, since other associated videos may contain relevant information. Integrating information across multiple videos of the same encounter is particularly difficult given the quality of auto-transcription and diarization, but even with this limitation, doing so is a promising avenue for improving predictive performance in the future.

Our findings are also shaped by the data available for this study. The *De Bour* classification analysis drew on a set of 2,858 encounter recordings from 2022-2024, while the consent search analysis examined one year of data from 2023. As discussed in Section III, shifts in *Terry* stop practices and new compliance efforts undertaken in recent years—including the launch of ComplianceStat in January 2024—may not be fully reflected in these findings. Other data characteristics affected certain analyses. For instance, the low number of consent searches of White individuals limits our statistical power to quantify the magnitude of differences in how the NYPD approaches these searches across racial groups; the fact that even within these constraints we identify some substantial differences, such as the greater use of "mind" requests with Black civilians and more frequent contextual commands with Black and Hispanic civilians, suggests the need for increased attention in this area.

But these data constraints are both a limitation and a significant opportunity. The work presented here demonstrates that key indicators of constitutional compliance can be analyzed computationally, across a large number of encounters. With access to more recent, more comprehensive data, these methods could support the Court, the Monitor, and the Department in evaluating the impact of ongoing efforts to bring the NYPD's policing practices into constitutional compliance.

E. Expanding Computational Monitoring

The methodology developed in this study establishes a foundation for expanded computational monitoring. The Stanford-affiliated research team has spent the past several years developing and refining this approach with two large cities in California, where police departments have shared full access to years of comprehensive BWC footage for analysis. That work has already produced concrete results: machine learning models that can automatically detect policy compliance, evaluations of trainings, and new insights into police-community interactions citywide. In total, this work has involved analyzing over 1.3 million videos representing more than 300,000 hours of interactions from thousands of officers.

Building on the collaboration between the Stanford team, the Department, and the Monitor over the past several years, the approach demonstrated in this report could be extended to analyze NYPD BWC footage more comprehensively. With broader access to footage, machine learning models could automatically identify footage that has been labeled by an officer as a Level 2 encounter but shares features that are more similar to a Level 3 stop. Although the misidentification of Level 3 stops as Level 2 encounters is quite consequential, it is also

relatively infrequent, which means that automating the process would save a substantial amount of time and resources. In addition, large-scale systematic assessments over time could determine whether (and how quickly) compliance is improving.

Machine learning models could examine large numbers of consent searches to flag potential compliance concerns. Such models could inform the Department of the exact manner in which officers request consent (e.g., directly or indirectly), how members of the public respond to such requests, and how those requests and responses may differ as a function of the race of the individual. More broadly, computational tools could supplement existing auditing functions. Models could present supervisors with footage of "edge cases"—cases that are on the border of Level 2 and Level 3—so that supervisor expertise is directed toward those interactions that would most benefit from review. Compared to random audits of footage, this approach would maximize supervisors' valuable time while reducing the time and resources required to achieve the same level of oversight.

The same classification approach could extend to other aspects of police-civilian encounters. Just as machine learning models can distinguish encounters that are likely Level 3 stops from those that are Level 2 interactions, similar models might be able to identify encounters that share linguistic or procedural features that have generated civilian complaints. This capability could complement the EIP, which the Court ordered to identify at-risk behavior before misconduct escalates.[133] Currently, the EIP is triggered only after complaints are filed, adverse findings are made, or other at-risk behavioral thresholds are reached.[134] AI analysis of BWC footage, implemented intentionally, could allow the Department to proactively identify patterns associated with encounters that generate complaints, supporting early intervention and more targeted supervisory review.[135] Additional data sources could further strengthen these capabilities. CCRB investigators already review BWC footage in their process, generating a potential source of labeled data, and civilian video collected through the Community Liaison could enrich these analyses as well.

Computational analysis also offers possibilities for evaluating training effectiveness. In California, the research team evaluated a procedural justice training by examining officers' footage before the training and again up to four weeks after. Post-training, officers were more likely to state the reason for the stop when they pulled drivers over and more likely to express concern for their safety.[136] A computational approach to training evaluation could be applied to

---

[133] *Floyd v. City of New York*, No. 1:08-cv-01034-AT (S.D.N.Y. June 2, 2020), ECF No. 767.

[134] Twenty-First Report of the Independent Monitor at 37-38.

[135] AI, like any technology or metric, must be deployed with care in organizations, to avoid adverse outcomes such as exploiting proximal assessments instead of primary outcomes. *See* Camp & Voigt, *supra* note 61.

[136] *See* Camp et al., *supra* note 67.

a wide variety of trainings. Going beyond commonly assessed elements like trainees' knowledge of the concepts taught or their enjoyment of the training, before-and-after analysis of BWC footage could determine whether a training actually achieved its desired outcomes during real encounters. If a training is effective, analysis could also determine for how long, providing data-informed guidance on the recommended cadence of training.

Finally, expanded computational monitoring could be designed with long-term sustainability in mind. Tools could be developed with the explicit intention of integrating them into the Department's existing workflows and operations. This approach would ensure that the NYPD could operate these analytical tools independently, supporting long-term sustainability and continuous improvement without ongoing dependence on external researchers.

VII. CONCLUSION

This report demonstrates that computational analysis of BWC footage can meaningfully strengthen constitutional compliance monitoring at the scale urban policing requires. The NYPD records millions of investigative encounter videos each year, yet quarterly Monitor audits and ComplianceStat meetings review only a few hundred BWC videos. As a result, even sustained monitoring efforts necessarily capture only a small fraction of officer-civilian interactions, leaving important compliance questions difficult to answer at Department scale.

The findings presented here highlight both immediate areas of concern and the broader potential of AI-powered computational approaches. Analysis of stops involving consent searches throughout 2023 shows that officers rarely use explicit consent language: the word "consent" appears in only 12.7% of encounters and "search" in just 46.0%. These patterns raise concerns under Fourth Amendment requirements and NYPD Patrol Guide policy and underscore the extent to which the constitutional validity of a consent search can depend on what officers actually say during interactions.

At the same time, the results demonstrate how computational tools can help systematically identify critical issues. Machine learning models distinguish low-level encounters from stops with strong performance and produce calibrated probability estimates that could support more targeted auditing. These findings suggest a path toward identifying under-documented stops that would be difficult to detect through traditional sampling alone.

Beyond documentation gaps, the analysis identifies patterns in how encounters are experienced. Encounters involving Black and Hispanic civilians that are documented as low-level encounters more often resemble detentions in their language, and even during consent searches—interactions that should be voluntary—Black and Hispanic civilians experience higher rates of command language than White civilians. These patterns suggest that disparities may arise not only in whether stops occur in the first place, but in how they are experienced and communicated once they do.

Importantly, these findings were produced using automatic transcripts and without direct access to underlying audio or video files. Improvements in transcription accuracy, direct audio processing, and broader data access would allow for more refined analysis while preserving the scalability demonstrated in the Stanford team's work analyzing more than 1.3 million videos in other jurisdictions.

Looking ahead, computational monitoring should be understood as a complement to—not a replacement for—human judgment. When paired with legal expertise and supervisory review, these tools could help focus attention on encounters that warrant closer scrutiny, assess whether training is changing officer behavior in practice, and track compliance trends over time.

Used thoughtfully, they offer a way to move beyond small-sample audits toward more systematic oversight.

The Court's 2021 approval of this study made it possible to assess whether these methods are feasible and informative in the *Floyd* monitoring context. The results suggest they are. With appropriate safeguards, collaboration, and infrastructure, computational analysis of BWC footage could help shift oversight from periodic sampling toward more frequent and systematic assessment—supporting earlier identification of problems, clearer measurement of reform progress, and the kind of accountability that strengthens both constitutional compliance and public trust.

VIII. APPENDICES

In this section we provide additional technical details supporting the analyses described throughout the report.

A. Automatic Transcription

Automatic transcripts used in all analyses were produced in a multi-step process.

First, BWC footage for the encounters in each sample was transferred by the NYPD to a separate instance of Evidence.com for use in the Monitor's studies. Administrative records for the Monitor-Assessed Sample and Consent Search Sample and encounter data from the ISLG Sample were linked to evidence serial numbers corresponding to those pieces of footage.

Automatic transcripts for each piece of footage were then generated via Axon's Application Programming Interface (API). This involved members of the study team receiving permissioned credentials, which allowed the sending of requests to Evidence.com to generate transcriptions. The resulting automatic transcripts and metadata associated with each video were downloaded to encrypted computers owned by members of the study team for analysis. Transcripts come in the form of JSON dictionaries segmented into turns with turn-level diarization and word-level timing information.

B. Human Transcription

The human transcriptions used in this report were generated by post-editing of automatic transcripts by professional transcribers from the transcription company Acolad[137] within the Evidence.com interface. This represented an unusual workflow since these transcribers generally perform their work in dedicated software in which they are expert. We arrived at this arrangement after extended discussions with representatives of the Department in which we were unable to come to an agreement regarding data security of audio and video data.

Members of the study team set up a workflow using the Evidence.com API to manage transcribers' access to footage, releasing footage to individual transcribers in batches of 20 videos at a time; when a transcriber indicated a batch was complete, access for those 20 videos was revoked and a new 20 videos were assigned. We helped transcribers troubleshoot challenges transcribing within Evidence.com, including the use of foot pedals which are standard in professional transcription. Even so, some transcribers were ultimately unable to

[137] *See* Acolad, Professional Transcription Services, ACOLAD, https://www.acolad.com/en/services/transcription.

access and work within Evidence.com; those who did reported that the task was very challenging and many chose to drop off the project.

Our initial goal was to perform post-editing on most or all videos included in the sample. These challenges made this goal infeasible in a reasonable time frame. Ultimately professional transcribers were able to post-edit 341 transcripts, which were used for analysis as well as model training and evaluation.

C. Evaluation of Automatic Transcription

We evaluated the automatic transcriptions generated using Evidence.com by comparison to the 341 human-transcribed pieces of footage we were able to obtain, under three key categories of performance.

1. Word-Level Transcription Performance

We first evaluated word-level transcription performance using the standard metric of word error rate (WER).[138] This metric provides a measure of how accurate the automatic transcriptions are at appropriately transcribing the words that were spoken in a piece of footage, independent of who spoke them. WER is calculated as the sum of word-level substitutions, insertions, and deletions in an automatic transcript, divided by the total number of words in the reference human transcript.

We find a mean WER of 0.31 and a median of 0.27. This is roughly interpretable to mean that on average, given 100 words in a reference human transcript, we expect that 31 of those words will be incorrect in the corresponding automatic transcript. This level of performance roughly aligns with that measured in Field et al. (2023) on traffic stop data.[139] We know from this existing work that we can expect higher WER for officer speech compared to civilian speech, so in general we should expect that any results reflecting officer speech that we present are relatively more robust. Overall, this level of word-level Automatic Speech Recognition (ASR) performance is usable for large-scale analysis but imposes limitations on fine-grained analysis.

2. Diarization Performance

Diarization refers to the speech processing task of identifying who is speaking when in a transcript. This includes the clustering of distant turns of speech and identifying that, for instance, they were spoken by the same or different speakers. We evaluated diarization

[138] *See* Daniel Jurafsky & James H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models, ch. 15.6 (3d ed., online manuscript released Jan. 6, 2026).

[139] *See* Field et al., supra note 66.

performance using the standard metric of diarization error rate for automatic transcription averaged across the 341 human-corrected transcripts and also reported more fine-grained results including diarization purity and coverage. These metrics provide measures of how well the automatic transcripts capture who was speaking which words. DER is calculated in terms of ratios of the total time containing errors, using the standard *pyannote* metrics package.[140]

Table 18. Average ASR diarization performance relative to human-corrected transcription.

Category	Metric	Score
Diarization Error Rate (DER) Overall Scores	Mean	0.706
	Median	0.649
Diarization Error Rate (DER) Sub-Components	Missed Detection	0.172
	False Alarm	0.281
	Correct	0.575
	Confusion	0.254

We find poor diarization performance, at a median error score of 0.649 out of 1, where higher is worse. We further calculated average diarization coverage (0.633) and purity (0.604); since both metrics are high, we cannot identify that poor performance is primarily due to one of over-segmentation or under-segmentation but rather a mix of both. We would characterize this level of performance as largely unusable. As described below, we attempt to use it where possible with an additional layer of officer identification prediction.

3. Impact on Predictive Performance

The above two performance metrics represent *intrinsic* measures of ASR performance since they directly evaluate the performance of the system internally. We can also evaluate ASR performance *extrinsically* by asking if human transcription relative to ASR has an impact on predictive performance. To do this, we evaluated our predictive performance at Task 3: *De Bour* Classification (Section IV.A.) within the subset of human-corrected transcripts versus the remaining automatic transcripts. We find the model achieves an accuracy of 87.6% on human transcripts, compared to 81.2% on automatic transcripts, representing an improvement of 6.4%

[140] *See* pyannote.metrics documentation, https://pyannote.github.io/pyannote-metrics/reference.html.

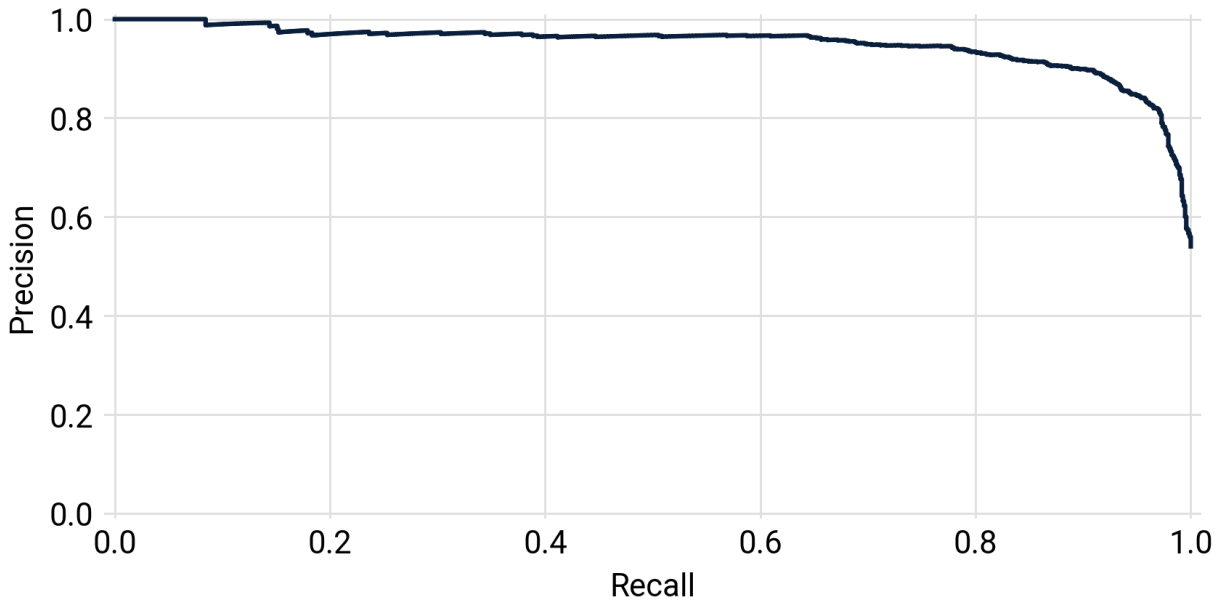
and a substantial error reduction of 34.0%. This leads us to believe that improving the quality of automatic transcription is a very promising avenue moving forward as we seek to build better predictive models as critical monitoring tools.

D. Officer Identification

In spite of the observed poor diarization performance from automatic transcription, we aimed to report some degree of findings regarding linguistic behavior specific to officers vs. civilians. To do this we used the 341 available transcripts with human-corrected speaker labels as a source of training data for a simple model aiming to predict, for a set of turns, whether the speaker is an officer or civilian. We featurized all the speech for a given human-corrected speaker with a TF-IDF (Term Frequency-Inverse Document Frequency) transformed bag-of-words approach and applied a binarized bigram SVM to classify the speaker as an officer or civilian.

Observing precision-recall curves for the trained models (see Figure 12), we found that high precision in this task was possible on human transcripts while maintaining moderate recall and tuned our prediction threshold to optimize for this setting. Our ultimate primary evaluation on these transcripts resulted in a precision of 94.7%, a recall of 72.3%, and an F1 score of 82.0%.

Figure 12. Precision-recall curve for officer identification task trained on human-corrected transcripts.



However, an additional complication in the actual application of this model is that the very poor diarization performance in the automatic transcripts could lead to poor performance in that setting. Therefore, we conducted a secondary evaluation of the performance of this model on automatic transcripts, concurrent with our manual labeling of “consent” and “search” instances

(Section V.C.). For each of these 100 instances we manually identified the speaker as an officer or civilian and evaluated the performance of automatic officer identification against these human labels. In this setting we found reduced, but still reasonable performance: precision of 91.3%, recall of 60.9%, and F1 score of 73.0%. This is likely reflective of the level of actual performance we can expect in the full dataset; that is, we can be confident that identified officer speech was in fact spoken by officers, at the cost of some recall such that some amount of officer speech may be missed. We applied these predicted labels to all transcripts and used them in some analyses as described in the body of the report; most analyses, however, were done with reference to any words spoken in transcripts and therefore ignoring predicted speakers.

E. Predictive Modeling

Importantly we note that all computational analysis took place locally, using open-source models, on encrypted computers belonging to members of the study team. No data was submitted to external or cloud-based AI services.

For statistical machine learning models (XGBoost and Linear SVM), analyses were predominantly performed using the *scikit-learn* library.[141] We transformed the input data with a tf-idf bag-of-words representation including unigrams, bigrams, and trigrams, limited to the 5,000 most common features by term frequency. For the Linear SVM we experimented with Bayesian hyperparameter optimization for regularization and alternative kernels but ultimately found it to have little impact, so results reported are with library default settings.

For LLM-based models, analyses were predominantly performed using the HuggingFace *transformers* library.[142] We experimented with alternative data representations, including casing and presenting speech in transcription format with prepended speaker names (e.g., “OFFICER: Can I search you? COMM: Yeah, go ahead.”), but found that these each reduced performance relative to directly passing ASR transcription content. All models were trained on a single Nvidia RTX 5090 with the AdamW optimizer. DistilBERT was trained for 4 epochs and a batch size of 64, ModernBERT for 6 epochs and a batch size of 2 with 16 gradient accumulation steps.

F. Pattern-Matching for Consent Searches

In Section V.B.2. we describe analyses contingent upon pattern-matching of request-like instances of “check” and “mind,” which we operationalize as regular expressions.

---

[141] *See* Fabian Pedregosa et al., Scikit-learn: Machine Learning in Python, 12 *Journal of Machine Learning Research* 2825–2830 (2011).

[142] *See* Thomas Wolf et al., Huggingface’s Transformers: State-of-the-art Natural Language Processing, arXiv preprint arXiv:1910.03771 (2019); Transformers, Hugging Face at https://huggingface.co/docs/transformers/en/index.

For “check” we required that it either be preceded by an instance of “mind,” as in “do you mind if I check,” or preceded by a first-person pronoun (“I” or “we”) and followed by a pronoun, noun, or preposition characterizing the thing or direction to be checked, in the following list: “it,” “you,” “in,” “on,” “inside,” “that,” “pocket,” “pants,” “shirt,” “jacket,” “waistband.”

For “mind” we require an instance of “mind” to be followed in the same utterance by a verb in the following list: “search,” “check,” “look,” “see,” and “open.”

For both “check” and “mind,” we checked that these are high-precision patterns, with all instances taking the form of requests in a manually reviewed sample of 20 cases each. We found that other instances not followed by these specific search-related terms were far less likely to take the form of requests, although these more restrictive patterns may miss some real cases where “check” or “mind” are used more ambiguously or indirectly.