Not Your Typical Government Tipline: LLM-Assisted Routing of Environmental Protection Agency Citizen Tips

Sharanya Majumder[1,*] **Zehua Li**[1,2,*] **Derek Ouyang**[1] **Kit Rodolfa**[1]
Julian Nyarko[1] **Elena Eneva**[1] **Daniel E. Ho**[1]

[1] Stanford University [2] Harvard University

Abstract

Regulatory agencies often operate with limited resources and rely on tips from the public to identify potential violations. However, processing these tips at scale presents significant operational challenges, as agencies must correctly identify and route relevant tips to the appropriate enforcement divisions. Through a case study, we demonstrate how advances in large language models can be utilized to support overburdened agencies with limited capacities. In partnership with the U.S. Environmental Protection Agency, we leverage previously unstudied citizen tips data from their "Report a Violation" system to develop an LLM-assisted pipeline for tip routing. Our approach filters out 80.5% of irrelevant tips and increases overall routing accuracy from 31.8% to 82.4% compared to the current routing system. At a time of increased focus on government efficiencies, our approach provides a constructive path forward by using technology to empower civil servants.

1 Introduction

Regulatory agencies guard society against threats to public health and safety. With limited enforcement resources, they often rely on tips from the public to identify on-the-ground regulatory violations. However, the volume of tips these agencies receive imposes a significant administrative burden.

The U.S. Environmental Protection Agency (EPA), for instance, receives thousands of citizen tips annually. Any incoming tip is routed heuristically, based on submitter inputs to structured fields regarding violation characteristics, to either the Civil or Criminal Division, the latter of which handles serious environmental crimes and intentional violations. In practice, this heuristic routing creates substantial inefficiencies by failing to filter out tips not falling under the EPA's jurisdiction and by frequently misrouting civil matters to the Criminal Division, resulting in wasted effort on irrelevant tips, delayed civil enforcement actions, and diverted Criminal Division resources from investigating serious environmental crimes. Despite its limitations, the tipline generates roughly a thousand open criminal leads every year.

At a time of heightened focus on government efficiencies, we provide a constructive path forward by demonstrating how advances in large language models (LLMs) can streamline tip routing within regulatory agencies and empower civil servants. Building on our long-term collaboration with the EPA[1], we analyze thousands of previously unstudied citizen tips from the agency's tipline system. Leveraging agency feedback, we develop a two-stage LLM-assisted classification pipeline that first excludes tips outside the agency's scope and then routes relevant ones to the appropriate division. We show that this approach automatically filters 80.5% of non-actionable tips and increases overall routing accuracy from 31.8% to 82.4%, in turn reducing unnecessary workload and re-routing.

2 Related Work

Effective regulatory enforcement requires not only well-defined legal frameworks but also institutional capacity to detect and respond to violations. Agencies often struggle to monitor compliance at scale due to resource constraints, oversight, and fragmented authority (Stephenson, 2006; Biber, 2009; Shimshack, 2014; Short, 2021). This is especially acute in environmental regulation, where most violations go undetected and disproportionately affect vulnerable communities (Gray and Shimshack,

[*] Equal contribution.

[1] This research was conducted through a Cooperative Research and Development Agreement (CRADA) with the U.S. Environmental Protection Agency (EPA). This data was provided by and belongs to the EPA. Any further use of this data must be approved by the EPA. Points of view or opinions contained within this document are those of the author and do not necessarily represent the official position or policies of the United States Environmental Protection Agency.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 35296–35304
November 4-9, 2025

2011). In the absence of sufficient monitoring capacity, citizen tips offer a vital channel for surfacing potential violations and extending the regulatory reach of under-resourced agencies (Yadin, 2023). Tips can reveal unknown issues, support enforcement with documentation, and enable public participation (Maniloff and Kaffine, 2021), but may also concern irrelevant matters that waste review time (Friel, 1999). A public tipline's efficacy depends on the agency's ability to adequately screen and triage incoming tips.

Regulatory agencies face substantial burdens in handling the volume of public submissions. For example, the U.S. Consumer Financial Protection Bureau has received over 1.5 million complaints since 2011, and the U.S. Securities and Exchange Commission processes more than 40,000 tips annually (Engstrom et al., 2020; Duara, 2022). Large-scale public input can strain limited staff resources, leading to delays and missed violations. When citizens feel their reports are ignored, trust and engagement suffer, potentially triggering a cycle of participatory fatigue that undermines both oversight and public accountability (Fung, 2015).

To address routing bottlenecks, agencies have begun exploring using machine learning (ML) methods to categorize and prioritize citizen tips (Madyatmadja et al., 2023; Engstrom et al., 2020; Heilweil, 2024). Yet, despite growing interest, there is limited empirical research on LLM-assisted triage using real-world regulatory data. One exception is the use of a fine-tuned BERT model to improve the routing of public service complaints in Portugal (Caldeira et al., 2022). Extending beyond that study's technical considerations, our approach was developed in collaboration with the EPA and informed by interviews with agency desk officers, allowing us to design a classification pipeline that aligns with existing workflows and reflects the agency's operational priorities and decision-making patterns.

3 Institutional Background

The "Report a Violation" system, which receives thousands of annual submissions, is the EPA's public tipline for environmental violations (U.S. Environmental Protection Agency, 2024a). Approximately 8% of currently open criminal dockets originated from the tipline, including a case that resulted in a $100,000 fine for a New Jersey biodiesel company, signifying the role of tips in surfacing

Routing	Tip Text
Criminal Division	Intentionally dumps automotive fluids including fuel, coolant, and hydraulic fluid in the creek on the east side of the building. The grass is dead and there is evidence of harmful fluids to wildlife.
Civil Division	Hydraulic fluid spill in the parking lot and into the storm water drain. Drum of slate fuel near outside smoking area.
Irrelevant to EPA	Major generator under constant use, creating noise disturbance to the community[.] Fumes are emitted 24 hours a day, preventing neighbors from sleeping.

Table 1: Example tips submitted to the EPA tipline.

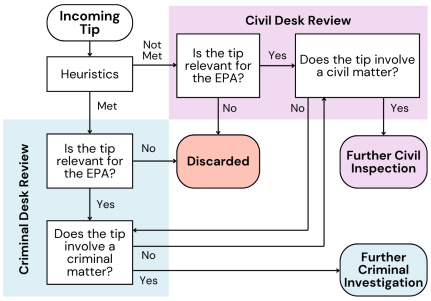

Figure 1: Diagram of EPA's internal tip routing. We replace the heuristic step with an LLM-assisted approach.

actionable leads (Donnelly, 2018). Each tip contains structured fields (*e.g.*, violation method, intent, involved parties) and a free-text field where the submitter can describe the issue (see Table 1).

Once submitted, tips are routed to either the Civil or Criminal Division for desk review, based solely on structured form fields. Tips marked by the submitter as "intentional" violations or as containing certain high-risk characteristics are routed to the Criminal Division, while all others go to the Civil Division. Desk review officers verify each tip's proper classification by examining free-text descriptions for contextual factors like environmental impacts and culpability evidence that structured fields may not capture. Irrelevant tips are discarded, misclassified tips are re-routed, and correctly routed tips advance on to enforcement, *i.e.*, Criminal Division tips are referred to an EPA Field Agent for in-depth investigation (see Figure 1). According to a time-use survey we conducted with desk officers, on an average day, officers spend over an hour removing irrelevant tips and two hours re-routing and processing relevant tips.

The existing heuristic step has two major limitations: It fails to filter out irrelevant tips that fall outside EPA jurisdiction; it also frequently misroutes civil matters to the Criminal Division, diverting time from critical investigations which require substantially more personnel (*i.e.*, desk officers often

serve dual roles as field agents), evidence building, and prosecutor coordination (U.S. Environmental Protection Agency, 2024b, 2025).

4 Methodology

Our data comes from three tipline datasets provided by the EPA: 1,371 unique tips labeled as irrelevant (2023-2024), 12,958 unique tips ultimately relevant to the Civil Division (2018-2023), and 4,780 unique tips ultimately relevant to the Criminal Division (2018-2023), totaling 19,109 tips in multiple languages.

Our two-stage pipeline addresses both heuristic limitations sequentially: First, a relevance model filters tips that do not pertain to the EPA; second, a triage model routes the remaining tips to either the Civil or Criminal Division for desk review. Using a 70/10/20 split, we train the models on the free-text portion of submissions. The relevance model trains on the entire training set, while the triage model trains only on tips desk officers labeled as relevant.

Both the relevance and triage models use Mistral-7B, chosen for its effectiveness on domain-specific tasks (Jiang et al., 2023; Arbel et al., 2024; Bolton et al., 2024), fine-tuned with LoRA adapters. Both models perform binary classification (*i.e.*, relevant or irrelevant, Criminal or Civil Division), compute confidence scores from the first generated token, and use F1-optimized thresholds that can be individually adjusted based on EPA priorities. To further evaluate our triage model, we compare the Mistral model against five alternative models: TF-IDF+SVM as a traditional statistical approach for short texts (Tong and Koller, 2001), DistilBERT (67M parameters) as an efficient transformer model, RoBERTa (125M parameters) as a strong pre-trained language model, Llama-3.1-8B as a state-of-the-art open-source foundation model, and Qwen-2.5-7B as a competitive multilingual language model (Liu et al., 2019; Sanh et al., 2020; Grattafiori et al., 2024; Yang et al., 2025). We use 5-fold cross-validation where each fold independently optimizes thresholds on validation data and evaluates on held-out test data. Within each fold, the relevance model trains on the entire training portion, while the triage model trains only on tips desk officers labeled as relevant. To test our models against a more realistic distribution of tip types, we normalize each dataset to annual rates, assuming each is comprehensive for its time period. Our re-weighted test set reflects approximately 1,645 (31.4%), 2,592 (49.4%), and 1,006 (19.2%) submitted tips per year for irrelevant, civil, and criminal matters, respectively. Reported metrics represent means ± standard deviations across the 5 folds. [2]

Selection Depth (%)	Threshold	F1 Score	Accuracy (%)	Precision (%)	FPR (%)	FNR (%)
80.0 ± 1.0	0.19	0.920 ± 0.006	88.1 ± 0.9	85.5 ± 1.0	37.1 ± 3.2	0.4 ± 0.1
76.9 ± 1.9	0.53	0.933 ± 0.009	90.3 ± 1.5	88.4 ± 1.9	28.6 ± 5.3	1.1 ± 0.3
75.0 ± 2.5	0.73	0.939 ± 0.011	91.3 ± 1.7	90.0 ± 2.5	24.0 ± 6.7	1.8 ± 0.6
72.8 ± 2.5	**0.86**	**0.943 ± 0.004**	**91.9 ± 0.6**	**91.6 ± 1.8**	**19.5 ± 4.6**	**2.8 ± 1.3**
71.0 ± 2.9	0.93	0.939 ± 0.005	91.4 ± 0.7	92.4 ± 1.9	17.4 ± 5.0	4.5 ± 2.0
60.5 ± 7.1	0.99	0.894 ± 0.039	86.6 ± 4.2	95.9 ± 2.0	8.4 ± 4.7	15.7 ± 8.2

Table 2: Performance of Mistral relevance model, by threshold. Selection depth is the proportion of all incoming tips predicted as relevant. All metrics use relevant tips as the positive class. All metrics represent cross-validated results with confidence intervals

Model	Threshold	F1 Score	Accuracy (%)	Precision (%)	FPR (%)	FNR (%)
Mistral-7B	0.335 ± 0.065	**0.758 ± 0.005**	85.7 ± 0.2	73.7 ± 1.6	11.2 ± 1.3	21.9 ± 2.4
TF-IDF+SVM	0.322 ± 0.024	0.667 ± 0.008	79.0 ± 0.5	61.0 ± 1.4	18.9 ± 1.5	26.4 ± 3.0
DistilBERT	0.321 ± 0.042	0.715 ± 0.009	82.5 ± 0.7	66.8 ± 1.5	15.3 ± 1.0	23.1 ± 0.8
RoBERTa	0.368 ± 0.083	0.742 ± 0.015	84.6 ± 1.0	71.3 ± 2.5	12.6 ± 2.2	22.3 ± 4.0
LLama-3.1-8B	0.372 ± 0.118	0.759 ± 0.014	**85.8 ± 1.2**	**74.1 ± 3.6**	**11.0 ± 2.3**	22.1 ± 2.7
Qwen2.5-7B	0.393 ± 0.092	0.746 ± 0.004	84.6 ± 0.4	70.5 ± 1.8	13.4 ± 1.6	**20.5 ± 2.6**

Table 3: Performance of triage model, by model. All metrics use Criminal Division as the positive class. All metrics represent cross-validated results with confidence intervals.

5 Results

Our Mistral relevance model using F1-optimal thresholds achieves 91.9% accuracy, and we achieve 97.2% recall for relevant tips, ensuring nearly all matters requiring the EPA's attention are captured. Table 2 situates the F1-optimal threshold among alternatives which may be deemed more appropriate by the EPA given the actual asymmetric costs of missing relevant tips in practice. Using our optimal threshold, the EPA would have to review 321 irrelevant tips per year (20.8% of all irrelevant tips) and would miss 101 relevant tips per year (2.8% of all relevant tips). Alternatively, if missing a relevant tip were considered 20 times as costly as processing an irrelevant tip, then a threshold of 0.14 would minimize overall cost, only filtering out 8 relevant tips per year (0.2% of all relevant tips) while processing 632 irrelevant tips per year (38.4% of all irrelevant tips), reducing the total tip review workload by 19.5%. [3]

[2] In total, we spent eight hours for training and two hours for inference with one NVIDIA A100 GPU.

[3] We also tested model performance across different training data fractions (25%, 50%, 75%, 100%) and found that the relevance model achieved consistently high performance regardless of training set size, while the triage model showed substantial improvement from 25% to 50% training data with diminishing returns thereafter. Full ablation results and cross-validation analysis can be found in Table 5 in the Appendix.

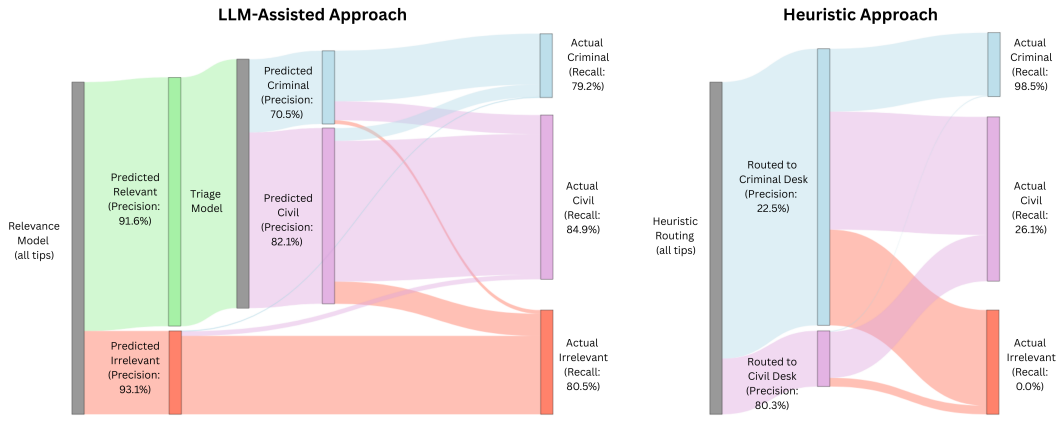

Figure 2: Flow of tips through our LLM-assisted pipeline, compared to the existing heuristic-based pipeline.

Table 3 presents the performance of our triage models on Civil versus Criminal Division classification, excluding irrelevant tips that incorrectly passed the relevance filter. The Mistral-7B approach achieves the highest F1 score of 0.76 at a threshold of 0.34. While Llama-3.1-8B minimizes civil matters incorrectly sent to Criminal Division, Mistral still maintains comparable performance in this regard. We chose Mistral over Llama because of its superior probability calibration and more reliable confidence distributions, which are critical for threshold-based decision-making in operational deployment.

Figure 2 illustrates our complete two-step classification workflow with F1-optimal thresholds. Note that these pipeline evaluation metrics capture errors from both stages: relevant tips wrongly filtered and irrelevant tips wrongly routed. [4] The existing heuristic approach routes 83.9% of all tips to the Criminal Division despite only 19.2% of all tips being genuine criminal matters. This creates a disproportionate burden on Criminal Division desk officers who must review and re-route 77.5% of their assigned tips. In contrast, our pipeline increases Criminal Division routing precision from 22.5% to 70.5%, significantly reducing the amount of time Criminal Division desk officers spend on irrelevant or civil matters. The number of civil matters that Criminal Division desk officers must re-route decreases dramatically from approximately 1,914 to just 298 per year, representing an 84.4% reduction. While Criminal Division routing recall (*i.e.*, the proportion of actual criminal matters correctly routed on the first pass) decreases from 98.5% to 79.2%, the overall Criminal Division workload is reduced by 74.2%, from 4,399 to 1,132 tips per year.

Simultaneously, the Civil Division benefits from more accurate initial routing, with correctly routed civil matters rising dramatically from 26.1% to 84.9%. Our system maintains strong Civil Division routing precision despite processing over triple the tips at the outset (2,684 compared to 842 per year), instead of after delayed re-routing. Furthermore, our relevance model successfully filters 80.5% of irrelevant tips that would otherwise consume valuable staff time. Overall, our pipeline increases the end-to-end routing accuracy from 31.8% to 82.4%, effectively addressing both limitations of the current system: irrelevant tips are filtered before reaching either division, and the remaining tips are more accurately routed, significantly reducing the effort spent on irrelevant and misrouted tips.

These precision and efficiency gains come with trade-offs that warrant consideration. Our LLM-assisted system's recall for criminal matters falls below the heuristic system's 98.5% recall, as 19.9% of criminal matters are incorrectly routed to the Civil Division, delaying response to potentially serious violations. However, by lowering the triage threshold to match the heuristic system's 98.5% recall, our pipeline would still achieve almost double its precision (41.7% compared to 22.5%). Notably, only a small proportion of tips ultimately result in prosecution, given the extensive process beyond

[4] The full metrics for the pipeline evaluation can be found in Table 6 in the Appendix

Routing	Tip Text
Criminal Division	Selling programs on how to increase torque and horsepower by intentionally altering ECMs/PCMs, oxygen sensors, and as an outcome all altered vechicles are not EPA compliant
Civil Division	RV park owner has openly refused to obey regulatory laws regarding protecting watershed by allowing dirt and debris into the small waterway that passes through the property. Heavy equipment has leaked diesel fuel onto thee ground within 20 feet of the waterway. there is still a stain on the gravel where the fuel leaked. There is no visible posted permit anywhere to be seen on or around this site

Table 4: Examples of interpreting triaging predictions with feature ablation. Features attributed as supporting criminal investigations are colored in red, and those supporting civil investigations are colored in green.

initial intake. Additionally, our relevance filter incorrectly discards 1.0% of criminal matters and 3.6% of civil matters as irrelevant before reaching any desk officer. Although missing relevant tips may have serious costs, these may be outweighed by the substantial reduction in processing irrelevant tips. Finally, while the Civil Division experiences an increased workload, they benefit from seeing fewer irrelevant tips and receiving more civil matters immediately, without the delay of re-routing.

6 Qualitative Analyses

To better understand our models' decision-making patterns and validate their practical utility, we conducted two detailed qualitative analyses. We first used feature attribution to understand how the classification results can be attributed to the input tips. We also manually reviewed classification errors, examining legitimate tips incorrectly filtered as irrelevant, irrelevant tips incorrectly retained for review, and tips misrouted between divisions.

Attribution analyses. We conduct attribution analyses with Captum by Kokhlikyan et al. Specifically, we visualize how each word in a given tip contributes to the model's relevance or triage predictions through feature ablation and integrated gradients (Sundararajan et al., 2017; Miglani et al., 2023). As shown in Table 4, attribution methods reveal how words can push the model to route a tip to the criminal division or label it as less urgent. After reviewing feature attribution for the triage model's 100 randomly selected classification results, we find the model capable of leveraging words that suggest the allegation's severity and criminality in line with the EPA guideline for tip routing.

Classification error review. We find that the relevance model very rarely misclassifies tips involving serious environmental violations. Of all 159 tips incorrectly filtered at the relevance stage, only 13 were legitimate Criminal Division matters, including whistleblower complaints and corruption cases. The majority of incorrectly filtered Civil Division tips consisted of conspiracy theories, incoherent submissions, and complaints outside EPA jurisdiction, including chemtrail theories and neighborly disputes. Although labeled as Civil Division matters in our ground truth data, the model does not appear to be missing serious Civil violations. Among the 12 seemingly actionable tips from 146 incorrectly filtered Civil cases, most concerned chemical spraying on properties and contamination spills.

In analyzing 100 randomly selected triage errors, we found that criminal tips misclassified as civil often lacked clear signals of intentionality, while civil tips mistaken for criminal typically included explicit references to deliberate dumping or repeat violations. Because all tips passing the relevance filter are reviewed by desk officers, triage errors result in re-routing rather than missed violations. The concentration of errors near decision thresholds offers operational flexibility: agencies can adjust thresholds to balance false negatives against workload, reducing re-routings under capacity constraints. Relative to the heuristic system, our approach yields far fewer re-routings while allowing greater flexibility for tuning.

7 Discussion

By implementing a two-stage relevance and triage model pipeline, we filter out most irrelevant tips, increase Criminal Division desk officer efficiency, and improve overall routing accuracy from 31.8% to 82.4%, significantly reducing the unnecessary re-routing of the current heuristic approach.

Our work contributes to the emerging field of algorithmic governance, and the broader responsible ML research community, by demonstrating a real-world implementation that provides substantial efficiency gains while preserving the regulatory workflow structure of a consequential federal agency. As agencies face increasing volumes of public input with constrained resources, this approach offers a pathway for scaling capacity without sacrificing quality. By improving tip processing efficiency, LLM-based systems can help ensure citizen reports receive appropriate attention, potentially strengthening public trust while preserving meaningful public participation.

8 Limitations

Our study has several important limitations that inform both the interpretation of our results and future work in this area.

(1) Misclassification costs. While we demonstrate a promising application of LLMs to regulatory triage, future work would benefit from a field trial component to better measure the asymmetric costs associated with different error types, given the substantial existing workloads of agency staff. In regulatory contexts, misclassification costs are rarely symmetric – failing to identify a serious environmental violation likely carries greater social cost than unnecessarily reviewing an irrelevant tip, and different error types carry different time costs. We tested multiple hyperparameter configurations and chose the option that gives agencies the most flexibility in choosing thresholds that reflect their perceived asymmetric costs. We also cannot quantify the potential acceleration in processing timelines for civil matters that our system would enable by routing them directly to the Civil Division without delays from unnecessary Criminal Division desk review. Without direct quantification of these tradeoffs, our optimization metrics may not perfectly align with the agency's implicit utility function. However, as we've underscored previously, our pipeline is designed to be inherently flexible for agencies to tailor to their own cost calculus (*i.e.*, by simply selecting different thresholds), and to their evolving priorities over time. Future work would benefit from incorporation of new data sources to establish baseline processing times for different tip categories and error types and collaboration with regulatory staff to calibrate error penalties that better reflect the agency's enforcement priorities

(2) Explainability. One limitation of our approach is the lack of interpretability inherent in fine-tuned LLMs, which creates challenges in regulatory contexts where transparency and accountability are paramount (Calo and Citron, 2020). While our system achieves strong performance metrics, it offers minimal explainability, making it difficult for both agency staff and affected parties to understand the reasoning behind specific classification decisions. We emphasize that our approach, in keeping humans in the loop with desk review, mitigates many of these concerns, relative to a system that completely automates the decision to open a field investigation, or even to assess fines. Recent advances in LLM interpretability, such as chain-of-thought prompting and structured reasoning frameworks, could further address these concerns (Wei et al., 2022; Khattab et al., 2023). We did explore chain-of-thought prompting, but found that when the system could generate reasoning text, it struggled to consistently produce the structured binary classifications required for deployment. Given that effective regulatory reasoning must be grounded in specific statutory frameworks, ensuring that generated explanations align with established legal principles would require validation from EPA officers. The datasets we were given did not include documentation of why routing decisions were made, making it difficult to evaluate whether generated explanations reflected proper regulatory reasoning. Future work would benefit from validation from EPA domain experts as to the proper alignment with regulatory reasoning.

(3) Multimodal methods. Our models are trained solely on the free-text descriptions provided in complaints, neglecting potentially valuable structured fields in the tipline system. Only a small percentage of tips identified specific facilities, limiting our ability to link tips with the EPA's compliance and enforcement history databases. Facility identifiers could enable integration with historical violation data, providing contextual information about repeat offenders, past fines, or ongoing enforcement actions that might influence routing decisions. Future research could explore multimodal models that combine both unstructured text and entity-specific compliance records.

(4) Labels as proxy for ground truth. Our evaluation relies on historical routing decisions as ground truth, which may incorporate existing biases or inconsistencies in EPA practice. We are not able to assess whether these historical decisions reflect optimal enforcement priorities or resource allocation. The mismatch between labels and the underlying outcomes they are a proxy for is an important general challenge for LLM-based interventions across many settings (Obermeyer et al., 2019). Future work could incorporate outcome-based evaluations that track not just agreement with historical decisions but also whether automated routing improves detection of actual violations, compliance with existing guidelines, and equity of enforcement actions across different communities and environmental issues.

References

Iftach Arbel, Yehonathan Refael, and Ofir Lindenbaum. 2024. TransformLLM: Adapting Large Language Models via LLM-Transformed Reading Comprehension Text. Preprint.

Eric Biber. 2009. Too Many Things to Do: How to Deal with the Dysfunctions of Multiple-Goal Agencies. *Harv. Envtl. L. Rev.*, 33:1.

Elliot Bolton, Betty Xiong, Vijaytha Muralidharan, Joel Schamroth, Vivek Muralidharan, Christopher D. Manning, and Roxana Daneshjou. 2024. Assessing the Potential of Mid-Sized Language Models for Clinical QA. Preprint.

Francisco Caldeira, Luís Nunes, and Ricardo Ribeiro. 2022. Classification of Public Administration Complaints. In *11th Symposium on Languages, Applications and Technologies (SLATE 2022)*, pages 9–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Ryan Calo and Danielle Keats Citron. 2020. The Automated Administrative State: A Crisis of Legitimacy. *Emory LJ*, 70:797.

Frank Donnelly. 2018. N.J. Fuel Company Admits to Dumping Contaminated Wastewater into Arthur Kill.

Nigel Duara. 2022. From Scandal to Scrutiny: How Intense Citizen Oversight Reshaped Oakland Police.

David Freeman Engstrom, Daniel E. Ho, Catherine M. Sharkey, and Mariano-Florentino Cuéllar. 2020. Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies. *NYU School of Law, Public Law Research Paper*, (20-54).

Brian Friel. 1999. Rule Makes Dismissing EEO Complaints Easier.

Archon Fung. 2015. Putting the Public Back into Governance: The Challenges of Citizen Participation and its Future. *Public Administration Review*, 75(4):513–522.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Wayne B. Gray and Jay P. Shimshack. 2011. The Effectiveness of Environmental Monitoring and Enforcement: A Review of the Empirical Evidence. *Review of Environmental Economics and Policy*, 5(1):3–24.

Rebecca Heilweil. 2024. The FBI Is Using AI to Mine Threat Tips, but Isn't Sharing Much Detail.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. DSPy: Compiling Declarative Language Model Calls Into Self-Improving Pipelines. *Preprint*, arXiv:2310.03714.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for pytorch. *Preprint*, arXiv:2009.07896.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Preprint*, arXiv:1907.11692.

Evaristus D. Madyatmadja, Corinthias PM. Sianipar, Cristofer Wijaya, and David JM. Sembiring. 2023. Classifying Crowdsourced Citizen Complaints through Data Mining: Accuracy Testing of K-Nearest Neighbors, Random Forest, Support Vector Machine, and AdaBoost. *Informatics*, 10(4):84.

Peter Maniloff and Daniel T. Kaffine. 2021. Private Monitoring and Public Enforcement: Evidence from Complaints and Regulation of Oil and Gas Wells. *Journal of Environmental Economics and Management*, 108:102473.

Vivek Miglani, Aobo Yang, Aram H Markosyan, Diego Garcia-Olano, and Narine Kokhlikyan. 2023. Using captum to explain generative language models. *arXiv preprint arXiv:2312.05491*.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science*, 366(6464):447–453.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *Preprint*, arXiv:1910.01108.

Jay P. Shimshack. 2014. The Economics of Environmental Monitoring and Enforcement. *Annual Review of Resource Economics*, 6(1):339–360.

Jodi L. Short. 2021. The Politics of Regulatory Enforcement and Compliance: Theorizing and Operationalizing Political Influences. *Regulation & Governance*, 15(3):653–685.

Matthew C. Stephenson. 2006. The Strategic Substitution Effect: Textual Plausibility, Procedural Formality, and Judicial Review of Agency Statutory Interpretations. *Harv. L. Rev.*, 120:528.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. *Preprint*, arXiv:1703.01365.

Simon Tong and Daphne Koller. 2001. Support Vector Machine Active Learning with Applications to Text Classification. *Journal of Machine Learning Research*, 2(Nov):45–66.

U.S. Environmental Protection Agency. 2024a. Report an Environmental Violation: General Information.

U.S. Environmental Protection Agency. 2024b. Strategic Civil-Criminal Enforcement Policy. Memorandum from the Assistant Administrator for Enforcement and Compliance Assurance, U.S. EPA.

U.S. Environmental Protection Agency. 2025. Basic Information on Enforcement.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Sharon Yadin. 2023. The Crowdsourcing of Regulatory Monitoring and Enforcement. *The Law & Ethics of Human Rights*, 17(1):95–125.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 Technical Report. *Preprint*, arXiv:2412.15115.

A Appendix

Training Data	Relevance F1 Score	Relevance AUROC	Triage F1 Score	Triage AUROC
25%	0.981 ± 0.002	0.945 ± 0.011	0.538 ± 0.077	0.723 ± 0.113
50%	0.983 ± 0.001	0.961 ± 0.009	0.730 ± 0.020	0.913 ± 0.009
75%	0.981 ± 0.003	0.938 ± 0.025	0.740 ± 0.004	0.923 ± 0.003
100%	0.983 ± 0.002	0.961 ± 0.013	0.721 ± 0.072	0.906 ± 0.048

Table 5: Cross-validation performance across different training data fractions for both the relevance model and the triage model. Results show mean ± standard deviation across 5 folds. We find that the relevance model achieves consistently high performance regardless of training set size, while the triage model benefits substantially from additional training data, with performance plateauing around 50% of the full dataset.

Stage	Metric	Heuristic System	LLM-Assisted System
Relevance	Relevant Precision (%)	N/A	91.6 ± 1.8
	Relevant Recall (%)	N/A	97.2 ± 1.3
	Irrelevant Precision (%)	N/A	93.1 ± 2.4
	Irrelevant Recall (%)	N/A	80.5 ± 4.6
Criminal Division	Criminal Precision (%)	**22.5 ± 0.1**	**70.5 ± 2.5**
	Criminal Recall (%)	**98.5 ± 0.5**	**79.2 ± 2.9**
	Tips Re-routed to the Civil Division (tips/year)	1914 ± 15	298 ± 43
	Irrelevant Tips Processed (tips/year)	1494 ± 27	38 ± 13
	Legitimate Tips Lost to Filter (tips/year)	0	10 ± 6
	Annual Caseload (tips/year)	4399 ± 18	1132 ± 83
Civil Division	Civil Precision (%)	**80.3 ± 3.3**	**82.1 ± 2.5**
	Civil Recall (%)	**26.1 ± 0.6**	**84.9 ± 2.2**
	Tips Re-routed to the Criminal Division (tips/year)	16 ± 5	200 ± 28
	Irrelevant Tips Processed (tips/year)	151 ± 27	283 ± 75
	Legitimate Tips Lost to Filter (tips/year)	0	92 ± 42
	Annual Caseload (tips/year)	843 ± 19	2684 ± 145
	End-to-End Accuracy (%)	31.8 ± 0.4	82.4 ± 0.8

Table 6: Performance Comparison: Heuristic vs LLM-Assisted System. Values shown as mean ± standard deviation across 5-fold cross-validation.