CaseLinker: Interpretable ML Approaches for Analyzing Internet Crimes Against Children Reports

Mrinaal Ramachandran
Graduate Student, Department of Computer Science
University of Massachusetts Amherst
mramachandra@umass.edu
Independent Research

March 2026

Technical Report Series #2 · github.com/mrinaalr/CaseLinker · Live Demo ·

Abstract

This is the second in a monthly series of technical reports tracking CaseLinker's development as an open-source platform for cross-case analysis of Internet Crimes Against Children (ICAC) reports. Each report documents progress toward scaling from roughly 100 to thousands of publicly available case reports while integrating interpretable ML functionality, visualizations, and filtering tools that allow investigators to engage deeply with case patterns—without requiring repeated direct exposure to potentially disturbing material [2, 3]. This report covers: (1) integration of Named Entity Recognition (NER) via Stanza (Stanford NLP) into a hybrid deterministic-ML processing pipeline; (2) dataset expansion from 47 to 207 cases across seven source PDFs spanning 2011–2024; and (3) early patterns emerging from the expanded dataset, including the distributed network of over 215 law enforcement organizations spanning multiple states and years, along with recurring trends in platform usage and investigation structure across jurisdictions. Evaluation demonstrates 87.9% agency extraction coverage across the expanded dataset. As CaseLinker scales toward thousands of cases, integrating AI thoughtfully—in ways that remain auditable and that keep investigators in control—becomes both the central technical challenge and the most important design commitment. Automated insights and clustering will continue to improve, but they are not a substitute for the judgment of an experienced investigator or journalist working directly with this data. Expert intuition—the ability to recognize meaningful patterns from context and experience— remains something algorithms do not yet replicate [5]. CaseLinker is built to support that judgment.

Keywords: child exploitation, interpretable machine learning, named entity recognition, cross-case analysis, analyst well-being, ICAC investigations

Contents

1 Introduction 2
1.1 Monthly Report Series . . . 2
1.2 What Changes at Scale . . . 2
2 Design Philosophy and the Role of Interpretable ML 3
2.1 The Problem with Scaling Without ML . . . 3
2.2 What NER Unlocks . . . 3
2.3 Patterns Immediately Discovered in the Expanded Dataset . . . 3
2.4 Interpretability as a Non-Negotiable . . . 5
2.5 Future ML Integrations . . . 5
3 Audit and Performance 6
3.1 Verification Before Storage . . . 6
3.2 Interactive Audit Interface . . . 6
3.3 Performance . . . 6
4 Moving Forward 6
5 Roadmap 7
6 Conclusion 7

1 Introduction

1.1 Monthly Report Series

The initial CaseLinker technical report [1] documented a working prototype evaluated on 47 publicly available AZICAC case reports from 2011–2014. That report established the five-layer architecture, the deterministic extraction pipeline, and the evaluation baseline. This series picks up from there, releasing a technical update each month that documents what was built, what was learned, and what the data is starting to show.

The goal of the series is not to report a finished system. It is to build one transparently—showing the reasoning behind each technical decision, the tradeoffs accepted, and document findings that become visible as the system grows. Each report is also a checkpoint: a record of what CaseLinker can and cannot do at a given scale, as the project evolves in studying and extracting meaningful patterns from this data.

1.2 What Changes at Scale

ICAC case reports contain attributes worth studying systematically: the nature of the offense (sexual assault, possession, production, extortion), the victim's age and relationship to the perpetrator, which platforms or methods were involved, what investigative approach was used, whether the suspect was a repeat offender, a teacher, a family member, or a stranger, which agencies collaborated, how the case was prosecuted, and custom topics that can be defined by the researcher during analysis. The initial report demonstrated feasibility—structured extraction and cross-case clustering were demonstrably possible, and even that limited dataset surfaced patterns worth examining.

At 207 cases spanning 2011–2024, more comes into focus. Law enforcement agency involvement becomes visible at scale: **215 unique organizations** appear across 207 cases, with **171 (79.5%)** appearing in only one case—demonstrating a wide, distributed network of local agencies converging around a stable institutional core of federal partners and NCMEC. ICAC task forces appear in **51.2%** of cases, the FBI in **9.2%**, with its specialized Child Sexual Exploitation Unit accounting for roughly half of those appearances. That federal-local structure is not obvious from any single case; it only emerges in aggregate.

Other patterns are equally instructive. **Stranger perpetrators account for 91.8%** of cases—a figure that cuts against common assumptions about who commits these crimes. Possession dominates over production (**66.2%** versus **2.9%**), yet **37.7%** of cases involve sexual assault, and **6.8%** involve victims under 12. The average age gap between reported perpetrator and victim age is **22.6 years**. Platform data is present in only 32.4% of cases, but where it appears, online methods (21.3%), chat platforms (8.2%), and Facebook (4.8%) lead—with Snapchat and Discord beginning to appear in more recent cases.

These are early signals and informal statistics easily derived from this database, not conclusions. The research questions they open will scale alongside the dataset—and more will emerge as it grows, including: how do federal-local coordination structure vary by offense type or geography? Do proactive versus reactive investigation split predict prosecution outcomes? How have platform patterns shifted between 2011–2014 and 2022–2024, and do newer platforms correlate with different severity profiles? Are there common investigative methods that correlate with better outcomes for the youngest and most severely harmed victims? The infrastructure to find answers to these and more questions in this landscape is being built.

This is publicly available data containing no personally identifiable information, analyzed under a Not Human Subjects Research determination (HRPO #7668) from the University of Massachusetts Amherst Human Research Protection Office, confirming the research contains no private or identifiable information under federal regulations [45 CFR 46.102(f)(1), (2)].

2 Design Philosophy and the Role of Interpretable ML

2.1 The Problem with Scaling Without ML

The initial system used exclusively deterministic, regex-based extraction. This was the right choice for a 47-case prototype: interpretable, auditable, no training data required. But regex does not scale gracefully to an environment with hundreds of agencies, inconsistent reporting formats, and 13 years of evolving language. The cost of maintaining exhaustive pattern lists grows with every new source PDF added. More importantly, patterns you did not anticipate do not get extracted at all—they simply disappear from the data.

The 207-case dataset makes this concrete. Before NER integration, agency extraction was limited to patterns explicitly encoded in the regex library. After integration, the system identified 215 unique law enforcement organizations—dramatic increase over the limited regex-only baseline—by recognizing organizational entities from linguistic context rather than direct matching alone.

2.2 What NER Unlocks

Named Entity Recognition using Stanza (Stanford NLP) adds a semantic layer to extraction. Where regex asks "does this text match a known pattern," NER asks "what kind of thing is this text referring to." For a dataset covering 61 ICAC task forces, federal agencies, county prosecutors, state police, and international partners, the difference is significant.

The extraction pipeline processes four entity types: `ORG` (filtered to law enforcement agencies), `DATE` (separating ages and event dates using contextual heuristics), `LOC`/`GPE` (geographic entities), and `PER` (excluded to avoid extracting reporter names). The result is a richer, more complete data that we can capture and analyze.

2.3 Patterns Immediately Discovered in the Expanded Dataset

With 207 cases now processed, several patterns have emerged that were not visible at smaller scale. These are observations from the current dataset, not conclusions—the sample remains limited and some figures will shift as coverage grows.

The collaborative network. The 215 unique organizations span a wide distribution: **171 (79.5%)** appear in only one case, while 44 appear in multiple cases. The most frequently recurring are ICAC Task Forces (106 cases, 51.2%), NCMEC (48 cases, 23.3%), and the FBI (19 cases, 9.2%), including its specialized Child Sexual Exploitation Unit (10 cases). A structural detail worth noting: despite comprising only ∼10 unique organizations, federal agencies account for ∼80 total case appearances—comparable to the ∼375 case appearances from ∼195 unique local agencies. Federal partners appear less often in count but recur heavily across cases, with an average of ∼8.5 cases per federal agency compared to ∼2 cases per local agency.

Multi-agency coordination. The 44 organizations appearing in multiple cases form a recurring institutional core, but the majority of the network is non-recurring. That 79.5% long tail of single-appearance agencies suggests each investigation draws on a highly localized set of partners—a pattern that may demonstrate how ICAC task forces operate in practice, but one that has not previously been quantifiable from public records alone.

Geographic and temporal distribution. Cases span multiple states and 13 years. Platform data is sparse—present in only 32.4% of cases—but where it appears, the mix shifts noticeably between 2011–2014 and 2022–2024, with Snapchat and Discord appearing in more recent cases alongside the persistent dominance of generic online and chat references. Quantifying this shift systematically is a goal of Report #3.

Investigation structure. The investigation type is unknown in **46.9%** of cases—not a data error, but a reflection of inconsistent documentation in public summaries. Of the cases

where type is recorded, proactive investigations (6.3%) outnumber reactive ones (1.0%), though the sample is too small to draw reliable conclusions about outcome differences. This remains an open analytical question as the dataset grows.

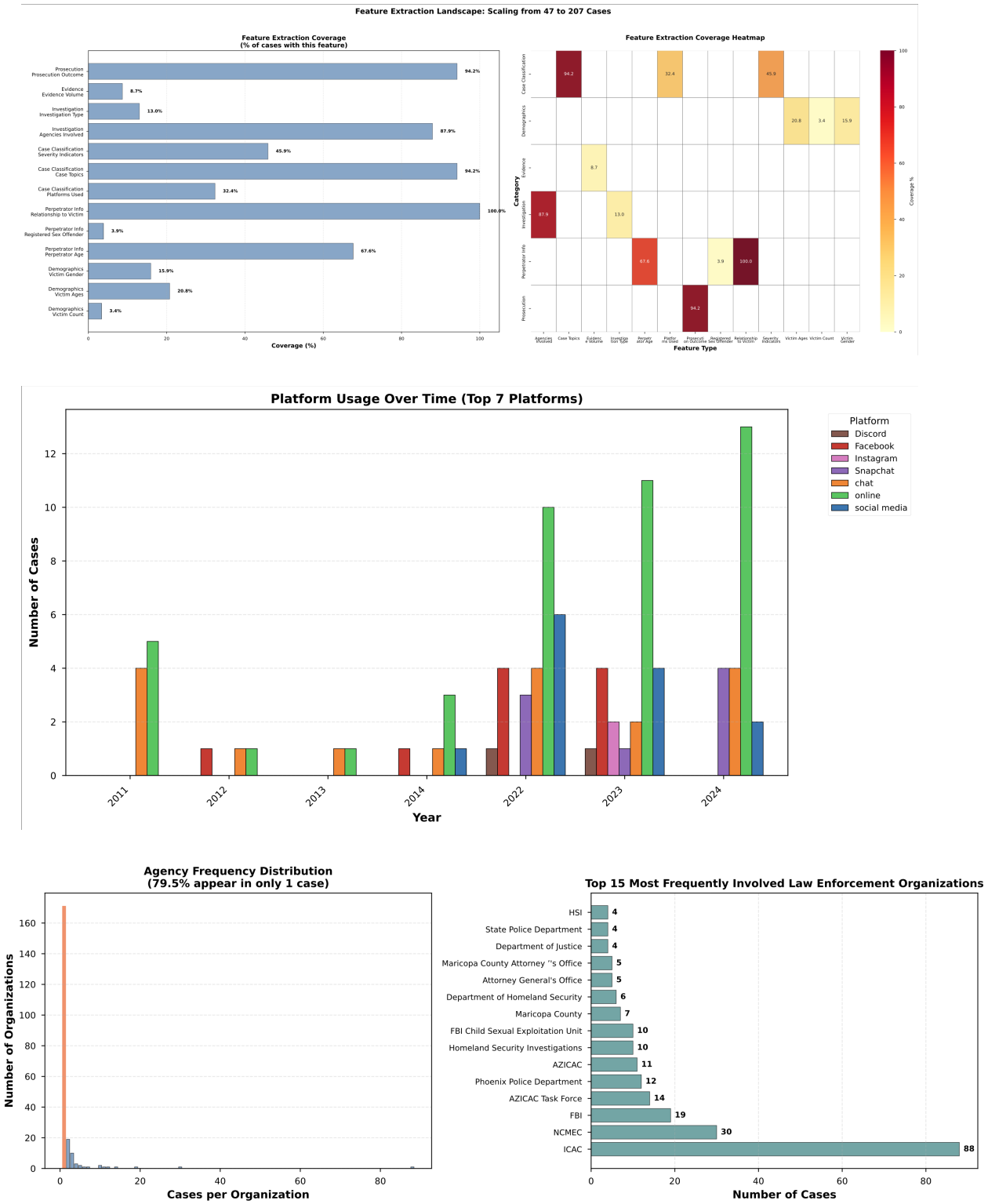

Figure 1: Initial exploratory analysis from the CaseLinker dataset. These figures illustrate feature extraction coverage, law enforcement organization participation patterns, and platform usage trends across ICAC case reports spanning 2011–2024.

2.4 Interpretability as a Non-Negotiable

Adding ML to CaseLinker does not mean accepting opaque black-box outputs. Every extraction is subject to the same auditability requirement as the deterministic pipeline: an analyst must be able to trace any feature back to its source text and understand how it was identified. For NER specifically, this means the merge logic is explicit and inspectable—pattern-based extraction takes precedence in conflicts, and NER contributions are flagged as such in the database. The goal is a system where an analyst can always ask "why does this case show this agency" and get a straightforward answer.

A simpler alternative—uploading case narratives directly to a commercial large language model—is not viable for this domain. Beyond the sensitivity of the content itself, commercial LLM APIs may temporarily retain submitted inputs and outputs for abuse monitoring and safety purposes, under data-usage policies that vary by provider [4]. Outputs are probabilistic and cannot be traced to source text [8], violating the auditability requirement for legal and research contexts. There is no established chain of custody for LLM-derived extractions, and models may reproduce or surface sensitive details in unpredictable ways. CaseLinker's architecture exists precisely because the standard shortcuts are not appropriate here.

This matters not just for analyst trust but for legal and reporting contexts. Case analysis that informs investigative decisions needs to be explainable. A feature or statistic extracted by an opaque model that cannot be traced to source text is not useful in that context—it is a liability.

2.5 Future ML Integrations

The ML roadmap for CaseLinker follows the same principle as the NER integration: each addition must be optional, independently verified, and clearly separated from the deterministic core. Three capabilities are under active development.

Semantic similarity. Current case clustering uses weighted Jaccard similarity to determine sub-groups across 5 cluster topics. Sentence transformer embeddings (`all-mpnet-base-v2`) would enable similarity based on meaning rather than keyword overlap—finding cases that describe the same offender methodology even when the surface language differs. For investigative narratives with legal terminology, a domain-adapted model such as nlpaueb/legal-bert-base-uncased [7] will be evaluated against the general-purpose baseline before any production integration.

Enhanced grouping. Semantic embeddings also make it possible to group cases and create diverse new clusters that traditional keyword methods miss—for example, similar platform usage patterns described differently across task forces or consistent investigative approaches that do not share exact terminology. Entity networks derived from NER output could additionally visualize relationships between organizations, platforms, and case characteristics across the full dataset.

Content sanitization. Summarization models (BART/T5) may eventually generate analytical case summaries that preserve investigative facts while omitting explicit detail—reducing direct exposure to disturbing material during review. This capability requires careful evaluation: a summary that loses legally relevant detail is worse than no summary. It is listed here as a research direction, not a planned release.

All ML dependencies are maintained in a separate `requirements-ml.txt` and are not required for core system operation. The system reverts gracefully if ML components are unavailable, ensuring the deterministic pipeline remains the reliable foundation.

3 Audit and Performance

3.1 Verification Before Storage

All ML-extracted features are merged with deterministic outputs and validated before being written to the database. The merge logic enforces explicit rules: pattern-based extraction takes priority for fields where regex achieves high precision, and NER fills gaps rather than overwriting confident extractions. Normalization handles surface-form variations before any feature reaches storage—apostrophe inconsistencies, spacing artifacts, and known abbreviation errors (e.g., `ZICAC` → `AZICAC`) are resolved at the merge layer, not post-hoc.

The result is a database where every stored feature has passed through an explicit, inspectable processing chain. There are no silent failures and no opaque model outputs stored without a corresponding source.

3.2 Interactive Audit Interface

CaseLinker's audit interface allows case-by-case review of all extracted features alongside the original case text. Hovering over any extracted feature highlights the corresponding source passage, making the extraction traceable without requiring access to the underlying code. This is available for every case in the database and covers the full feature schema: perpetrator and victim demographics, platforms, agencies, investigation type, severity indicators, charges, and evidence volume.

The audit interface serves two purposes. For research use, it enables verification that the extraction pipeline is behaving correctly as new source PDFs are added. For any future practitioner use, it provides the transparency required to trust automated outputs—an analyst can spot-check any case, confirm any feature, and identify where extraction falls short.

3.3 Performance

NER processing runs during ingestion, not at query time. All clustering, triage, and insight generation are precomputed and stored, so production query performance is unaffected by the ML layer. Current benchmarks on the 207-case dataset:

Table 1: Processing Performance (207 cases, 7 PDFs)

Operation	**Time**
Full ingestion with NER (7 PDFs)	30–60 seconds
Full analysis pipeline	2–5 seconds
API endpoint queries	<100ms

Projected at 2,000 cases, ingestion is estimated at 5–10 minutes (one-time cost per batch), with query performance maintained via precomputed clusters.

4 Moving Forward

What this dataset is starting to show is that child exploitation investigations are not isolated events handled by a single agency. They are distributed, collaborative, and patterned—across states, years, platforms, and organizations. That structure is only visible in aggregate, and I am very interested in studying it.

CaseLinker is the primary vessel for this motivation. Not to replace investigative judgment and time spent reading cases, but to provide analysis tools and a platform to make answering key, challenging questions as easy as possible.

The (fun) challenge is that scaling this responsibly takes time. Each new source PDF requires validation. Each new ML component requires evaluation against ground truth. The audit interface exists precisely because these checks need to be visible and ongoing, not assumed. The monthly report series is part of that commitment—a public record of what the system can do at each stage, and what it cannot yet.

5 Roadmap

- **3 months (Report #4, June 2026).** Target: 1,000+ cases. The NCMEC CyberTipline annual reports alone contain **roughly 2,800 publicly available cases** across 2022–2024, of which 160 are currently ingested. Analyzing this existing source and additional reports from at least one other task force beyond AZICAC will bring the dataset past the 1,000-case threshold without requiring any new data acquisition. Primary engineering focus: extending the ingestion layer to handle formatting variation across task forces. ML focus: temporal trend analysis across the full 2011–2024 span, quantifying platform evolution and investigation method shifts. Visualization: platform will remain responsive and visually accessible, with multiple avenues for analysis as the dataset grows.

- **6 months (Report #7, September 2026).** Target: 2,000+ cases. Introduce sentence transformer embeddings for semantic case similarity—opt-in, clearly separated from the deterministic core, evaluated against the weighted Jaccard baseline. Begin severity classification experiments using the labeled severity indicators already in the dataset as ground truth. If institutional partnerships and user studies are desired, integrate feedback from practitioner review.

- **12 months (Report #13, March 2027).** Target: Comprehensive coverage of publicly available ICAC reports across all a wide array of online sexual exploitation instances. Full temporal analysis across all available years. Network analysis for multi-agency and multi-jurisdiction case connections. Peer-reviewed publication submission documenting the system and findings. Active user study with researchers or analysts if institutional partnerships allow.

6 Conclusion

Version 2 of CaseLinker advances the system on two fronts. The hybrid NER-deterministic pipeline achieves 87.9% agency extraction coverage on a 207-case dataset spanning 13 years and two source types, directly addressing the coverage limitation of the regex-only baseline. More importantly, the expanded dataset has started to reveal insights: a distributed collaborative network of 215 law enforcement organizations, 80% of which appear in only one case—a pattern that is invisible without cross-case analysis and that has real implications for understanding how these investigations are conducted.

The deeper goal of this project is not technical metrics. It is giving investigators and researchers tools that let them understand the landscape of child exploitation at scale—platform trends, offender patterns, investigation strategies, jurisdictional networks—without requiring them to carry the psychological weight of repeated direct exposure to case material. Every technical decision in CaseLinker is evaluated against that standard: does it help people go further, while protecting them from going through more?

Report #3 will document progress toward 1,000 cases and introduce temporal trend analysis across the 2011–2024 dataset.

CaseLinker is available at github.com/mrinaalr/CaseLinker released under the MIT License.
Live demo: web-production-13a2.up.railway.app.

References

[1] Ramachandran, M. (2026). *CaseLinker: An Open-Source System for Cross-Case Analysis of Internet Crimes Against Children Reports. Technical Report.* University of Massachusetts Amherst. doi:10.5281/zenodo.18743600

[2] Perez, L. M., Jones, J., Englert, D. R., & Sachau, D. (2010). Secondary traumatic stress and burnout among law enforcement investigators exposed to disturbing media images. *Journal of Police and Criminal Psychology*, 25(2), 113–124.

[3] Burns, C. M., Morley, J., Bradshaw, R., & Domene, J. (2008). The emotional impact on and coping strategies employed by police teams investigating internet child exploitation. *Traumatology*, 14(2), 20–31.

[4] OpenAI. (2023). *API Data Usage Policies.* https://openai.com/enterprise-privacy/

[5] Klein, G. (1998). *Sources of Power: How People Make Decisions.* MIT Press.

[6] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. *Proceedings of ACL 2020, System Demonstrations.*

[7] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. *Findings of EMNLP 2020.*

[8] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT).*