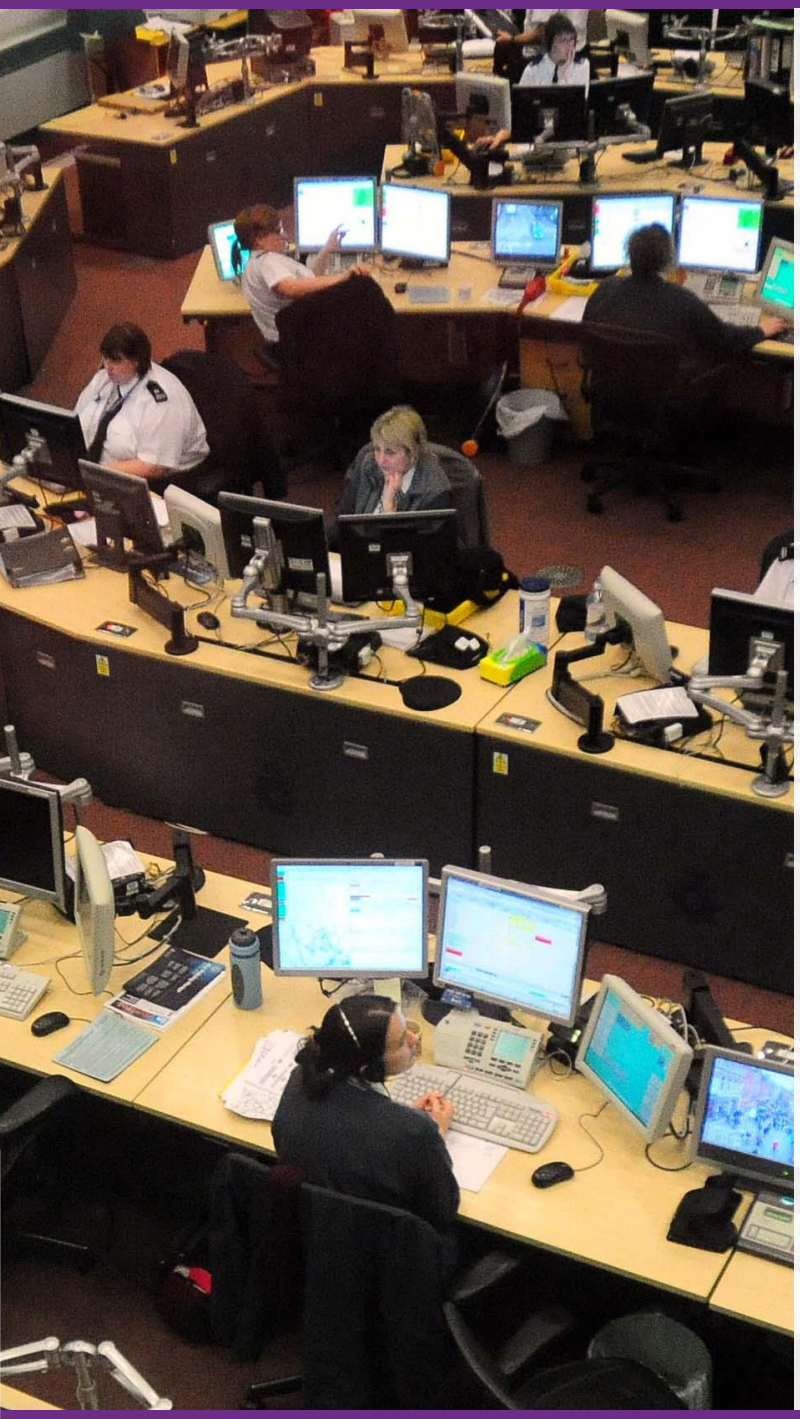
Vulnerability & Policing Futures Research Centre

Developing AI tools to help quantify police involvement with vulnerable people



Key points

- The team developed and evaluated an approach that used large language models (LLMs) to analyse free-text descriptions of interactions between the police and the public in police incident data. The team instructed the LLMs to identify four vulnerabilities: mental ill health, substance abuse, homelessness, and alcohol dependence.
- Both LLMs and human coders were given the same qualitative codebook, allowing the team to compare classifications directly.
- The LLMs' outputs closely aligned with human judgements and were particularly effective at ruling out cases without vulnerabilities.
- The team sees considerable potential in using LLMs as screening filters with humans evaluating ambiguous cases. This could significantly reduce resource demands and enable large-scale analysis previously only possible through exhaustive human coding.
- The team makes recommendations for responsible deployment.

Summary

Police often engage with individuals experiencing a range of vulnerabilities. Estimating the prevalence of these interactions is challenging – current methods rely on either counting "flags" in police information systems that may be inconsistently applied, or resource-intensive manual analyses of typically small samples of incidents.

Our research instructs large language models (LLMs) to identify vulnerability indicators in unstructured police incident logs and evaluates how they perform compared to human coders. Using publicly available incident logs from Boston Police Department, we tested our method by prompting both humans and instruction-tuned LLMs (IT-LLMs) to identify several vulnerabilities (mental ill health, substance misuse, alcohol dependence, and homelessness). Results show that IT-LLMs effectively screen out narratives where vulnerabilities are absent and demonstrate minimal demographic bias, offering a promising approach for analysing large-scale unstructured police data that could be applied to a much broader range of policing actions and contexts.

Background

Police frequently encounter vulnerable individuals throughout the course of their everyday duties. Despite growing recognition of this aspect of policing, measuring the extent of these interactions remains challenging due to limitations in data collection methods.

Current estimates typically rely on either categorical "flags" in call and dispatch systems, or requires manually analysing or observing small samples of incidents. The limitations of these and other approaches are reflected in significant variations in published estimates of police involvement with vulnerability. To illustrate, the UK's 2022 Policing Productivity Review suggested that between 5% - 9% of incidents involved mental ill health, while evidence submitted to a UK Parliamentary Inquiry estimated 20% of police time was spent on mental health-related calls. Similarly, a systematic review of North American studies found estimates ranging from 1% - 9% depending on the measurement method used. Understanding these patterns is key to inform evidence-based problem and demand analyses, training, and inter-agency coordination.

A potentially valuable but underutilised source of information exists in the narrative text that police officers or call handlers write when documenting incidents. These written accounts typically contain rich details about circumstances and behaviours that often aren't captured in standardised data fields. They could provide deeper insights into police-public populations. However, analysing these narratives has traditionally required labour-intensive manual review that becomes impractical when dealing with thousands of reports.

Recent advances in artificial intelligence, specifically IT-LLMs, offer new possibilities for automating this analysis. These models can understand and follow complex instructions to analyse text, potentially enabling systematic, large-scale qualitative analysis of free-text data without specialised training. Our research explores a generalizable methodology using IT-LLMs for qualitative analysis of police incident narratives, using four specific vulnerabilities (mental ill health, substance abuse, homelessness, and alcohol dependence) as example categories to evaluate the approach's effectiveness.

What we did

We investigated if LLMs could reproduce insights generated by human coders through identifying potential indicators of vulnerability in police incident narratives.

Using narrative police reports from Boston Police Department, we developed a codebook defining four vulnerabilities: mental ill health, substance misuse, alcohol dependence, and homelessness. We used a three-tiered labelling scheme (positive, inconclusive, negative) to account for ambiguity in the narratives.

We tested three language models of varying sizes: two open-source models (Llama 8B and 70B) and one proprietary model (GPT-4o). For each model, we explored different sets of instructions, either using codebook definitions designed for humans or custom-engineered prompts optimised through iterative refinement. We instructed the models to classify each narrative multiple times to estimate certainty.

For evaluation, we selected 500 narratives that were independently coded by two human coders using our codebook. We then compared the LLM-generated labels with human labels to assess performance.

Also, we tested for potential demographic biases by conducting counterfactual analyses, i.e. systematically altering the race and sex descriptors in narratives while keeping all other content identical. We conducted statistical tests using these counterfactual narratives, to quantify the effect of varying these demographic features.

Key findings

Our results demonstrate that IT-LLMs can effectively support qualitative coding of police narratives.

The most promising capability was the models' effectiveness to act as "negative filters", capable of reliably identifying cases without indicators of vulnerability (as defined by humans). In this role, all model configurations performed well, with precision for negative classifications consistently exceeding 90%. When the models were in strong agreement (i.e., all classifications gave the same answer), their accuracy was even higher. Both custom prompted open source models (Llama 8B and 70B) showed between 95% and 100% alignment with human classifications, while classifying 51% and 72% of examples as negative, GPT-4o reached 99% and 100% alignment on 49% and 63% of cases.

Models were less reliable in positively identifying vulnerability. When they unanimously classified a case as involving substance misuse or homelessness, GPT-4o demonstrated over 90% precision, but these cases made up a small percentage of the total (only 6–11%). The least reliable area was inconclusive classifications, where the models struggled to maintain precision above 70%. This suggests that while LLMs are effective at ruling out vulnerability, they are less consistent at confidently identifying when it is present or uncertain. Importantly however, in the data analysed we found that most cases didn't contain indicators of vulnerability.

These results indicate that a human-LLM collaboration where model and human labels are combined may provide a promising approach, significantly reducing the workload associated with manual review while ensuring expert oversight in areas where the models are weakest.

Larger models performed better with standard codebook instructions, but custom prompts significantly improved smaller model performance. Surprisingly, the smallest open source model with custom prompts achieved comparable or superior performance to OpenAI's GPT-4o using codebook instructions for all vulnerabilities studied. This finding has important practical implications, as smaller models can be deployed locally when data security concerns preclude sharing sensitive information with third-party providers.

Our counterfactual analyses found minimal bias in model outputs when race and sex were manipulated. After correcting for multiple comparisons, we found few statistically significant demographic effects, and where present, their magnitudes were generally small (less than 5% change in classification probability).

Next steps

Our research demonstrates promising applications of IT-LLMs for analysing police incident narratives but practical implementation requires careful consideration of several factors.

The approach offers substantial benefits over manual coding. IT-LLMs can process thousands of narratives rapidly, enabling analysis at scales previously impractical due to resource constraints. The process inherently requires researchers to fully articulate classification criteria through explicit codebooks and formal prompts, making analytical decisions more transparent and replicable. While human coders may unconsciously supplement written definitions with implicit knowledge, IT-LLMs work consistently from instructions provided. Having demonstrated the viability of this approach, our next step is to apply it to a large dataset that can be used to estimate how many vulnerability-related incidents take place in police interactions — a task that would be prohibitively resource-intensive using traditional methods.

Careful thought needs to be given before deploying IT-LLMs, particularly regarding data security. While larger proprietary models like GPT-4o performed well, sharing sensitive police data with third-party servers raises significant concerns. Our finding that smaller open-source models with custom prompts achieved comparable performance offers a promising alternative, allowing deployment on local secure infrastructure. Future research should systematically document the prompt engineering process to identify which elements most contribute to performance improvements and establish generalizable methods for adapting prompts across different domains.

Implementation should be limited to research applications by practitioners capable of properly evaluating results. While our counterfactual analyses showed minimal demographic bias, these models are developing rapidly and professionals with appropriate expertise will need to continually monitor this to ensure models are deployed responsibly. Importantly, these approaches should be confined to exploratory aggregate-level analyses rather than operational decision-making. The non-deterministic nature of LLMs and their occasional inconsistencies make them unsuitable for case-specific judgments where outcomes depend on precise assessments.

IT-LLMs have the potential to augment traditional qualitative methods. Their ability to rapidly process large volumes of text while maintaining consistent criteria makes them valuable for initial screening and filtering tasks. However, the need for human oversight of ambiguous cases, combined with challenges around transparency and stability, indicates they should complement rather than supersede expert judgment. Used thoughtfully within these constraints, IT-LLMs offer promising capabilities for expanding the scope and scale of research involving qualitative judgements, enabling unique insights to be derived from vast quantities of unstructured data that has to date remained largely inaccessible.

Authors

The research team were:

Sam Relins, University of Leeds; Dan Birks, University of Leeds; and Charlie Lloyd, University of York.

For further information

Scan the QR code to read more about the project.

Lead investigator:

Professor Dan Birks
d.birks@leeds.ac.uk

Researcher:

Sam Relins
s.relins@leeds.ac.uk

UKRI Economic and Social Research Council

University of York

University of Leeds