
Monograph:
Birks, D. and Relins, S. (2026) Using AI to estimate how often UK police encounter vulnerable people. Report. University of Leeds , Leeds, UK.

https://doi.org/10.48785/100/503

Vulnerability & Policing Futures Research Centre

Using AI to estimate how often UK police encounter vulnerable people

Key points

- The project team applied a large language model (LLM) to just under 3,000 anonymised incident logs from a UK police force to identify indicators of four vulnerabilities: mental ill health, substance misuse, alcohol dependence, and homelessness. The model was hosted locally within a secure research environment, reflecting the conditions under which sensitive police data must be analysed.
- Following human review and statistical adjustment, we found indicators of mental ill health were present in more than one in five incidents (23%), with alcohol dependence at 8%, substance misuse at 5%, and homelessness at 3%. These are, to the team's knowledge, the first estimates of this kind derived from routine UK police incident narratives.
- Producing these estimates required repeated classification runs, structured human review, and statistical adjustment. The model's uncorrected output overstated the prevalence of every vulnerability examined, indicating that LLM outputs should not be treated as measurements without substantial methodological support.

Summary

Police regularly encounter people experiencing vulnerability, but the extent of these interactions is not well understood. Structured administrative data provide limited insight, and the narrative accounts of incidents are too resource-intensive to analyse at scale.

Building on earlier research testing whether LLMs could reproduce human coding of publicly available US police reports, we applied a fine-tuned large language model (LLM) to just under 3,000 anonymised incident logs from a UK police force to identify indicators of mental ill health, substance misuse, alcohol dependence, and homelessness.

Following human review and statistical adjustment, we estimate that approximately 23% of incidents contain indicators of mental ill health, with lower prevalence for the remaining three vulnerabilities. The raw estimates from the model overstated prevalence in every category, indicating that LLMs can support analysis at this scale but require substantial methodological safeguards.

Background

Police frequently encounter individuals experiencing mental ill health, addiction, and homelessness. Despite widespread recognition that this forms a significant part of frontline work, there is little reliable evidence about how often it occurs. Forces hold the information needed to answer this, but in free text entries that cannot readily be analysed at scale.

Published estimates of police involvement with vulnerability vary considerably. Officers themselves report that between 40% and 75% of their work involves vulnerable individuals, while analyses of structured police data tend to produce lower, though still substantial, figures. Variation of this magnitude limits the usefulness of these estimates when making decisions about resourcing, training, and partnership arrangements with health and social care agencies. It also limits broader policy discussions about where police responsibilities should begin and end, an issue that has gained greater prominence with the national rollout of Right Care, Right Person.

The most readily available source of information for estimating how often UK police encounter vulnerable people is the categorical qualifier flags recorded against incidents in police information systems. However, these were designed to help police manage incidents in real time, not to produce a reliable record of how often vulnerability is present. They are applied under considerable time pressure, and recording practices are known to vary between individuals, between forces, and over time.

Police forces also hold, in considerable volume, unstructured narrative text. Incident logs are written by control room operators and attending officers as events unfold, and describe the circumstances, behaviours, and stated needs of those involved in far greater detail than structured fields allow. This material is likely to provide a more accurate picture of vulnerability-related demand, however, it has historically been difficult to analyse at scale. Manually reviewing and coding thousands of narratives requires substantial time and resources, and earlier automated approaches to text analysis were not sophisticated enough to interpret the information reliably.

Large language models offer a potential solution. Because they can follow written classification instructions and assess ambiguous evidence in context, analysis that previously required expert manual coding becomes practical. Whether their outputs are sufficiently reliable to support estimating the prevalence of vulnerability is a separate question, and one this study addresses alongside the first.

What we did

This study extends earlier research in which we assessed whether LLMs could reproduce human qualitative coding of vulnerability indicators, using publicly available narrative reports from Boston Police Department. That work found the models were highly reliable at identifying reports containing no evidence of vulnerability, but considerably less reliable at identifying its presence.

We applied the same approach to three months of STORM incident logs - records generated by a force's computer-aided dispatch system - comprising just under 3,000 incidents from a single Basic Command Unit. The logs were anonymised within the force before transfer and analysed in a secure research environment. As data of this sensitivity cannot be processed using external services, classification was performed by a locally hosted Llama 3 8B model, fine-tuned for the task.

Because individual classifications vary between runs, each passage of text was classified five times for each vulnerability. We then reviewed 250 passages manually against the model's labels to establish the frequency and direction of classification errors, and used this to statistically adjust the estimates across the full dataset.

Key findings

Our results indicate that LLMs can produce meaningful prevalence estimates from police incident narratives, but that considerable methodological work is required to make those estimates defensible. The model's uncorrected output overstated the presence of every vulnerability examined, and this tendency to overestimate was not immediately apparent from its responses.

Prevalence of vulnerability indicators

Following adjustment, indicators of mental ill health were identified in 22.9% of incidents (95% CI 19.4–26.7%), alcohol dependence in 8.2% (6.3–10.3%), substance misuse in 5.0% (3.7–6.5%), and homelessness in 2.9% (2.1–4.0%). We compared these classifications with the force's own qualifier flags, where both were available, and found substantial disagreement. For mental ill health, a third of incidents with force-assigned flags contained no relevant narrative content, while 220 incidents containing explicit references to mental ill health carried no flag.

Reliability of negative classifications

The model performed consistently well in identifying narratives that contained no evidence of vulnerability. Human review agreed with more than 96% of these classifications across all four vulnerabilities, and the labels were stable across repeated runs. This replicates the principal finding of our earlier study, and indicates that the approach can reliably reduce the volume of material requiring manual review.

Over-sensitivity in positive classifications

Where the model identified evidence of vulnerability, its output was less stable and systematically over-sensitive. Manual review of the model's outputs identified recurring errors, including hotel and hospital stays classified as homelessness, hospital attendance classified as mental ill health, and alcohol consumption classified as substance misuse. Adjustment reduced the estimated prevalence of substance misuse from 12.8% to 5.0%, and homelessness from 9.7% to 2.9%. These errors were not detectable from the model's output alone, and would not have been identified without manual review.

Next steps

Our findings indicate that this approach can generate evidence that is not currently available through any other scalable method, but that its responsible use depends on the methodological framework within which it is applied. Several considerations follow for police forces, partner agencies, and researchers considering similar work.

Application at greater scale

The principal value of this approach lies in its application to larger datasets. Extending the analysis across whole forces and multiple years would support more robust demand modelling, and provide an evidence base for decisions about resourcing, workforce development, and multi-agency safeguarding arrangements. Applying a consistent method across several forces would also make it possible to compare the scale and nature of vulnerability-related demand, supporting benchmarking and strategic planning across the policing system. Analyses of this kind are not achievable through manual coding.

The role of human review

The estimates reported here depend on all three stages of the process described above. The simplest approach is to classify each narrative once, but this can produce different results each time the model is run, with no clear way to identify which classifications are wrong. Classifying repeatedly and aggregating improves stability, but does not address over-sensitivity: the aggregated output placed substance misuse at 12.8%, against an adjusted estimate of 5.0%. A discrepancy of this size could materially distort operational priorities or resource allocation if taken at face value. It was identified only through structured human review, and corrected only through statistical adjustment.

Limitations for operational use

The estimates are reliable when used to describe the overall level of vulnerability across a large population of incidents, because the model's errors follow predictable patterns that can be corrected statistically. For individual incidents, however, there is no comparable way to correct mistakes, and classification errors remain both frequent and unpredictable. Proposals to use LLM outputs to triage, flag, or prioritise individual cases are therefore not supported by these findings, whether based on a single classification or an aggregate of many.

Implications for practice

LLMs are capable of extracting information from unstructured police data at a scale that was not previously achievable. Their outputs do not, however, constitute reliable evidence in themselves. The kinds of error observed here are less visible than those of more familiar research instruments, and are not immediately apparent from the model's responses. Responsible use in this context depends less on model performance than on the rigour of the processes within which the model is embedded.

Authors

The research team were:

Sam Relins and Dan Birks (both University of Leeds).

https://doi.org/10.48785/100/503

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

The support of the Economic and Social Research Council (ESRC) is gratefully acknowledged. Grant reference number: ES/W002248/1.