Probabilistic AI Systems in Law Enforcement Futures

**PROBabLE Futures Policy Brief**

# Don't Let AI Do the Policing: Findings from a Multi-Stakeholder Workshop on Evaluating Large Language Models for Policing

**Nneoma Ogbonna, Mackenzie Jorgensen, Muffy Calder, and Marion Oswald**

## 1. Introduction: The Rationale for Our LLM Evaluation Workshop

The Home Office has announced reforms in UK policing, including the creation of the National Centre for AI in Policing (PoliceAI), driving a major expansion in AI adoption across UK policing.[1] Police forces are already increasingly adopting a range of AI tools, including large language models (LLMs), across multiple stages of the criminal justice process,[2] and the national-level rollout is expected to accelerate and expand this adoption. In policing, LLMs could be used to transcribe and translate interviews, generate witness statements, prioritise and triage information and intelligence, and generate summaries to support investigations and charging decisions.[3]

The adoption of LLMs in policing introduces several risks, such as omissions, misinterpretations and erroneous outputs.[4] These risks can affect case progression, compromise the integrity of evidence and ultimately undermine procedural fairness and public trust in the criminal justice system.[5] The high-stakes nature of policing, where errors can have profound consequences, requires that AI adoption be approached responsibly,[6] and an integral pathway to this is through continuous and holistic evaluation of LLM design, development and adoption within policing workflows.

In November 2025, PROBabLE Futures, the National Police Chief's Council, and the Home Office ran a 1.5 day workshop with stakeholders from academia, government, industry, and policing to discuss what is needed in a LLM evaluation framework for policing use cases, and to test the actual systems in practice through a hackathon. LLM evaluation here refers to the process of assuring LLMs are fit for purpose and testing how well they work within policing workflows to support responsible adoption. **The aim of the workshop was to develop and gather the information required for practical guidance to support police forces to evaluate LLM systems, drawing on the insights and recommendations of stakeholders.** This guidance will inform the rigorous testing of LLMs used in policing, in the interest of justice. Rigorous testing is required for evidence-based decisions about AI adoption.

We proceed by highlighting the LLM use cases discussed and tested during the workshop in Part 2. In Part 3 we present the substantive insights, grouped into five themes: (1) Importance of Robust and Representative Data, (2) Operational Scenarios and Consent, (3)

[1] Home Office, *From Local to National: A New Model for Policing* (White Paper, CP 1489, January 2026) <https://assets.publishing.service.gov.uk/media/69779267276692606c013862/260125_White_Paper.pdf> accessed 11 February 2026.
[2] Calder, M., Oswald, M., McClory-Tiarks, E., Sevegnani, M. and Taka, E., 2026. Responsible AI in criminal justice: LLMs in policing and risks to case progression <https://arxiv.org/pdf/2603.18116> accessed 12 April 2026; Taka, E., Lawal, T., Calder, M., Sevegnani, M., Kotsoglou, K., McClory-Tiarks, E. and Oswald, M., 2025. Mapping the Probabilistic AI Ecosystem in Criminal Justice in England and Wales <https://arxiv.org/pdf/2512.04116> accessed 11 March 2026.
[3] Calder et al n 2.
[4] Calder et al n 2.
[5] Calder et al n 2.
[6] Gunasekara, L., El-Haber, N., Nagpal, S., Moraliyage, H., Issadeen, Z., Manic, M. and De Silva, D., 2025. A systematic review of responsible artificial intelligence principles and practice. *Applied System Innovation*, *8*(4), p.97.

*PROBabLE Futures is supported by the Engineering and Physical Sciences Research Council [Grant Number EP/Y009800/1 AI KP0003] through funding from Responsible AI UK*

Performance Metrics, (4) The (Questionable) Effectiveness of Human Oversight, and (5) Continuous and Holistic Evaluation and Selection of the Right Evaluators. Twenty detailed recommendations, based on these insights and themes, are presented in the Appendix. Part 4 concludes with reflections on what effective LLM evaluation in policing ultimately demands.

## 2. The LLM Use Cases Discussed and Tested

During our workshop, we focused on four LLM use cases, each aligned with a specific LLM system that we had provided for the hackathon:

- Use Case 1 – Translation and transcription: An officer reviewing a suspect's phone download identifies an audio file believed to be relevant to the investigation. As the recording is in a non-English language, the officer uploads it to the translation system to generate a translated transcript.
- Use Case 2 – Transcription: An officer conducts a recorded interview in English with a vulnerable victim and uploads the audio to a transcription system in order to produce an interview transcript for inclusion in the case file.
- Use Case 3 – Video summarisation: An officer attending a public order incident records events using body-worn video and uploads the footage to the summarisation system to generate a concise text summary of what occurred.
- Use Case 4 – Text summarisation and information retrieval: An officer takes over an investigation from a colleague and uses an AI chatbot to quickly understand the key details and context of the case. This enables an effective handover by summarising information from case files in the police record management system.

## 3. Five Key Findings for a Policing LLM Evaluation Framework

The workshop generated a rich body of qualitative insights drawn from participants' discussions and the hackathon. These initial insights have been synthesised into recurring themes, which are discussed in this section and underpin the evaluation and policy recommendations presented in the Appendix. In a subsequent publication, we will outline further themes and findings.

### *3.1. Importance of Robust and Representative Data*

Workshop participants emphasised that using data drawn from actual policing systems, such as Record Management Systems (RMS), is critical for evaluations. **Synthetic or overly "*clean*" datasets may fail to capture demographic or linguistic variability, which can mask errors or biases that would only emerge in practice**.[7]

Test data for evaluations must reflect **real-world conditions**, particularly where application contexts differ from the data used to train models. Participants highlighted that, to reflect real-world conditions, **test data must capture a wide range of variation, including edge cases and common scenarios**. As one participant in the transcription use case no. 2 explained:

[7] Gholami, S. and Omar, M., 2023. Does synthetic data make large language models more efficient? *arXiv preprint arXiv:2310.07830*.

*PROBabLE Futures is supported by the Engineering and Physical Sciences Research Council [Grant Number EP/Y009800/1 AI KP0003] through funding from Responsible AI UK*

> *"Once you start to add in things like regional accent, things like emotional state, shouting, whispering, multiple speakers, background noise, less predictive speech, ungrammatical speech, disfluent speech, that sort of thing, all of those factors are going to contribute to impact the performance."*

When real data drawn from actual policing systems are used to test the LLM systems, police forces must ensure compliance with data protection principles, such as the requirements that processing be lawful, fair, necessary, and proportionate.[8]

### *3.2. Operational Scenarios and Consent*

LLMs should also be tested in operational scenarios where they interact with live workflows; as one participant put it, *"testing in a genuine real environment".* This should be done in a layered approach, **beginning with less sensitive cases, such as shoplifting, where risks to individuals are lower, before moving to higher-stakes scenarios, like sexual offences**. As a participant recommended: *"Let's start with a specific type of crime where clearly the person is not upset. It is not very personal. It is not very complicated."*

When testing and deploying LLM systems in live workflows, the need for **consent** was emphasised. *"First of all, you should ask consent"* from individuals whose data is being processed, such as when a victim calls to report a crime and the interaction may be transcribed or handled by an AI system.

Consent from individuals must be "freely given, specific, informed and unambiguous",[9] but attaining this can be challenging in high-stress or sensitive contexts, such as use case no. 2. For one, individuals in distress may not be in a position to fully understand or freely consent to their data being processed by AI systems:

> *"I would bet that most people in the forces don't even know what AI is doing, so let alone people calling. If it's in relation to serious sexual offences or offenders, you're going to have far more problems."*

Consent can and should also be easily withdrawn,[10] which may complicate the testing and use of LLM systems in live operational scenarios. One attendee remarked: *"If you implicitly consent, which of course is the easiest to remove, what happens if they say, 'Actually, I'm now withdrawing my consent'?"* The potential for individual consent to be easily withdrawn suggests that alternative legal bases for processing, such as necessity for policing functions,[11] may need to be considered in collaboration with legal advisers.

**Relying on necessity as the legal basis for data processing must remain consistent with the broader principle of "policing by consent",[12] under which the legitimacy of policing derives not from the explicit consent of each individual, but from the common trust and approval**

[8] The first to sixth data protection principles for the processing of personal data for any law enforcement purposes are specified in sections 35, 36, 37, 38, 39 and 40 of the Data Protection Act, 2018.
[9] Section 33(1A) Data Protection Act, 2018; Section 69(2) Data (Use and Access) Act 2025.
[10] Section 40A(6) Data Protection Act, 2018; Section 69(4) Data (Use and Access) Act 2025.
[11] Section 35(2)(b) Data Protection Act 2018.
[12] Jackson, J., Bradford, B., Hough, M. and Murray, K., 2012. 2 Compliance with the Law and Policing by Consent: Notes on police and legal legitimacy. In *Legitimacy and compliance in criminal justice* (pp. 29-49). Routledge; Robertson, A., 2016. *Policing by consent: Some practitioner perceptions* (Doctoral dissertation, University of Sunderland).

*PROBabLE Futures is supported by the Engineering and Physical Sciences Research Council [Grant Number EP/Y009800/1 AI KP0003] through funding from Responsible AI UK*

**of the public.** Maintaining that trust requires **transparency**. Where LLM systems are tested in live workflows, both **the wider public and individuals who interact directly with the police must be clearly informed about how their data will be used in the development and testing of these systems**.

### *3.3. Performance Metrics*

Measuring performance is imperative when LLMs are used in high stakes policing environments. **Errors from LLMs can mislead investigations, impede justice outcomes and compromise public trust.**[13] This risk is exacerbated by the tendency of LLMs to appear confident and coherent even when factually incorrect or using inappropriate sources.[14] As one participant in the text-summarisation use case no. 4 explained:

> *"It may come up with something that isn't correct and if that does follow all the way through to handover, the next officer has that information, that is just going to persist…."*

Several participants acknowledged that assessing technical performance of LLMs in policing contexts is particularly challenging because LLM performance is highly context-dependent and language-related tasks are prone to subjectivity,[15] making it difficult to define a clear "ground truth". One participant queried, *"What does good enough look like from a human perspective?"* For example, what constitutes an ideal summary may vary across officers. Nonetheless, there was consensus that human experts should serve as the baseline benchmark for evaluating the performance of LLMs, as they reflect the standards and expectations of policing practice.

Even in the evaluation of LLM outputs, such as transcripts, ***"different metrics tell you different stories".*** For example, word error rate (WER) can be used to calculate the percentage of errors of a transcript generated by a speech recognition model, by comparing that transcript to a "ground-truth" reference.[16] While the WER metric is useful as a simple, standardised measure of word-level accuracy, it *"raises a red flag."* As one participant said:

> *"If you're using a typical metric like word error rate… you're going to have loads and loads of errors…."*

This is because WER treats complex errors, such as missed speech, incorrect speaker attribution, or poor segmentation, as basic word-level mistakes.[17] As a result, it inadequate for evaluating performance in realistic, multi-speaker scenarios that reflect the variability and edge cases seen in actual policing contexts. The WER metric should not be used on its own and the errors it finds need manual review by human experts to understand the severity of the errors.

---

[13] Calder et al n 2; Gohil, N., 2026, January. West Midlands police chief apologises after AI error used to justify Maccabi Tel Aviv ban <https://www.theguardian.com/uk-news/2026/jan/14/west-midlands-police-chief-apologises-ai-error-maccabi-tel-aviv-ban> accessed 11 February 2026.

[14] Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S., 2021, March. On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

[15] Calder et al n 2; Basile, V., Caselli, T., Balahur, A. and Ku, L-W., 2022. Editorial: Bias, Subjectivity and Perspectives in Natural Language Processing. *Frontiers in Artificial Intelligence*, 5, p.926435.

[16] von Neumann, T., Boeddeker, C., Delcroix, M. and Haeb-Umbach, R., 2025. Word Error Rate definitions and algorithms for long-form multi-talker speech recognition. *IEEE Transactions on Audio, Speech and Language Processing* <https://arxiv.org/pdf/2508.02112> accessed 11 February 2026.

[17] von Neumann et al n 16.

Other metrics such as Bilingual Evaluation Understudy (BLEU),[18] Recall-Oriented Understudy for Gisting Evaluation (ROUGE)[19] and Bidirectional Encoder Representations from Transformers (BERTScore),[20] also have limitations, as they can reduce complex, context-dependent errors to measures of surface or semantic similarity, without fully capturing meaning, intent or impact.[21] BLEU is mainly used in translation and assesses how closely a system's output matches a reference text.[22] ROUGE is commonly used in summarisation tasks, as it measures how much of the reference content is captured.[23] BERTScore assesses whether different texts convey the same meaning, even if their wording differs.[24]

Overall, participants highlighted that **evaluating LLM systems requires more than simply ascribing numerical measures to transcripts, translations or summaries.** As one participant in the transcription use case no. 2 stated:

> *"95% accuracy sounds good, but in that 5% it could be the difference between a confession and something innocuous … 80% accurate could be just missing ums and ahs."*

This quote highlights the need to investigate what the errors actually are and if they are higher risk (e.g., missing or making up a confession) or lower risk (e.g., missing an "um"). The **impact and severity of errors should always be considered, as they may have serious consequences that evaluation metrics cannot capture.**

Furthermore, the design of performance evaluations must begin before LLM deployment and continue throughout adoption. **Measures of success and agreement on what is "good enough" performance must be defined in the early stages of designing or procuring the system.** As one participant shared, *"you really need to understand how you're going to measure success from the beginning."*

### *3.4. The (Questionable) Effectiveness of Human Oversight*

It is widely assumed that keeping a human "in the loop" is sufficient to prevent and detect errors propagated by LLM systems.[25] However, we know from many audits and industry failures, that despite the integration of human oversight, AI systems can cause automation bias, whereby the user trusts or accepts the output because it was automatically generated.[26] **Human oversight, is a fragile safeguard[27] that does not succeed in a vacuum; it is shaped by**

[18] Papineni, K., Roukos, S., Ward, T. and Zhu, W.J., 2002, July. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).
[19] Zhou, Q., Yang, N., Wei, F., Huang, S., Zhou, M. and Zhao, T., 2018, July. Neural document summarization by jointly learning to score and select sentences. In *Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long papers)* (pp. 654-663).
[20] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q. and Artzi, Y., 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
[21] Fabbri, A.R., Kryściński, W., McCann, B., Xiong, C., Socher, R. and Radev, D., 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, *9*, pp.391-409; Sai, A.B., Mohankumar, A.K. and Khapra, M.M., 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, *55*(2), pp.1-39.
[22] Fabbri et al n 20; Papineni et al n 17.
[23] Fabbri et al n 20; Zhou et al n 18
[24] Fabbri et al n 20; Zhang et al n 29
[25] Beck, J., Eckman, S., Kern, C. and Kreuter, F., 2025. Bias in the loop: How humans evaluate AI-generated suggestions. *arXiv preprint arXiv:2509.08514*.
[26] Beck et n 25.
[27] Rinta-Kahila, T., Penttinen, E., Salovaara, A., Soliman, W. and Ruissalo, J., 2023. The vicious circles of skill erosion: A case study of cognitive automation. *Journal of the Association for Information Systems*, *24*(5), pp.1378-1412; Romeo, G. and Conti, D., 2026. Exploring automation bias in human–AI collaboration: a review and implications for explainable AI. *AI & Society*, *41*(1), pp.259-278.

*PROBabLE Futures is supported by the Engineering and Physical Sciences Research Council [Grant Number EP/Y009800/1 AI KP0003] through funding from Responsible AI UK*

**system design, workload pressures and cognitive limitations.**[28] Discussions from our workshop reinforce this concern about automation bias and the unclear effectiveness of human oversight. As one participant expressed:

> *"We're talking about giving the officer a tool to do their job faster but it's the overreliance that's the risk."*

Meanwhile, participants emphasised that **human review is essential because what LLMs *"find important"* may diverge from what experienced officers consider operationally significant.** Others stressed that **policing relies on intuition and contextual knowledge, such as familiarity with local patterns, that cannot be readily encoded or reproduced by LLMs.** This highlights the need to move beyond generic notions of oversight and precisely define who the 'human' is, when they intervene, and what their role entails. This challenge is captured by one participant who asked:

> *"How do you ensure a human remains in the loop. Which human? When are they in the loop? So, lots of questions around that. Can't just be any human."*

Seasoned officers with extensive experience are likely better skilled to determine the right prompts for LLMs and extract meaningful information, using their tacit knowledge to make informed judgments. As one participant stated, *"seasoned officers will know the right things to ask to trigger the system."* **Deployment should therefore prioritise highly experienced officers steering the system to best support *their* police work**. At the same time, **less-experienced officers should be supported in developing their skills, both within and outside of LLM integration, before being allowed to rely fully on a LLM for aspects of their work.**

Participants also highlighted the importance of structuring the LLMs to perform tasks in ways that support effective human oversight. LLMs should prioritise **flagging, triaging, or organising** information, while **humans determine meaning and actions to follow**. This requires **building in a degree of friction or cognitive effort as a feature of a model rather than a flaw,[29] encouraging officers to actively engage with and construct their own interpretation of outputs rather than passively accepting them**. In the video summarisation use case, participants showed a **strong preference for structured, time-stamped (therefore sourced and verifiable) bullet points over polished narratives, which may be misleading.** As one participant explained:

> *"You'd rather have the points than an essay. I think because... we invent narratives. We'll read them and we'll infer things from them. Whereas if what we've got is just plain bullet points, we're less likely to construct that narrative."*

This preference for structured output over narratives was reinforced by another participant who added: *"you'd get at least a sort of a contour map of the recorded data, but without any*

[28] JUSTICE, *AI in Policing: International Lessons and Domestic Solutions* (December 2025) <https://cdn.prod.website-files.com/67becde70dae19a9e5ea2bc3/692de2e6bba4243c8489b11c_AI%20in%20Policing%20Report%20-%20Dec%202025.pdf> accessed 11 February 2026; Rinta-Kahila et al n 27; Lamchek, J. and Trieu, V.H., 2026. Risk Regulation of Generative AI: A Case Study of Microsoft Copilot in the Australian Government. <https://doi.org/10.71265/esg7cf11> accessed 11 February 2026.

[29] Cabitza, F., Natali, C., Famiglini, L., Campagner, A., Caccavella, V. and Gallazzi, E., 2024. Never tell me the odds: Investigating pro-hoc explanations in medical decision making. *Artificial intelligence in medicine*, *150*, p.102819.

*PROBabLE Futures is supported by the Engineering and Physical Sciences Research Council [Grant Number EP/Y009800/1 AI KP0003] through funding from Responsible AI UK*

*narrative placed upon it."* **Overly interpretive outputs risk directing police attention toward certain elements while obscuring others and may encourage overreliance on the system rather than independent judgement**. In addition, LLM systems should not include moral judgement in their outputs for policing as this could mislead or bias the officers who are investigating a case. Outputs should be factual and descriptive, e.g., "the suspect walked in the room at 5:03pm", rather than interpretive, e.g., "the suspect walked suspiciously into the room". Commercial providers should apply these insights when **designing** LLM systems used for policing, building tools that actively support and augment human oversight rather than subtly displacing it.

Furthermore, overreliance is closely intertwined with deskilling.[30] **Police forces should ensure a balanced use of LLMs by providing clear guidance on the types of tasks they may be used for.** In the video summarisation use case no. 3, one participant stated, ***"Let it not do the policing but use it as an information sieve."*** The skills of police officers should be continuously engaged to prevent 'skill atrophy'.[31] This could be supported through policies requiring officers to complete the task on their own, periodically, without using the LLM system. As participants emphasised, **where officers become dependent on the technology, their work may be hampered if the system fails, is unavailable or '*broke*'.**

Police forces should also be aware that **time-poor, inexperienced or unskilled users may more easily slip into irresponsible usage**,[32] potentially relying on LLM outputs without adequate verification. Operational guidelines should therefore account for these factors when prescribing how LLM systems are integrated into workflows. Where human oversight is not actively safeguarded, one participant expressed concern that there is a risk that *"over time, the quality of services gets eroded".*

### *3.5. Continuous and Holistic Evaluation and Selection of the Right Evaluators*

Evaluation must be embedded throughout the AI development, deployment and post-deployment phases, ensuring that system performance, workflow impacts, and operational risks are monitored continuously rather than assessed only at a single point.[33] Continuous evaluation and monitoring is therefore essential to ensuring the responsible integration of LLMs into policing workflows. If the LLM and workflow are no longer working as they should, the force should then seriously consider taking it offline.

**LLM evaluation also requires a holistic approach[34] that extends beyond technical accuracy, encompassing workload effects, impacts on case progression,[35] and implications for fundamental rights.** Police forces must consider the *"systems-wide impact"* of LLM adoption rather than the *"niceties of how it works."* What constitutes 'good' adoption must therefore be clearly defined, reflecting the wider picture. As one participant noted:

[30] Rinta-Kahila et al n 27.
[31] Ganuthula, V.R.R., 2024. The paradox of augmentation: A theoretical model of AI-induced skill atrophy. *Available at SSRN 4974044*. <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4974044> accessed 15 April 2026.
[32] Lamchek and Trieu n 28.
[33] Suresh, H. and Guttag, J., 2021, October. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1-9).
[34] Lamchek and Trieu n 28; Suresh and Guttag n 33.
[35] Calder et al n 2.

*PROBabLE Futures is supported by the Engineering and Physical Sciences Research Council [Grant Number EP/Y009800/1 AI KP0003] through funding from Responsible AI UK*

> *"It's not clear what good looks like. Is it efficiency or is it outcomes? I think that's (what) needs to be decided. Because if it's outcomes, then the data needed isn't just the deployment of the tool, but it's use in… it's long-term use, yeah, it's systems-wide impact."*

Another participant reinforced this system-level perspective:

> *"You'd probably want to have a look at the entire process… not just how does this do but you might need to say well how does this affect the entire process, and the sort of process engineering perspective."*

To drive continuous and holistic evaluation, the appropriate evaluators need to be involved. A key concern raised by participants was, ***"Are we getting the right people evaluating?"*** This question must be carefully addressed, as effective and thorough evaluation depends on having the appropriate experts at the forefront. It is, as one participant noted: *"A lot of work to test properly!"*

Actors within the justice system, such as police officers, are particularly valuable evaluators. However, participants emphasised that **evaluators should not be limited to law enforcement personnel.** Concerns were raised about *"potential sabotage by evaluators protecting their jobs",* thus highlighting the need for external perspectives to mitigate potential biases. **Prosecutors from the Crown Prosecution Service (CPS) and lawyers are important potential evaluators, given their positions within the criminal justice system and considering their understanding of justice and evidential impacts, as well as operational realities.** Their expertise makes them well placed to identify and categorise the risks of errors and assess the potential downstream impacts of LLM systems on justice outcomes. Likewise, **evaluators from academia, industry, and the third sector can provide valuable socio-technical contributions, bringing diverse insights and mitigating potential biases.** An ideal collaborative evaluation arrangement could involve independent actors from academia, industry, and the third sector who provide the scaffolding for testing the LLM system, while the police forces introduce the real-world, sensitive data for the evaluation.

Furthermore, **participatory approaches should be adopted, bringing in perspectives from affected populations, especially where AI use may disproportionately impact specific communities (e.g., survivors of gender-based violence).**[36] This mixed stakeholder approach helps ensure that errors are assessed with a well-rounded awareness of legal and operational contexts, ultimately strengthening evaluation and contributing positively to justice outcomes.

## 4. Conclusion

Unlike conventional roundtables, our workshop extended beyond discussion to include hands-on use of real-world LLM systems. Workshop participants engaged directly with LLM systems through a hackathon, testing these tools in use-case scenarios reflecting how they may be used in real-world policing. This hands-on engagement, alongside conversations,

[36] Dunkwu, J., Paterson-Young, C. and Maher, M., 2025. Scoping review of frameworks for effective participatory oversight of AI in Law Enforcement. (PROBabLE Futures 2025) <https://research.northumbria.ac.uk/probablefutures/wp-content/uploads/2025/12/Participatory-Framework-Scoping-Report-05122025.pdf> accessed 12 February 2026; Birhane, A., Isaac, W., Prabhakaran, V., Diaz, M., Elish, M.C., Gabriel, I. and Mohamed, S., 2022, October. Power to the people? Opportunities and challenges for participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1-8).

*PROBabLE Futures is supported by the Engineering and Physical Sciences Research Council [Grant Number EP/Y009800/1 AI KP0003] through funding from Responsible AI UK*

formed the basis for practical guidance aimed at supporting police forces in adopting reliable LLM systems.

It goes without saying that the introduction of LLM systems into policing is not simply a matter of technological capability or efficiency. These systems shape how decisions are made, affect case progression, and have direct implications for the integrity of evidence and procedural fairness, and as such, bear on fundamental rights and freedoms.

On this basis, effective evaluation must be rigorous, continuous and holistic. Technical performance metrics, while necessary, are insufficient on their own. Evaluation must extend to operational realities, human oversight, workflow impacts, and system-level outcomes, and must be grounded in data that reflects the full diversity of people and circumstances that policing encounters. Throughout this process, transparency is essential.

Effective evaluation is a collective responsibility. Police forces, PoliceAI, legal professionals, independent experts and affected communities must all work together on LLM evaluations. Ultimately, how LLM systems are evaluated in policing is a matter of justice and of public trust in law enforcement and criminal justice system, therefore, evaluation must be accorded the resources and due diligence that it demands.

**Appendix**

*Recommendations: Importance of Robust and Representative Data*

i. Police forces should draw on **patterns observed in real cases and operational experience** to identify the kinds of edge cases, and challenges systems may encounter in practice. This can help inform decisions about the data needed for testing LLM systems, ensuring it reflects real-world policing.

ii. LLM systems should be evaluated to determine how different quality or types of data (real-world or synthetic) affect model outputs across variables such as accent, dialect, and other demographic or contextual factors. This can help shape the evaluation criteria needed to determine whether a system is fit for deployment.

iii. Wherever possible, police forces should **prioritise real-world policing data when testing LLM systems**, while ensuring compliance with data protection principles, such as the requirements that processing be lawful, fair, necessary, and proportionate.

iv. PoliceAI should develop mechanisms that enable police forces to evaluate LLM systems **collaboratively**, including through shared repositories of anonymised real-world or synthetic evaluation data.

*Recommendations: Operational Scenarios and Consent*

v. Police forces should adopt a **layered approach to operational testing**, beginning with lower-sensitivity case types (such as shoplifting), before progressing to higher-stakes scenarios (such as sexual offences), allowing risks to be identified and managed incrementally.

*PROBabLE Futures is supported by the Engineering and Physical Sciences Research Council [Grant Number EP/Y009800/1 AI KP0003] through funding from Responsible AI UK*

vi. Police forces should ensure individuals interacting directly with the police are clearly informed when AI systems are involved in handling their data, for example through verbal disclosure at the point of contact and online transparency information. Where required (such circumstances should be determined in consultation with legal advisers) AI tools should only be used where consent has been given freely.

vii. PoliceAI and police forces should jointly ensure that the use of LLM systems in policing is communicated transparently to the public and that there is public awareness of how personal data may be used in the testing and development of LLM systems.

*Recommendations: Performance Metrics*

viii. PoliceAI and police forces should d**efine measurements of LLM success from the beginning** and carry out performance evaluation before LLM deployment and continue throughout adoption.

*ix.* Performance metrics should reflect operational standards, benchmarking AI outputs against human performance within policing workflows.

*x.* Evaluators should apply harm-based weighting of errors, so that those with serious investigative or justice implications receive closer scrutiny, without discounting the significance of other deficiencies.

*xi.* Evaluators should also stress-test LLM systems against edge cases to expose failure modes.

*Recommendations: The (Questionable) Effectiveness of Human Oversight*

xii. Evaluation should extend to the operational impact of LLM use to fully understand how human oversight shapes output quality in practice. This includes assessing how officers engage with, interpret, and act on system outputs, beyond considerations of efficiency gains or technical performance. Evaluators should also consider the extent to which officers meaningfully exercise the *"final say",* rather than simply deferring to LLM outputs.

xiii. **Evaluators should not solely focus on productivity or time-based metrics to decide if an LLM system is worthwhile and working as it should.** Time-based metrics can be misleading when assessing the effectiveness of human oversight. In the early stages of deployment, workload may temporarily increase as officers learn to interrogate, verify, and contextualise AI outputs; this should be understood as a necessary part of responsible use rather than a failure.

xiv. Evaluators should **be cautious of treating time savings as evidence of successful adoption**, as faster workflows may reflect uncritical acceptance of LLM outputs. Conversely, the absence of time savings may indicate that officers are engaging in reflective thinking through LLM use, thus supporting their training.

*PROBabLE Futures is supported by the Engineering and Physical Sciences Research Council [Grant Number EP/Y009800/1 AI KP0003] through funding from Responsible AI UK*

xv. PoliceAI must set clear design standards for commercial providers to build in **deliberate frictions** that actively encourage meaningful human oversight. For example, in the video summarisation use case, developers could structure outputs as **time-stamped bullet points** rather than narrative summaries to prompt officer reflection rather than passive acceptance.

*xvi.* PoliceAI and police forces **should jointly establish** operational guidelines specifying **what meaningful human oversight requires in practice,** when and how officers should use LLM tools, and which officers should use them. Officers with extensive operational experience should be given priority access to LLM systems as they may be more skilled to critically assess outputs. PoliceAI should provide guidance on the skills that should be developed by police officers before being permitted to use LLM tools in their work.

xvii. PoliceAI and police forces should take proactive measures to guard against skill erosion and atrophy, including policies that require officers to periodically complete tasks on their own without LLM assistance, allowing them to exercise their skills.

*Recommendations: Continuous and Holistic Evaluation and Selection of the Right Evaluators*

xviii. PoliceAI should establish clear evaluation benchmarks and metrics for a variety of policing use cases. These should reflect a holistic view of system performance, extending beyond technical accuracy to include workload effects, impacts on case progression, and implications for fundamental rights.

xix. PoliceAI should clarify the roles, responsibilities, and best practices for evaluators, ensuring evaluation draws on a diverse range of expertise. This should include experienced officers, legal professionals such as prosecutors and lawyers, and independent experts from academia, industry, and the third sector. Participatory approaches should also be mandated, bringing in perspectives from affected communities, particularly where LLM use may disproportionately impact specific groups.

xx. PoliceAI should set clear standards and timelines for continuous evaluation across development, deployment, and post-deployment phases, ensuring that system performance, workflow impacts, and operational risks are monitored on an ongoing basis and formally documented.

*PROBabLE Futures is supported by the Engineering and Physical Sciences Research Council [Grant Number EP/Y009800/1 AI KP0003] through funding from Responsible AI UK*