Technical Comments on Guidelines 02/2026 on Anonymisation

Automated PII detection and agentic access controls in anonymisation workflows

Rohith Uppala | Independent AI Security Researcher

August 19, 2026

Executive summary

I welcome the European Data Protection Board's practical framework based on No Record Isolation, No Linkage, and No Inference. I offer a narrow technical comment on automated personally identifiable information (PII) detection used during anonymisation of unstructured text.

Automated PII detection can help identify simple cases under paragraphs 50-51, but it is not itself proof of anonymisation. It addresses only some explicit or context-dependent identifiers and does not establish No Record Isolation, No Linkage, or No Inference. Its errors also vary materially across domains, languages, scripts, and PII types.

I recommend that the final Guidelines:

1. State expressly that automated PII detection or redaction is an input to, not a substitute for, the three-criteria anonymity assessment.
2. Specify minimum validation and documentation requirements when automated detectors are used.
3. Require performance to be disaggregated by relevant domain, language, script, and PII type.
4. Require residual-risk testing on representative post-transformation data, including linkage and inference testing.
5. Require periodic revalidation after changes in models, thresholds, input distributions, intended recipients, or reasonably available re-identification capabilities.
6. Distinguish prompt-based behavioral restrictions from independently enforced authorization when agentic AI can access personal source data, auxiliary datasets, or analysis tools.

These additions would operationalise paragraphs 36, 40-44, 50-51, 83-88, and 92-93 without changing the legal standard articulated in the Guidelines.

Commenter's relevant research and scope limitation

My comments draw on an independent evaluation of OpenAI's Privacy Filter (OPF), a 1.5 billion parameter PII detector, across 32 benchmarks, 14 languages, and five domains. The study compared zero-shot OPF with other PII and entity-detection systems and examined domain transfer, script effects, PII-type errors, and fine-tuning data requirements.

The study did **not** evaluate whether any dataset satisfied the GDPR standard for anonymisation. It did not test the complete No Record Isolation, No Linkage, or No Inference framework. Its PII-annotated benchmarks contained synthetic or machine-generated PII. I therefore cite it only as empirical evidence that an automated detector's ability to find identifying spans is conditional and must be validated before it is relied upon in an anonymisation workflow.

Comment 1: Clarify that automated PII detection is not proof of anonymisation

Relevant provisions: paragraphs 40, 43-44, and 50-52.

Paragraph 50 correctly identifies names, email addresses, telephone numbers, and similar direct identifiers as simple cases in which information is clearly not anonymous. Paragraph 51 recommends early data mapping. In practice, controllers may use named-entity recognition, PII detectors, pattern matching, or generative models to perform this mapping on free-form text.

The final Guidelines should prevent a common category error: concluding that text is anonymous because an automated tool did not detect any PII, or because all detected spans were redacted. A detector's negative output is not evidence that the text contains no identifying information unless the tool has been validated for the relevant input distribution. Even a highly accurate direct-identifier detector would not establish that the transformed data satisfies No Record Isolation, No Linkage, and No Inference.

Suggested insertion after paragraph 51

> Where automated tools are used to identify or remove direct or indirect identifiers, their output should be treated as one input to the assessment and not as proof that the given data is anonymous. The absence of a detected identifier may reflect a limitation of the tool, its training data, its label schema, or its application to a different domain, language, script, or data format. The transformed data should still be assessed under No Record Isolation, No Linkage and No Inference, using representative residual-risk testing.

Comment 2: Define minimum validation evidence for automated detectors

Relevant provisions: paragraphs 41, 43-44, and 83-87.

Paragraph 41 appropriately requires documentation of the anonymisation process and testing of supposedly anonymous datasets. Paragraphs 83-87 require assessment of re-identification techniques and their accuracy. Comparable specificity would be valuable for automated preprocessing tools that determine which identifiers are removed before the three criteria are applied.

In the cited evaluation, OPF achieved span F1 of 0.855 on AI4Privacy but 0.464 on synthetic medical PII. Performance dropped to 0.04-0.40 on general named-entity recognition benchmarks. Person-name detection was approximately 0.04 F1 on Arabic script and 0.03 on Cyrillic. Structurally regular PII was easier to detect: email and telephone numbers achieved aggregate F1 of 0.78 and 0.76, while person names and addresses achieved 0.40 and 0.49.

These figures come from different evaluation settings and are not directly comparable: AI4Privacy is a structured synthetic PII benchmark, whereas the Arabic- and Cyrillic-script results use PER-only MultiCoNER general-NER splits to examine person-name detection under domain and script shift.

These results do not establish expected performance on any controller's data. They demonstrate why an aggregate score from a different benchmark should not be accepted as deployment evidence.

Suggested addition to paragraph 41

When an automated detector or redaction model is used, retained documentation should include at least:

- The model, software, rules, and version used.
- The PII or identifier taxonomy the component can and cannot detect.
- Thresholds, decoding settings, and any human-review rules.
- The provenance and representativeness of validation data.
- Precision, recall, and false-negative counts with sample sizes and uncertainty, where feasible.
- Results disaggregated by relevant domain, language, script, data source, document type, and identifier type.
- Known unsupported conditions and the escalation procedure for them.
- The date of evaluation and the events that require revalidation.

Recall is particularly important where a missed identifying span can defeat the intended transformation. Precision should also be recorded because excessive redaction can remove useful information and may obscure whether the transformation remains proportionate and fit for its stated purpose. Neither metric should be considered a substitute for testing linkage and inference risk in the resulting data.

Comment 3: Require representative residual-risk testing after transformation

Relevant provisions: paragraphs 43-44, 51-53, and 83-88.

Validation of the detector and testing of the resulting data answer different questions. Detector validation asks whether a component correctly finds labelled identifying spans. Residual-risk testing asks whether the post-transformation data permits record isolation, linkage, or specific and meaningful inference by relevant entities using means reasonably likely to be used.

The Guidelines should make this distinction explicit. For unstructured text, a useful test set should preserve difficult cases such as culturally variable names and addresses, narrative references, indirect identifiers, rare attribute combinations, mixed scripts, transliteration, misspellings, and identifiers embedded in technical or domain-specific language. Testing only regular patterns such as email addresses and telephone numbers can give a misleading picture of residual risk.

Suggested insertion in subsection 3.5.1

> Where anonymisation includes automated detection or redaction, the effectiveness of that component should be tested on representative pre-transformation data, and the resulting data should be separately

tested for residual record isolation, linkage and inference. Testing should include difficult and outlying cases and should reflect the languages, scripts, domains, formats and identifier types reasonably expected in use. A test limited to average performance or common identifier types may fail to reveal a technique that succeeds against a single individual, as described in paragraph 87.

Comment 4: Disaggregate results and define escalation paths

Relevant provisions: paragraphs 36, 44, 47-49, 83-88, and 94.

The contextual approach requires attention to the capabilities of relevant entities. The same discipline should apply to the transformation technology. A tool can be reliable for one subgroup of data and unreliable for another. Aggregation can conceal this difference.

The cited evaluation found a recall-biased profile in several PII domains, with precision between 0.31 and 0.54 and recall between 0.70 and 0.85. This means that the practical cost of errors differed by application: missed spans create privacy risk, while false positive redactions reduce data utility. The correct operating threshold and review policy therefore depend on the processing context.

Paragraph 36 states that a dataset should be considered anonymous only where anonymisation is effective for all included individuals. The cited evaluation found person-name F1 of 0.038 for Persian written in Arabic script and 0.027 for Ukrainian written in Cyrillic on PER-only MultiCoNER splits. These results do not establish that any dataset failed the GDPR anonymity test, and they do not measure a complete anonymisation process. They demonstrate why a controller should not infer effectiveness for every subgroup from aggregate or Latin-script validation. Materially represented languages and scripts should be validated separately, with a documented fallback for subsets outside the validated operating range.

Controllers should define an escalation path for inputs outside the validated operating range. Appropriate responses may include restricted human review, routing to a different validated detector, applying multiple complementary methods, withholding release, or treating the output as personal data. Unsupported inputs should not silently inherit the performance claim of the overall system.

Suggested clarification

> Validation results should be disaggregated wherever aggregate results could conceal a materially higher residual risk for a relevant subset of the data. Inputs outside the validated operating range should be identified and subject to a documented fallback, escalation or non-release procedure.

Comment 5: Require periodic revalidation and change control

Relevant provisions: paragraphs 35, 42, 88, and 92-93.

The Guidelines correctly recognise that relevant entities' capabilities and the cost of re-identification can change over time. Automated transformation systems also change. Model updates, new thresholds, altered label schemas, new source systems, data drift, and additional languages can invalidate earlier performance evidence even when the nominal product name remains the same.

The cited study also found that model choice can change after adaptation to representative labelled data. Fine-tuned XLM-RoBERTa surpassed zero-shot OPF with approximately 500 labelled examples on English synthetic PII, approximately 100 examples on non-English Kiji customer-support data, and approximately 1,000 examples on synthetic medical PII. These counts are not universal thresholds and should not be transplanted to other contexts. They illustrate that local validation and adaptation may materially change the appropriate technical control.

Suggested addition to paragraph 88

> Reassessment should also be triggered by material changes to an automated anonymisation component or its operating environment, including changes to the model or rules, thresholds, identifier schema, input distribution, language or script coverage, intended recipients, accessible auxiliary data, or available re-identification techniques. Controllers should retain versioned evidence showing which assessment applies to each released dataset.

Comment 6: Distinguish behavioral restrictions from enforced authorization for agentic AI

Relevant provisions: paragraphs 28 and 92-93.

Paragraph 92 recognises that agentic AI may reduce the time and cost of deploying re-identification techniques. Where such agents can access original personal data, auxiliary datasets, or analysis tools, controllers should document whether those access boundaries are enforced independently of model output.

A separate study evaluated unauthorized tool selection, not re-identification, anonymised datasets, or personal-data access. Across three language models and 200 adversarial tasks, explicit prompt-based allowlists left Unauthorized Invocation Rate between 4% and 37%. A governed MCP proxy that filtered unauthorized tools at discovery and checked authorization again at invocation produced 0% UIR across the evaluated models and tasks under the study's stated assumptions, including a trusted identity issuer, uncompromised registry, and correctly enforced policy.

This is adjacent evidence rather than evidence that a particular dataset is or is not anonymous. It supports a limited governance distinction: behavioral instructions to an agent are not equivalent to technical authorization controls enforced independently of the agent's output.

Suggested addition to paragraph 92

> Where agentic AI can access personal source data, auxiliary information, or tools capable of linkage or inference, the assessment should document how those access boundaries are enforced. Instructions expressed only through prompts or model-readable policies should be distinguished from technical controls that independently restrict data and tool discovery and verify authorization before execution.

Conclusion

The draft Guidelines provide a strong legal and technical structure for determining anonymity. The recommended additions would make that structure more operational for organisations using automated systems on unstructured and multilingual data.

The central principle is simple: **automated PII detection can support data mapping and transformation, but it cannot establish anonymity by itself.** Controllers should validate the component on representative inputs, document disaggregated errors, test the transformed data against No Record Isolation, No Linkage, and No Inference, and revalidate when the system or its context changes.

Where agentic AI operates on or near the relevant data, controllers should also document whether access to personal source data, auxiliary datasets, and consequential tools is limited by controls enforced independently of model output.

References

European Data Protection Board. (2026). *Guidelines 02/2026 on Anonymisation*, Version 1.0, adopted 7 July 2026 for public consultation.

Uppala, R. (2026). ”OpenAI Privacy Filter: A Cross-Lingual, Cross-Domain PII Evaluation Across 32 Benchmarks.” *arXiv preprint* arXiv:2608.02616v2. https://arxiv.org/abs/2608.02616. https://doi.org/10.48550/arXiv.2608.02616

Uppala, R. (2026). ”Prompts Don't Protect: Architectural Enforcement via MCP Proxy for LLM Tool Access Control.” *arXiv preprint* arXiv:2605.18414. https://arxiv.org/abs/2605.18414