Comment on the EDPB Template for Personal Data Breach Notification

Submitted by: Meem Arafat Manab
Affiliation:
Marie Skłodowska-Curie Doctoral Fellow, Ontology Engineering Group, Universidad Politécnica de Madrid (HARNESS Doctoral Network, Horizon Europe);
Research Associate (pro bono), School of Religion, Theology, and Peace Studies, Trinity College Dublin
Consultation: Template for Personal Data Breach Notification, adopted 8 June 2026
Deadline: 5 August 2026

Summary

These comments address three gaps in the template that are not incidental but structural: they concern a category of personal data breach that the template, as drafted, cannot capture. The category is the exposure of personal data absorbed in AI model weights, whether through deliberate adversarial attack or incidental query-induced regurgitation. Two authoritative sources establish that this is not a theoretical risk. The EDPB itself, in Opinion 28/2024, determined that AI models trained on personal data cannot, in all cases, be considered anonymous, and that extraction of personal data through queries or attacks is a realistic risk. The Regional Court of Munich, in *GEMA v. OpenAI* (November 2025), has held that song lyrics were reproduced in model weights in a legally meaningful sense, because simple prompts reproducibly elicited near-identical outputs. This was a finding with direct implications for personal data. The UK High Court's contrary finding in *Getty Images v. Stability AI* (also from 2025) is not applicable: the United Kingdom is not a member of the European Union, and that judgment carries no authority in EU law.

Taking EDPB Opinion 28/2024 and *GEMA v. OpenAI* together, the following chain of reasoning is now grounded in authoritative EU-proximate sources: AI models trained on personal data are presumptively non-anonymous; personal data is retained in weights in a legally meaningful sense; a prompt that reproducibly elicits personal data from a model is therefore a personal data breach. The template must be able to capture this breach type. If you could have a close look, you would see it currently cannot capture any such breach at all.

Comment 1: Field 60 (Nature of the incident)

Field: 60 (Nature of the incident)
Current options (relevant): (c) Hacking / malware activity detected; (f) Unauthorised access to personal data in IT systems; (g) Data exfiltration / unauthorised extraction of personal data; (t) Technical malfunction; (y) Other

Issue:

None of the 25 options in Field 60 captures the extraction or regurgitation of personal data from AI model weights via query or adversarial attack. The closest candidates are, upon inspection, (g) data exfiltration and (y) other. Neither of them is adequate (I mean, other is accurate for every case out there, but that is not particularly meaningful). What is exfiltration? It is an act of extraction from a discrete data store by an external actor. Regurgitation from model weights is structurally very distinct: your data is not in a store; it is absorbed in parameters; it is elicited rather than extracted; and the actor may be an ordinary and simple end user asking an ordinary question rather than a malicious attacker.

The EDPB's own Opinion 28/2024 identifies the relevant attack vectors with precision: membership inference, exfiltration, regurgitation of training data, model inversion, and reconstruction attacks (para. 55). The template's incident taxonomy does not reflect any of these (we cannot tell why). This is a gap that will grow as AI models become a standard processing pipeline for most of our tasks.

GEMA v. OpenAI (Regional Court of Munich, November 2025) provides a concrete worked example. The court found that simple prompts of the form "What are the lyrics of (song title)?" reproducibly elicited near-identical reproductions of songs that were used as training content. The court held that this was sufficient to establish that the content was retained in the model in a legally meaningful sense. Transposed to personal data: a prompt that reproducibly elicits personal data from a model trained on that data, like a named individual's address, bank transactions, medical history, or some identifying details, is, on *GEMA* logic, a personal data breach. The template should, I believe, be able to classify it as such.

Proposed addition to Field 60:

> (z) Personal data extracted or regurgitated from AI model weights, where personal data relating to an identifiable natural person is elicited from an AI model through queries, prompts, or adversarial attacks exploiting data retained in model parameters (including membership inference, model inversion, regurgitation, reconstruction, or exfiltration attacks against model weights or inference interfaces).

This new incident type may be accompanied by a tooltip referencing EDPB Opinion 28/2024, paragraph 55, for the relevant attack taxonomy.

Comment 2: Field 70 (Type of breached data)

Field: 70 (Type of breached data)
Current options (relevant): (a) Basic data; (b) Contact details; (k) Health data; (r) Profile data; (u) Additional data types

Issue:

Field 70 lists data categories by type. Every category presupposes that the breached data exists as a discrete record in a defined data store. This presupposition, again, fails for personal data whose existence is absorbed in an AI model's weights rather than a database table or file.

Personal data absorbed in model weights during training does not correspond to a record that can be identified, isolated, or erased by conventional means. This is thoroughly documented in the academic literature on the right to be forgotten in machine learning systems (see, e.g., Manab, "Eternal Sunshine of the Mechanical Mind", 2024, or Zhao, “The Application of the Right to be Forgotten in the Machine Learning Context: From the Perspective of European Laws”, 2022), and it is for this very reason EDPB Opinion 28/2024 treats the anonymity question as non-trivial rather than, let’s say, resolved by default.

What are the consequences? A controller notifying a breach of this type cannot accurately complete Field 70 using existing categories, because the breach does not involve the exposure of a record of a known type. What has been exposed is personal data whose type may be unknown until the prompt is examined. Its source is a training dataset that may no longer exist in its original form, and its scope cannot be determined by inspecting a database. Without a recognised category, the notification cannot adequately inform the data subject of what has been exposed, which in turn prevents the data subject from exercising their rights under Article 17 GDPR with respect to the exposed data.

Proposed addition to Field 70:

> (v) Personal data absorbed in AI model weights, where personal data relating to an identifiable natural person was used in the training, fine-tuning, or updating of an AI model and has been exposed through query-induced regurgitation, adversarial extraction, or inference attack.

The tooltip for this field should note that the scope and type of data exposed in this category may be uncertain at the time of initial notification, and that this uncertainty does not preclude notification under Article 33(4) GDPR. We may also add a note that this category may overlap with all the other categories listed above and controllers should select all applicable categories.

Comment 3: Field 75 (Measures in place when the breach occurred)

Field: 75 (Relevant measures in place when the personal data breach occurred)
Current options: (a) Pseudonymisation; (b) Backup / Recovery plan; (c) Data encryption; (d) Data protection and information security policies; (e) Data protection and security training; (f) Incident log; (g) Levels of access to data; (h) Logical access control (e.g. MFA); (i) Periodic audits; (j) Physical access control; (k) Up-to-date IT systems; (l) Other measures

Issue:

The measures listed in Field 75 are designed for conventional data processing infrastructure. None of them capture the technical mitigations that are specifically relevant to AI model training and deployment, the infrastructure in which the breach type described in Comments 1 and 2 occurs. This design failure haunts all European data processing approaches, where we are talking spreadsheet languages in an age where we have trillion-parameter AI at our disposal.

EDPB Opinion 28/2024 had identified a specific and materially distinct set of technical mitigations relevant to AI model anonymity and resistance to extraction attacks (paras. 49–55). These included the following approaches: selection criteria for training data sources; data preparation and minimisation techniques; privacy-preserving training methods, including differential privacy; regularisation approaches to reduce overfitting and memorisation; output filtering and guardrails (although they don't use the exact term); and structured testing against membership inference, exfiltration, regurgitation, model inversion, and even reconstruction attacks.

None of these, sadly, appear in Field 75. A controller using conventional pseudonymisation and access control on a model that was trained without differential privacy, without regularisation against memorisation, and without adversarial testing does have some measures in place, but not the measures that matter for the breach type that this template cannot currently classify. The current Field 75 options would allow such a controller to tick several boxes accurately while providing absolutely no clear information relevant to assessing whether this type of breach was preventable.

The fix does not have to replicate the full technical discussion of Opinion 28/2024 in a tooltip. It can be adding a catch-all option that directs AI model controllers to the relevant framework.

Proposed addition to Field 75:

> (m) AI model privacy-preserving technical measures, including but not limited to: differential privacy techniques applied during training; regularisation methods to reduce memorisation; adversarial testing against membership inference, model inversion, regurgitation, and reconstruction attacks; output filtering and guardrails; and data minimisation applied to training datasets. See EDPB Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, paragraphs 49–55, for the relevant framework.

Closing remarks

Our three comments address a single structural gap: the template was designed for a data processing landscape that predates the widespread deployment of AI models trained on personal data. That landscape has considerably changed. The EDPB has already acknowledged, in Opinion 28/2024, that AI models present distinct and non-trivial data protection risks. The Munich court has already held, in *GEMA v. OpenAI*, that training content is retained in model weights in a legally meaningful sense. This template should reflect both.

The proposed additions are minimal. They do not restructure the template in any measure. They add one incident type, one data category, and one measures option, each with a tooltip or cross-reference to existing authoritative guidance. The cost of not adding them will be that a growing class of personal data breach will be systematically misclassified or perhaps left unclassified by every controller in every EU jurisdiction that adopts this template. This is the burden we have to bear.