External research commissioned by the Information Commissioner's Office

Appropriate data protection safeguards for generative AI/foundation model training in the context of the research provisions

December 2025

Authors

Michael Veale, Professor of Technology Law and Policy, Faculty of Laws, UCL

Luc Rocher, Associate Professor, Oxford Internet Institute, University of Oxford

ico. Information Commissioner's Office

This research may not reflect the ICO's views or policies (and we cannot guarantee its accuracy) but is intended to help us understand the nature and range of matters we may need to address and it may inform our decision making. It may also be used to help us create relevant policy, guidance and information to address current issues of concern.

Executive Summary ---- 3
1. Introduction ---- 6
1.1. Technical safeguards ---- 7
1.2. Report structure ---- 8
1.3. Considerations for training datasets that may contain personal data ---- 9
1.3.1. Understanding the identifiability of high-dimensional data ---- 9
1.3.2. Adversary consideration ---- 16
2. Evaluation of current and emerging methods for data minimisation in a scientific research context ---- 18
2.1. General metadata removal ---- 19
2.1.1. Soft bindings ---- 19
2.1.2. Hard bindings ---- 20
2.2. Audio data ---- 22
2.2.1. Speech datasets ---- 23
2.2.2. Disentangling different aspects of a piece of speech ---- 24
2.2.3. Speech redaction and resynthesis ---- 26
2.2.4. Techniques to reduce the identifiability of voice data ---- 28
2.2.5. Text-Based Speech Editing ---- 32
2.3. Text data ---- 35
2.3.1. Personal data in text-based LLM training datasets ---- 35
2.3.2. Methods for detecting and replacing identifiers ---- 42
2.3.3. Stylometry and profiling de-identification techniques ---- 54
2.4. Image datasets ---- 60
2.4.1. Facial data and its data protection risk ---- 61
2.4.2. Soft biometrics and ephemeral identifiers ---- 66
2.4.3. Geolocation inference ---- 70
2.4.4. Scene text ---- 75
2.4.5. Metadata ---- 83
2.4.6. Hash databases (eg for CSAM) ---- 83
2.5. Other sensor data ---- 85
2.5.1. Special features of sensor data ---- 86
2.5.2. Removing direct and indirect identifiers ---- 87
2.5.3. The limitations of classical tabular anonymization ---- 87
2.5.4. Best practices safeguards ---- 88
3. Common themes and tensions ---- 90
3.1. At a generative AI scale, techniques are computationally costly ---- 90
3.2. Considering an ML model's purpose to determine a dataset's utility ---- 91
3.3. Public–private tensions in training data ---- 92
3.4. Lack of relevant benchmarks ---- 93
3.5. Lack of end-to-end pipelines and evaluations of them ---- 95
3.6. Documentation requirements ---- 96

3.7. Prioritisation ---------- 96
3.8. Considering techniques and threat models side by side ---------- 97
4. Suggested approach and trajectory ---------- 97
4.1. Tier 1 — Removing metadata ---------- 97
4.2. Tier 2 — Direct Identifiers ---------- 98
4.3. Interaction with organisational safeguards ---------- 99
Glossary ---------- 101

Executive Summary

In the report below, we consider how technical safeguards can be applied to datasets that are produced and shared by researchers for research purposes under UK data protection law. In particular, for the development of machine-learning tools, these datasets are often large and rich and contain personal data both in directly identifiable form and in a wide variety of indirectly identifiable ways.

We focus on audio, text and image data, with some consideration of other data from sensing systems. We conclude that **there are many technical opportunities based on current research and open-source tools to provide defences for such data**.

In this report, we consider a wide variety of tools and techniques that might serve as safeguards — tools that allow voices to be swapped or edited; tools that allow text in audio to be changed; tools that allow entire transcripts to be resynthesised with convincing emotional melody and cadence; tools that allow text to be redacted or have its style altered and resynthesised; and tools that allow rewriting of text at different levels of generality in order to reduce the privacy risk posed by more specific disclosures. We consider tools that allow faces to be replaced with fictitious or altered faces to protect the identities of people in images while maintaining the information that faces encode about an image, such as gaze or emotion. We consider tools that allow entire bodies to be resynthesised within, for example, pedestrian scenes, to avoid recognition through clothing or gait. We consider tools that add noise to try to protect against methods in which machine-learning systems attempt to identify the location of an image from clues in the photos. We consider tools to erase or alter text present in images, and tools that allow sensor data to be changed and to have more privacy-preserving features based on statistical alteration. In each of these areas, we consider the limitations of these tools, their maturity, and other considerations regarding their utility and feasibility for researchers applying them today.

However, the extremely rich nature of these data means that **these defences can only be regarded as safeguards and will not typically remove the personal data potential from many of these datasets without significantly affecting their utility**. Furthermore, the nature of research is such that "utility" can be hard to define (because research questions are not always predictable). There are significant gaps in the literature on assessing how much applying these safeguards will damage research potential.

Much research is also happening in silos, while the law requires the identifiability of personal data to be considered from many angles at once. As a result, we have **few evaluation methods that combine multiple risks simultaneously and consider how utility affects the ability to defend against these risks**. For example, we have limited information about the effect of face swapping or text editing in photos on the training of downstream models. We have no evidence that these technologies are being used in concert or of how they might interact. Yet such pipelines are going to be crucial for these safeguards to be applied in practice.

Preserving the utility of data for subsequent research when privacy safeguards are applied is often very **computationally intensive,** as many of these tools rely on generative AI systems to create convincing alternatives to the data that has been removed, such as faces, bodies, text in images, or voices. Furthermore, the growth of sophisticated attack vectors against such datasets also involves machine learning (ML) systems, which means that both defending against them and adapting data in response to these attacks has become more computationally intensive than before. Large datasets may find that the most advanced defences designed to preserve utility would cost tens or hundreds of thousands of pounds in computer costs. These are, however, research-grade tools; price is likely to reduce significantly when commercialised, as researchers are often not exploring the question of computational efficiency. Incentive from the regulator to explore this is likely to help foster an industry in privacy-enhancing technologies.

We suggest a **three-tier approach to safeguards** for consideration. The first tier focusses on low-effort and low-cost safeguards that we should always expect to be applied, such as the removal of unnecessary metadata from files, and the scanning for CSAM content using widely available tools. The second tier focusses on the removal of direct identifiers, such as emails or phone numbers from text, or similar text from images (such as a photo of a passport), spoken direct identifiers from audio, and in the case of images, replacing faces of individuals in them. The third tier focusses on indirect identifiers, which uses a wide variety of techniques — many of them more nascent and with more research required — to minimise the possibility of indirect reidentification from inference, reasoning over multiple aspects of data, style or voice, or similar. The impact of this third tier on the utility of data for research may be greater and is generally less studied. Moving up the tiers protects data against more actors, but may create greater utility issues, and as a result, the more extensive technical safeguards may be more appropriate for data intended for wide or public release, while researchers wishing to

apply fewer technical safeguards (such as only those from lower tiers) due to their specific utility effects may need to rely on more organisational safeguards (such as contractual controls or trusted research environments).

1. Introduction

The research, archiving and statistical purposes provisions (hereafter RAS provisions) permit eligible data processing to occur with modified rights and obligations for data subjects and controllers. Analysis of these modifications is not in scope of this report. This report concerns the stipulation in UK law that the RAS provisions cannot be used if appropriate safeguards for the rights and freedoms of the data subject are not in place, which must include among them technical and organisational measures for the purpose of ensuring respect for the principle of data minimisation (such as pseudonymisation).[1]

The focus of this report is the technical **safeguards for training data for generative AI (including foundation models/FMs)**. We want to acknowledge this is a small proportion of the potential use of the RAS provisions. For example, research actors storing interview transcripts or recordings; survey microdata; biological samples — all of these also will be subject to the need for specific and appropriate safeguards where controllers seek to use the RAS provisions in relation to them. While some of these data types overlap with data types suitable for FM training, we do not propose safeguards only on the basis of data types — we have the focus on generative AI (including FM) training in mind.

In this report, we provide a literature review of the current approaches to safeguards that apply to common types of generative AI/FM training data in the context of research. We focus on audio data, text data, image data, and select other sensor data. Within each data type, we characterise the challenges that this data type presents for data protection and look at the types of safeguards that have been proposed, as well as connecting across literatures and filling the gaps where fewer safeguards have been suggested or analysed in scientifically rigorous detail. As we will discuss, our purpose is not anonymity, but data minimisation; ensuring that data is "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed".[2] This means that we have to consider the effect of these safeguards on the underlying purposes — in this case, qualifying research under the RAS provisions.

[1] UK GDPR, art 84C(4) (when/if commenced).
[2] UK GDPR, art 5(1)(c).

1.1. Technical safeguards

In this report, we look at a wide variety of safeguards that can be applied to data intended for generative AI and FM-related research. These safeguards use a variety of terms because they are typically drawn from computational literature.[3] One issue to note is that the methods we discuss often use the term 'anonymisation'. Where that term is used or referred to in the references of this report, it does not refer to the concept of anonymisation under the UK GDPR.

The anonymised information is no longer subject to data protection provisions, as it is no longer be personal data and would fall outside the material scope of the UK GDPR regime. Because it would fall outside the regime's material scope, the research provisions that modify rights and obligations would not apply, making those provisions largely superfluous.

In the report below, when we consider terms like 'anonymisation', we note that they often apply to narrow aspects of the data in question. For example, altering someone's voice in a dataset does not necessarily grant them anonymity in terms of data protection law, as the content of their speech might still reveal who they are, if, for example, they said their name.

We will explicitly highlight where we use terms with reference to specific legal contexts and legal tests, but we ask readers not to draw too many legal conclusions from the specific semantics used below (as opposed to our analysis and recommendations). Terms in law do different work from terms in computer science and need explicit bridging; they are not always the same, nor should they be. Further work will be needed to ensure the bridge between the law in areas we are not considering, such as anonymisation, and the content of these safeguards, is aligned.

The purpose of the safeguards in this report, as an overarching rule, is to minimise the personal data processed for research purposes, subject to not undermining the requirements (the purpose) of that research. These safeguards are not designed to, as anonymisation would, take data outside of the scope of the law. Data protection does not require anonymisation (anonymisation is a question of which data is within scope). However, data protection does require the application of safeguards in this situation; these safeguards are not intended to make the data non-personal, as this may be difficult or impossible given the

[3] Including in (overlapping) areas such as broad machine learning research, computer vision, computational linguistics, signal processing, privacy engineering, document processing, information retrieval, and statistics.

type of datasets generative AI systems and FMs are typically trained upon. As a result, we do not see the possibility of reidentification given these safeguards as fatal to their applicability. Instead, we look at the level of protection they provide, in the context of other trade-offs they may bring.

We also use technical terms — and provide a glossary at the end for readers to support them if they do not have a technical background in this area. Some of these terms may not resonate clearly with the law, but we use them to be faithful to the literature and to help bridge the gap between law and scholarship. In particular, we use the term 'attack', which is extremely commonly used in computer security. Computer security scholars studying privacy do not see the term attack as one loaded with intent, or one that can only refer to an attempt to reidentify an individual. An attack might leak any information that the scholar is trying to analyse the possibility of leaking, or design defences against leaking. For our purposes, as we will discuss through this report, we are not just interested in efforts that immediately reveal the identity of natural persons; inferences that provide further information (such as the inference of where an image was taken, or clustering people wearing the same outfit in a dataset together) can be an intermediate step to revealing the identity of a natural person.

We focus on technical safeguards. The law requires both appropriate technical and organisational safeguards to be applied; considering the necessary organisational safeguards is outside of the scope of this report.

1.2. Report structure

We start this report by looking at some cross cutting issues in data, and setting the scene with definitions and concepts. We include a generic set of considerations about identifiability in the type of data typically used to build FMs and generative AI systems more broadly.

We continue with considering current and emerging methods for personal data minimisation in scientific context (Section 2). In this section, we look at the different methods that can be used as safeguards by grouping the different categories of data — audio, textual, visual, and some generic high-dimensional sparse data (such as mobility data). Each method is accompanied with an appraisal of its feasibility, maturity and potential as a useful safeguard. This section contains some cross-referencing to avoid repetition.

We finish the main considerations with narrative considerations of common themes and tensions (Section 3), where we synthesise cross-cutting limitations and concerns.

1.3. Considerations for training datasets that may contain personal data

We analyse the particular types of datasets that may contain personal data within the meaning of the UK General Data Protection Regulation (UK GDPR) in each section below considering different data modalities. In this section, we introduce a few cross-cutting considerations for training data that may contain personal data. We work on the assumption throughout that the datasets we describe and modalities we consider will, in general, contain personal data. While specific, smaller datasets may exist that do not, FM and ML training often requires large datasets which are collected indiscriminately and can be considered to contain personal data under many approaches. These considerations are useful when looking across these datasets, as they will have implications for any detection methods and their evaluation.

1.3.1. Understanding the identifiability of high-dimensional data

Data that is useful in FM training is often *high-dimensional* in nature, at least in the format that it is fed into ML training processes as. High dimensional data has many attributes for each data point. For normal tabular data, this is quite clear — imagine a spreadsheet with a very large number of columns (representing features), regardless of the number of rows (representing records). This would indicate you have a lot of data on each record. But for less structured data, as is much of the data we consider in this report, this may need further explanation.

We might look at **text** data and think that it is quite simple: isn't "apple" in this sentence just a single data point? It is not, because "apple" in the previous sentence co-occurred with all the words around it — "in", "this", "isn't", as well as the order these words came in and the proximity they have to the word "apple". The sentence actually encodes a very large amount of data, and even more so when you take a paragraph, or a document. Computers do not ingest text for FM training in the linear way humans read it; it has to be transformed into a more tabular format which itself is very big and unwieldy, trying to capture much of this high dimensionality.

These high dimensional datasets mean that records of data can form a vast number of patterns, some of them unique. Consider all the words and the order of these words in a sentence: there is an enormous amount of combinations that *do not* appear. For example, while the word "apple" above co-occurs in its surrounding with the word "sentence", it obviously does not co-occur with all the other words in the dictionary. A table with all the words in the dictionary listed in the X axis and Y axis, and "TRUE" or "FALSE" in each cell wherever they co-occurred, would be massive, and have very few "TRUE" values and an enormous number of "FALSE" values.

Image data is similar — pixel values (eg R, G and B colour channels between 0 and 255) co-occur in some form of an order, surrounded by other pixel values. **Audio** data is similar, because for a computer to analyse it, it is broken down into small 'windows' of time, and a lot of features mathematically extracted from those windows (like the sound's constituent frequencies). Video data is of course a combination of these two and so is even higher in dimensionality than a single modality alone.

In these high dimensional situations, there are huge numbers of potential patterns. This intrinsically heightens any risk of identification of natural persons, as even where an identifier is removed, that record or section may be associated with a unique pattern, which could be connected back to a natural person. These rare patterns themselves become identifiers. While the statistical patterns of words in text or pixels in images are still a conceptual way away from something that might be indirectly identifying of a natural person, this is the backdrop against which these risks and safeguards operate.

1.3.1.1. The existence of identifiers in different modalities

More traditional identifiers also exist directly in the data and can be extracted using methods designed to reason about the data. These can include names, ID numbers, emails, IP addresses, passport numbers, or similar. In some situations, the ICO considers these personal data by default (eg where they are 'online identifiers' within the meaning of the UK GDPR[4]). At other times, a case-by-case test might be applied, such as where an ID number is released, but the corresponding data is not widely

[4] UK GDPR, recital 30 ("Natural persons may be associated with online identifiers provided by their devices, applications, tools and protocols, such as internet protocol addresses, cookie identifiers or other identifiers such as radio frequency identification tags. This may leave traces which, in particular when combined with unique identifiers and other information received by the servers, may be used to create profiles of the natural persons and identify them.")

available, and as such a test as to the reidentification likelihood would need to be applied. The ICO gives the example of DVLA numbers for this.[5]

Explicit identifiers can be present in text data, speech data and visual data — for example, written in text, part of a photograph (of eg a passport), or spoken in audio. The latter two modalities might require technical processing, such as automated speech recognition or optical character recognition, to be detected when they are embedded within media. However, they might also be embedded in the machine-readable metadata of different files. For example, JPEG, HEIC (iPhone image format), TIFF and RAW files can have author names or comments embedded in the EXIF (exchangeable image file format) data; audio data can have author names or revealing original file titles embedded in the ID3 tags; video files can have a large amount of metadata stored in them (location varies based on media type; XMP and MPEG-7 are the most widely used metadata formats).

1.3.1.2. All the data that can 'single out' people

The UK GDPR considers information personal data if it relates to an identified or identifiable natural person. In some cases, it may be less clear if information is personal data, and to determine that, it is important to consider a combination of factors, which might effectively single a person out. The ICO refers to these factors as 'significant criteria'.[6] Academic scholarship often calls these 'quasi-identifiers'.[7] These might include combined references of someone's age, postcode, ethnicity and gender, for example. Singling out might be possible from a single statement that is unique, such as a specific job role in a firm; the certain winner of a 2024 award; author of a certain work; a relation to a separate identifiable person.

Apart from the content of a file which can contain factors enabling singling people out, so can the metadata. For example, a multimedia file's metadata can contain:

---

[5] 'Can We Identify an Individual Indirectly from the Information We Have (Together with Other Available Information)?' (28 March 2025) <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/personal-information-what-is-it/what-is-personal-data/can-we-identify-an-individual-indirectly/> accessed 11 October 2025.
[6] 'Can We Identify an Individual Indirectly from the Information We Have (Together with Other Available Information)?' (28 March 2025) <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/personal-information-what-is-it/what-is-personal-data/can-we-identify-an-individual-indirectly/> accessed 11 October 2025.
[7] The term was first, to our awareness, used in Tore Dalenius, 'Finding a Needle In a Haystack or Identifying Anonymous Census Records' (1986) 2 Journal of Official Statistics 329.

- the GPS coordinates of its creation;
- the date and time of creation; and
- specific information about the (potentially quite unique) device used to record or edit the media.

This metadata is increasing rather than decreasing. For example, the C2PA standard for content provenance is emerging in light of difficult-to-authenticate AI content, and can contain large amounts of identifying information about a file.[8]

1.3.1.3. Inference of further criteria for singling out individuals

LLMs increase the possibility of singling out people by allowing some of these significant factors to be inferred rather than explicitly provided, at scale. This used to be difficult, particularly at scale, but LLMs increasingly make this easier in two ways:

1. Firstly, they increase the ability to connect significant factors from datasets to make predictions that can contribute to identifying people, such as their age, location, occupation or employer.;[9]
2. Secondly, they increase the ability to transform raw data (such as an image) into more structured data that can in turn produce further predictions or inferences, such as by classifying objects, identifying locations, or extracting text.[10]

Consider first the following statement from a Reddit post:

> "There is this nasty intersection on my commute, I always get stuck there waiting for a hook turn."

This can be analysed using abductive reasoning – a logical inference that seeks the simplest and statistically most likely conclusion from a set of evidence. Academics Staab and others provided the above specific example, noting that when prompting an LLM (which does the reasoning in this case), a 'hook turn' is an accidentally disclosing feature, as it is both Australian English and highly specific to Melbourne (the type of turn

[8] C2PA, 'Content Credentials: C2PA Technical Specification' (2025) <https://spec.c2pa.org/specifications/specifications/2.2/specs/C2PA_Specification.html#_hard_bindings>.
[9] We discuss this throughout the report below, such as in the section on Geolocation inference.
[10] We discuss this below for example when considering the ability to transcribe audio, or to analyse and extract text from images. See eg below in Text-Based Speech Editing and Scene text.

is legal barely anywhere else in the world).[11] Other examples provided in that research include the statements "watching Twin Peaks when they were coming home from school" (estimate: a date of birth in the mid-late 80s) and being "charged more now for a 34d" (estimate: a bra-purchaser of female gender).

Forensic phonetics and forensic linguistics more broadly constitute the field of - among other things - inferring information about someone from their speech and writing.[12] Inferences about speakers are not hypothetical techniques, but an extremely widespread part of audio analysis. Commercial AI systems already have speaker recognition built into their pipelines in order to. Popular open-source models used for this purpose include Nvidia's TitaNet[13] and Tencent's WeSpeaker model.[14] Models and benchmark datasets are also growing for niche dialect identification models.[15] In both audio and text, there are a variety of potential signals that disclose aspects about the speaker which might single them out further.[16]

Models can, in general, be applied to datasets in order to transform them into richer sources of information which might then support reidentification. For example, automatic speech recognition (ASR) or optical character recognition (OCR) technologies can turn audiovisual material into text content, which is then vulnerable to the identification methods that apply to text.

But it is not only text that can be extracted from images. Multi-modal models have demonstrated significant ability to geolocate imagery in order to provide more information for singling out the individuals taking them. Contemporary, commercially available multi-modal LLMs have been shown to geolocate images containing environmental scenes (such as the photos that people tend to take day-to-day) to a city level with around

[11] Robin Staab and others, 'Beyond Memorization: Violating Privacy via Inference with Large Language Models', *The Twelfth International Conference on Learning (ICLR 2024)* (2024) <https://arxiv.org/abs/2310.07298> accessed 11 October 2025.
[12] Malcolm Coulthard and Alison Johnson, An Introduction to Forensic Linguistics: Language in Evidence (Routledge 2010).
[13] Nithin Rao Koluguri, Taejin Park and Boris Ginsburg, 'TitaNet: Neural Model for Speaker Representation with 1D Depth-Wise Separable Convolutions and Global Context' (arXiv, 8 October 2021) <http://arxiv.org/abs/2110.04410> accessed 11 October 2025.
[14] Hongji Wang and others, 'Wespeaker: A Research and Production Oriented Speaker Embedding Learning Toolkit' (arXiv, 1 November 2022) <http://arxiv.org/abs/2210.17016> accessed 11 October 2025.
[15] See eg the models and datasets associated with the ACL VarDial series of workshops, the latest held in 2025.
[16] See below in Audio (Section 2.4).

40% accuracy.[17] Shadows further reveal the location of image capture, especially when combined with the time the image was captured, or in combination with multiple images in the same series.[18]

In these situations, models might be considered as other available information that might be combined in order to enrich and potentially identify data. For example, a model that can reason about an image in order to guess its location to the level of a residential address would appear to be providing other available information in order to turn an image into something that might be considered personal data (Section 2.8.3).

There are many examples of this. Image-captioning models (such as the now ubiquitous multi-modal LLMs) can caption a scene, and such annotations inferred from an image (such as what is in it, what the landscape looks like, what is visible) may then be subject to the abductive analysis above to work out where it is. Further information can be gleaned through source camera identification algorithms, which use specific patterns of noise and aberrations (eg manufacturing errors in a sensor or lens) to identify both the camera model and the specific camera which took a photograph.[19] Such models may themselves be revealing (eg specific to a certain market or, in combination with the date of the photograph, may indicate how old a mobile device is and say something about sociodemographic of the data subject as a result).

Audio data can include voice characteristics (eg loudness, pitch, nasality, roughness, hoarseness); speech characteristics (eg tempo, rhythm, pronunciation, accent, articulation, intonation, clarity of expression); non-speech human sounds (eg coughs, breaths, sighs) and background sounds (eg vehicle and natural soundscapes).[20] From all these data points, inferences may be made about a data subject's physical health (eg Parkinson's disease, asthma, common cold, flu); gender; mental health

[17] Oliver Grainge and others, 'Assessing the Geolocation Capabilities, Limitations and Societal Risks of Generative Vision-Language Models', AAAI Fall Symposium 2025 on AI Trustworthiness and Risk Assessment for Challenging Contexts (ATRACC) (arXiv 2025) <http://arxiv.org/abs/2508.19967> accessed 11 October 2025. See generally below in Geolocation inference.

[18] See below Shadows (Section 2.4.3.1).

[19] Jaroslaw Bernacki and Rafal Scherer, 'Algorithms and Methods for Individual Source Camera Identification: A Survey' (2025) 25 Sensors 3027.

[20] Jacob Leon Kröger, Otto Hans-Martin Lutz and Philip Raschke, 'Privacy Implications of Voice and Speech Analysis – Information Disclosure by Inference' in Michael Friedewald and others (eds), Privacy and Identity Management. Data for Better Living: AI and Privacy: 14th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2.2 International Summer School, Windisch, Switzerland, August 19–23, 2019, Revised Selected Papers (Springer International Publishing 2020) <https://doi.org/10.1007/978-3-030-42504-3_16>.

(eg depression, schizophrenia, post-traumatic stress disorder); age; personality traits (eg openness, extroversion, neuroticism); socioeconomic status; communication disorders (eg language disorders, voice disorders, speech disorders); geographical origin; impression on other people (eg perceived as charismatic, competent, emotionally stable); moods and emotions (eg happy, angry, sad, bored, frustrated, anxious); biometric identity; body measures (eg body weight, body height); intention to deceive; and medium-term states (eg sleepiness, intoxication).[21] Research indicates individuals' awareness about these potential disclosures from their voice recordings in the UK is low.[22] With multiple recordings, more can be inferred about a person, for example whether these transient states (moods, environments, intoxication) are indicative of something longer-term and more persistent. 'Personality' or health data, such as mental health information, may be able to be inferred from longitudinal data.

The above methods may be aggravated by the size of the datasets in question. If a large dataset is clustered to recognise the same person across multiple instances in the dataset, attackers can use the criteria that can be inferred from this much larger collection of data (than eg, an individual image) to identify someone. This is called unsupervised person re-identification (Re-ID).[23] These clustering methods can be undertaken without reference to any external dataset, based on common and publicly available tools.[24] We therefore suggest they must be considered as potential "means reasonably likely to be used" to identify a natural person,[25] *regardless* of "whose hands" the data is in in relation to other data.[26] However, the picture is also mixed — under some conditions,

[21] Jacob Leon Kröger, Otto Hans-Martin Lutz and Philip Raschke, 'Privacy Implications of Voice and Speech Analysis – Information Disclosure by Inference' in Michael Friedewald and others (eds), Privacy and Identity Management. Data for Better Living: AI and Privacy: 14th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2.2 International Summer School, Windisch, Switzerland, August 19–23, 2019, Revised Selected Papers (Springer International Publishing 2020) <https://doi.org/10.1007/978-3-030-42504-3_16>.
[22] Jacob Leon Kröger and others, 'Personal Information Inference from Voice Recordings: User Awareness and Privacy Concerns' [2022] Proceedings on Privacy Enhancing Technologies <https://petsymposium.org/popets/2022/popets-2022-0002.php> accessed 11 October 2025.
[23] Unsupervised person Re-identification: A review of recent works - ScienceDirect
[24] See eg the index of many tools for this, with GitHub repositories, here https://github.com/bismex/Awesome-person-re-identification?tab=readme-ov-file#CVPR2024.
[25] UK GDPR, recital 71.
[26] See eg in relation to pedestrians in a 'smart city' context, Zhaojie Sun and others, 'A Comprehensive Review of Pedestrian Re-Identification Based on Deep Learning' (2024) 10 Complex & Intelligent Systems 1733; more generally reviewing recent techniques,

larger data sets mean that people can effectively hide in the crowd due to being less unique.[27] Scale works in multiple ways — with more data over time, reidentification is likely to become easier as people's traces become more unique; with more data over population but not over time, that uniqueness may work in favour of individuals by making them less identifiable. The exact dynamics will depend heavily on the data type. For example, studies show that with more individuals, text authorship attribution becomes much harder, but facial recognition accuracy and browser fingerprint accuracy decrease at a much slower rate.[28]

What the research literature calls 'hard' and 'soft' biometrics might be used for such clustering. We do not use these terms in a legal sense in this report — data protection law does not make this same distinction or use these terms, and individuals should refer to the ICO guidance when considering the legal implications of biometric data.[29] In the language used in the research literature, hard biometric analysis (relatively high uniqueness and permanence) includes face and voice recognition; soft biometric analysis (lesser uniqueness and permanence, lower accuracy in isolation) might include factors such as gait, linguistics, or tattoos. Related to these categories are identifiers with some, limited permanence over time: ephemeral identifiers like clothing. Other proxies, such as an individuals' vehicle, might be factors that can be clustered upon.[30]

1.3.2. Adversary consideration

Many modalities of datasets are heavily used in ML training, and this data has intense potential to both identify people and to disclose sensitive or special category data about them. When considering privacy, data protection, and datasets for research, it is important to consider which kind of *adversary* you might be facing. This often involves considering the

see Andrea Asperti and others, 'A Review of Recent Techniques for Person Re-Identification' (2024) 36 Machine Vision and Applications 25.

[27] Luc Rocher, Julien M Hendrickx and Yves-Alexandre de Montjoye, 'A Scaling Law to Model the Effectiveness of Identification Techniques' (2025) 16 Nature Communications 347.

[28] Luc Rocher, Julien M Hendrickx and Yves-Alexandre de Montjoye, 'A Scaling Law to Model the Effectiveness of Identification Techniques' (2025) 16 Nature Communications 347.

[29] Information Commissioner's Office, 'Biometric Data Guidance: Biometric Recognition' (*ICO*, 13 November 2024) <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/biometric-data-guidance-biometric-recognition/>.

[30] See generally on the use of machine learning in this area, Soumyajit Gayen and others, 'Two Decades of Vehicle Make and Model Recognition – Survey, Challenges and Future Directions' (2024) 36 Journal of King Saud University - Computer and Information Sciences 101885 <https://doi.org/10.1016/j.jksuci.2023.101885>.

capabilities of what the ICO calls a *motivated intruder*.[31] The most important distinction is whether you are facing a human, or a human assisted by computer software — and if so, the capabilities of this software.

A computer-assisted adversary can search through huge amounts of data, and can structure and annotate it in ways that may disclose information about those within it, or to which the data relates. It can at times lead to 'superhuman' (at least compared to the average, non-specialised human) forms of inference, such as analysing a photo to understand where it might have been taken (see Geolocation inference, Section 2.4.3 below) or undertake stylometric analysis of text in order to discern the potential author.

Yet such machines can be vulnerable to defences and safeguards in ways that humans alone are not. For example, adversarial machine learning attacks, which perturb data (such as by adding noise) in order to induce incorrect classifications, are typically designed to be imperceptible (or near enough) to humans, and therefore are supposed have limited impact on what they think of an image or text. This means that you could have a safeguarded image that looks imperceptibly different to a human from an original image, but that for a machine, might be treated entirely differently.

Yet safeguards designed to fool humans cannot use the tricks you can use to only fool machines, and therefore defences designed to protect against unassisted human adversaries are likely to need to manipulate the data a lot more, which might come with significant utility cost for downstream research purposes. However, a human adversary is slow to look through, read or listen to data, and cannot easily manually go through thousands or more images, or lines of text, or minutes of audio, on their own — and in enormous datasets, it might not be the case that it is necessary to extensively safeguard against that kind of adversary.

[31] Information Commissioner's Office, 'Anonymisation: Glossary' (*ICO*, 9 July 2025) <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/anonymisation/glossary/>.

2. Evaluation of current and emerging methods for data minimisation in a scientific research context

In this section, we look through the data types for FM training and provide an overview and analysis of the methods for reducing, and ultimately minimising (with respect to a purpose) the presence of personal data within these datasets. We start with some general considerations. Each section contains information about the limitations of these approaches and a general appraisal, but the final section of this report contains more detailed information about limitations that are cross-cutting.

2.1. General metadata removal

Common to almost all file types is the need to strip and remove metadata which can unnecessarily reveal information about the media they are attached to — and potentially the data subjects they relate to.

The vast majority of metadata is currently **softly bound** to a file (such as EXIF data) - it is easy to remove. A soft binding is attached to a media asset, rather than integrated into the actual content of the media itself. This is equivalent to there being an envelope tucked into the back of a painting's frame — easily discarded without affecting the visual presentation of the painting. In contrast, a **hard binding** is embedded in the content itself, and usually is designed to be tamper resistant or evident. For example, a secret watermark hidden in an image, audio, video or text from which metadata can be computed.[32]

2.1.1. Soft bindings

For images and audio, removal of softly bound metadata is easy to do with open-source tools such as **ExifTool**.[33] For video, there may be additional need to use the open-source software **ffmpeg** to strip more advanced types of metadata.[34] combine some of these approaches are also available, such as the **Metadata Anonymisation Toolkit (MAT2)**.[35] Text documents such as .docx and PDFs are also supported by Exiftool but should be considered on a case-by-case basis as they may contain a variety of revealing metadata (such as comments, track changes, macros).

Some residual risk remains from these tools, which may not always remove the metadata that are applied by one's computer, or added by Cloud services as they upload and download media (and related to that upload or download).

Metadata removal should be the first port of call for any large research dataset. There will very rarely be a reason for supplying data with metadata that might constitute personal data, as this is very rarely used in ML training pipelines.

[32] See on binding types, C2PA, 'Content Credentials: C2PA Technical Specification' (2025) <https://spec.c2pa.org/specifications/specifications/2.2/specs/C2PA_Specification.html#_hard_bindings> s 9.
[33] Available at https://exiftool.org.
[34] Available at https://www.ffmpeg.org.
[35] Available at https://github.com/tpet/mat2.

Stripping soft bindings

Accessibility of methods: Very high.

Considerations for further research: Typically none.

Method maturity: Highly mature.

Key resources: *Exiftool*.[36]

Are these methods feasible in the future? Already extremely feasible.

2.1.2. Hard bindings

Stripping hard-bound metadata is difficult because:

- it typically requires modifying, and sometimes degrading, the quality of the media;[37]
- it might be designed to resist removal, in which case significant degradation is required;[38]
- is not always obvious it is there (for example, it might only be evident with a cryptographic key).[39]

Hard bindings are particularly being pushed due to the requirements of the EU AI Act to watermark synthetically generated content,[40] and the standards being pushed by industry in the C2PA (the Coalition for Content Provenance and Authenticity), an industry body looking to create methods to prove that certain content is not AI generated. We may see a world

[36] Phil Harvey, 'ExifTool' (30 September 2025) <https://exiftool.org/> accessed 15 October 2025.
[37] This is known as the fidelity quality in watermarking schemes, see generally Ingemar J Cox and others, *Digital Watermarking and Steganography* (Morgan Kaufmann 2008) ss 2.3.2., 2.5.2..
[38] See generally on the literature on *defensive watermarking* Andre Kassis and Urs Hengartner, 'UnMarker: A Universal Attack on Defensive Image Watermarking', *2025 IEEE Symposium on Security and Privacy (SP)* (2025) <https://ieeexplore.ieee.org/abstract/document/11023303>.
[39] Ingemar J Cox, Matthew L Miller and Jeffrey A Bloom, *Digital Watermarking* (Morgan Kaufmann 2002) ss 2.3.8., 2.5.9.
[40] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), art 50(2).

where, to avoid deep-fakes polluting what is true, hardware starts to watermark images it produces.[41]

If hard bindings become more popular, researchers may need to seek to remove or mitigate these bindings before publishing datasets for scientific research, as they might disclose the personal data (or support singling out) of the people who created the media.

Stripping hard bindings

Accessibility of methods: Low.

Considerations for further research: Can be destructive of content.

Method maturity: Immature (hard binding methods have also not been consolidated or standardised).

Key resources: Early research libraries and tools such as UnMarker.42

Are these methods feasible in the future? These are not easy to execute.

[41] See early proposals of this in GL Friedman, 'The Trustworthy Digital Camera: Restoring Credibility to the Photographic Image' (1993) 39 IEEE Transactions on Consumer Electronics 905.
[42] Andre Kassis and Urs Hengartner, 'UnMarker: A Universal Attack on Defensive Image Watermarking', 2025 IEEE Symposium on Security and Privacy (SP) (2025) <http://arxiv.org/abs/2405.08363> accessed 15 October 2025.

2.2. Audio data

In this section, we wholly focus on speech data, a type of audio data that we consider both more risky and pervasive than other types of aural personal data.[43] Speech data has always been significant for data protection purposes; for example, it has long been a considered part of UK data protection law that adding audio recordings to a CCTV camera will typically bring it outside of what would be considered proportionate for typical legal CCTV usage.[44]

Speech data brings a range of privacy and data protection risks to natural persons.[45] The content of a recording can reveal a lot about a natural person, such as their age, education level and where they live. It might also reveal information about their social status or employment. Specific things they say may further add to this profile; for example, if the voice data comes from a smart assistant, we might learn what they bought or the messages they wrote to friends, and in doing so be able to extract private attributes and build a profile of that person. We can also learn a great deal from the voice itself: we can mine information such as age and gender, accents that indicate geolocation, and even, in certain situations, information about height and health conditions.

Speech data might also reveal enough information that an individual's voice 'print' or 'template' can be leaked and recreated.[46] This is particularly concerning because voice recognition is used as an authentication mechanism and can be exploited in fraud or scam attacks against data subjects' loved ones. There is a further risk that identifying someone from their voice demonstrates they were in the dataset in question, and in certain contexts, such as medical recordings, this might reveal more about them — for example, if the dataset only contained people with specific diagnoses or who were enrolled in particular medical trials.

---

[43] For example, we might consider health-related audio data (eg breathing patterns or biological measurements like heartbeats); potential identifiers like footsteps as a measurement of gait; the cadence with which someone types, which has been proposed for many decades as an identification and authentication mechanism (see generally M Karnan, M Akila and N Krishnaraj, 'Biometric Personal Authentication Using Keystroke Dynamics: A Review' (2011) 11 Applied Soft Computing 1565).

[44] *Southampton City Council v Information Commissioner* [2013] UKFTT 2012_0171 (GRC).

[45] Jianwei Qian and others, 'Towards Privacy-Preserving Speech Data Publishing', IEEE INFOCOM 2018 - IEEE Conference on Computer Communications (2018) <https://ieeexplore.ieee.org/abstract/document/8486250>.

[46] On voiceprints, see Section 28.

To extract and represent speech data, practitioners do not have access to a single best method. Instead, they rely on a broad range of approaches, carefully choosing which one to use depending on which informational factors in speech are desired and which will be used.[47] For example, sometimes, emotion and cadence will matter — in other times, they will not. Understanding the privacy–utility trade-off, which is necessary for applying robust safeguards to research datasets for, amongst other things, AI training, is therefore not possible without careful consideration of the purpose(s) for which the data is being collected.

2.2.1. Speech datasets

A range of speech datasets for research exist which provide some indicative examples of what we are talking about in this coming section. Some are professionally transcribed, such as the **SPGISpeech** datasets, ~5000 hours of professionally transcribed and checked company earnings calls.[48]

We suspect that some datasets are not likely to create substantial privacy issues relating to the content, and are consensually recorded by individuals who, it should be hoped, are aware of the risks of having their voice in the public domain. An example of this is **Mozilla's Common Voice** dataset, where volunteers record themselves reading short text sentences, which are often sourced from public domain text like Wikipedia; other volunteers then validate these recordings by listening and confirming the audio matches the text via up– or down-voting them. However, Mozilla still attempts to put organisational safeguards on the dataset, requiring downloaders to tick and pledge that "You agree to not attempt to determine the identity of speakers in the Common Voice dataset". Similarly, the **LibriSpeech** dataset was created from the combination of the LibriVox open access audiobook dataset with the relevant texts from Project Gutenberg, using a local sequence alignment algorithm to map the sentences to the spoken text.[49] Meta later used a similar approach to make the **VoxPopuli** dataset by matching European Parliament text to multilingual plenary transcripts.[50]

[47] Jennifer Williams, 'Learning Disentangled Speech Representations' (PhD, University of Edinburgh 2022) <https://era.ed.ac.uk/handle/1842/38724> accessed 12 October 2025.
[48] See kensho/spgispeech · Datasets at Hugging Face.
[49] Vassil Panayotov and others, 'Librispeech: An ASR Corpus Based on Public Domain Audio Books', 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2015) <https://ieeexplore.ieee.org/abstract/document/7178964>.
[50] Changhan Wang and others, 'VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation' (arXiv, 27 July 2021) <http://arxiv.org/abs/2101.00390>.

More controversial are scraped datasets, as they contain information that individuals may not have expected to have been transcribed and distributed to the public or used in FM training.

VoxCeleb and **VoxCeleb2** contain non-consensually scraped 'celebrity' speech (eg from interviews). VoxCeleb2 contains roughly 1 million utterances from 6,000+ speakers. They are benchmarks used in many major research activities (such as research into the performance of voice recognition systems), and generally defended on the basis that these are public figures with low expectations of privacy.

More recent datasets such as **Gigaspeech** and **Gigaspeech 2**,[51] Nvidia's **GRANARY**,[52] and **YODAS**,[53] are controversial in other ways. These methods are effectively developing an automated pipeline to crawl, transcribe (using automated systems), and clean/refine labels (transcripts) for new audio from YouTube videos, podcasts and more. The hope is that even automated labelling will improve language modelling. This hope manifests particularly in relation to small and lower resourced languages (where less spoken content is easily found or available). Similarly, MLCommons' **Unsupervised People's Speech Dataset** is a dataset of 1m hours of speech from the Internet Archive, although this dataset is not labelled by the distributor. These approaches have already been mired in controversy — with public disputes between OpenAI and Google over terms of service violation through mass scraping of YouTube (owned by Google) to build datasets to train OpenAI's *Whisper* ASR system, amongst others.[54] It is unlikely that individuals in these videos realise their data, voices and similar are being used to train models by researchers and commercial firms.

2.2.2. Disentangling different aspects of a piece of speech

Information can be made less identifiable by generalising data and removing information (such as indirect identifiers) in data records that

[51] Yifan Yang and others, 'GigaSpeech 2: An Evolving, Large-Scale and Multi-Domain ASR Corpus for Low-Resource Languages with Automated Crawling, Transcription and Refinement' (arXiv, 27 May 2025) <http://arxiv.org/abs/2406.11546>.
[52] Nithin Rao Koluguri and others, 'Granary: Speech Recognition and Translation Dataset in 25 European Languages' (arXiv, 21 May 2025) <http://arxiv.org/abs/2505.13404> accessed 16 November 2025.
[53] Xinjian Li and others, 'YODAS: Youtube-Oriented Dataset for Audio and Speech' (arXiv, 2 June 2024) <http://arxiv.org/abs/2406.00899>.
[54] Maxwell Zeff, 'OpenAI May Have Broken YouTube Rules by Training ChatGPT on 1 Million Hours of Video' (Quartz, 8 April 2024) <https://qz.com/openai-chatgpt-youtube-videos-gpt4-google-ai-1851394868>.

may identify a specific individual. This may be achieved for speech data by separating voice characteristics from content, a process known as **disentanglement**.

Speech data contains two main signals that relate to personal data: inherent speaker specific traits (eg the voice, intonation, cadence) and speech content (eg what they say, the type of words they use). The ways these signals are entangled together are not fully understood and theorised, and can be difficult to express in a simple and consistent manner.[55]

In speech sciences, it is not generally known to what extent we can disentangle certain signals from each other. Can accent be cleanly disentangled from an individual's identity, from the cadence or pitch of their voice, or the vocabulary they choose to use? It is also not known how many different conceptual types of information are found in the speech signal, and if there are limits to how many of these might be neatly disentangled from each other. For example, if we were successfully going to disentangle 'speaker identity' from their accent, we would have to make the assumption that the speaker identity is only the voice of that individual, and does not include nuances about how particular parts of words are pronounced. Yet in practice, we know that speaker accent is both connected to identity and the way that certain words are pronounced.[56] Disentanglement is a messy and incomplete practice.

For the purposes of privacy, this matters when you try and swap or synthesise a new 'identity' for a speaker to make it harder to identify them from their speech. In practice, what exactly do you swap about a voice in order to keep it useful and realistic sounding, while also masking their identity? To give a deliberately stereotypical example, if you swap an East Midlands English accent for an Estuary English accent out to attempt to obscure an identity, it might be implausible and uncanny to hear the swapped voice greet the listener with "ay up!". Swapping accents while making speech uncanny might be reasonable for a very narrow purpose, eg to record the emotion and cadence of the material being communicated, but not if the purpose of the dataset would be to train speech detection systems which require a variety of accents with plausible utterances in a way that people actually speak.

[55] Tianchi Liu and others, 'Disentangling Voice and Content with Self-Supervision for Speaker Recognition' (2023) 36 Advances in Neural Information Processing Systems 50221.
[56] Jennifer Williams, 'Learning Disentangled Speech Representations' (PhD, University of Edinburgh 2022) <https://era.ed.ac.uk/handle/1842/38724> accessed 12 October 2025.

In computational privacy scholarship, the conceptual nature of disentanglement is somewhat less of a focus. Researchers mounting privacy attacks (as do real-world attackers) use the features that they can extract and get their hands on, regardless of conceptual neatness. If accent, speech, or vernacular are useful, then they can present a threat — attackers are unlikely to conceptually only choose to focus on a single aspect. Furthermore, automated systems that may support these attackers may be even more resourceful in this open-ended way. We see some of these new attacks appearing in the literature.[57] This is only recently beginning to be considered in protocols for vocal privacy, which have in the past often simulated quite low capability attackers.[58]

2.2.3. Speech redaction and resynthesis

We consider that the safest way to share a speech audio dataset with robust privacy is as follows. First, take a speech dataset and transcribe it using automated speech recognition, or ASR technologies. Second, run this dataset through an appropriate suite of text deidentification methods (described later in this report in Section 2.3), bearing in mind their limitations). Finally, synthesise new speech audio entirely from the transcripts using a text-to-speech (TTS) model while removing parts of speech that most identify the speaker and replacing them with a silence or beep.[59] This method should, in theory, produce speech data that is no more revealing than the underlying text data being transcribed.

When the resynthesis approach undermines the quality of speech too heavily, a data provider could instead consider keeping the speech audio dataset intact but simply replacing parts of speech that most identify the speaker with a silence or beep, using the transcript to first identify them. This method will however potentially keep a larger set of identifying signals in the end dataset, such as user accent, intonation, etc.

[57] See in response to Natalia Tomashenko and others, 'The First VoicePrivacy Attacker Challenge', *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2025) <https://ieeexplore.ieee.org/document/10888513>.
[58] Pierre Champion, 'Anonymizing Speech: Evaluating and Designing Speaker Anonymization Techniques' (arXiv, 1 March 2024) <http://arxiv.org/abs/2308.04455> accessed 11 October 2025.
[59] Jing-Xuan Zhang and others, 'Voice Conversion by Cascading Automatic Speech Recognition and Text-to-Speech Synthesis with Prosody Transfer' (arXiv, 3 September 2020) <http://arxiv.org/abs/2009.01475> accessed 12 October 2025.

2.2.3.1. Limitations

Despite their simplicity and promise, both speech redaction and resynthesis discard a lot of potentially useful signals and distort remaining signals, resulting in a degraded data quality. Adding bleeps and redactions will likely contaminate the model training that is, for example, trying to generate naturalistic speech. Removing a word or phrase entirely may make data that is nonsensical. Synthesising new data will remove much additional information, such as much of that relating to emotion, emphasis, cadence, specific voice, or similar.[60] It will not be very useful for training new ASR or TTS systems.

Many firms exist to transcribe and redact audio data using bleeps and similar automatically, integrating typically some of the simple personally identifiable information (which forms part of the data that data protection oversees) detection tools we cover in this report.[61]

No firms we are aware of sell a full resynthesis pipeline including privacy-preserving text redaction, but this is trivial to create once you are holding a transcript diarised by speaker using either open source tools (ASR using eg Whisper; TTS using eg Chatterbox, Coqui TTS) or any number of proprietary APIs (from all major cloud providers). Resynthesis pipelines without the text redaction components exist designed to replicate aspects of speech such as emotion, melody and intonation.[62] Depending on how much is reintroduced, this might reintroduce re-identifiability risk, for example if many other identified pieces of audio of the same person exists, enough information may be present in the emotion, melody and intonation for reclassification. This may particularly be an issue if the data is, for example, spoken data of a well-known person, with much public information and recordings about them, whose identity the researchers were trying to hide.

Nevertheless, these methods should not be discarded as a possibility. Entities holding audio datasets for scientific research should genuinely ask whether a transcript, or a read-out transcript, is good enough for their purposes — and it may well be. However, for many purposes that relate to AI training, this pipeline of techniques will come with a significant loss of utility.

[60] That being said, some could be transferred via a *prosody encoder*, which tries to annotate the transcript with those features, a little like stage directions in a play.
[61] For example see: https://caseguard.com/audio-redaction-software/.
[62] See eg Henry Li Xinyuan and others, 'HLTCOE JHU Submission to the Voice Privacy Challenge 2024' (arXiv, 17 September 2024) <http://arxiv.org/abs/2409.08913> accessed 13 October 2025.

Speech redaction and resynthesis

Accessibility of methods: High (well understood components to string together).

Considerations for further research: May undermine the data's utility for model training and performance.

Method maturity: Components have high maturity; pipeline maturity is low in the sense that it does not exist today for privacy purposes, but it is conceptually very simple to put the components together.

Key resources: Hugging Face leaderboards for ASR and TTS.63

Are these methods feasible in the future? Already feasible; high certainty that these will become better. The main issue is which text redaction technique is appropriate (see Section 2.4.2).

2.2.4. Techniques to reduce the identifiability of voice data

Voice anonymisation (the term the literature uses, rather than relating to DP law) links the identity of a data subject with their voice and attempts to obscure it. There are two main methods for doing this.

Signal-level modification methods process the data frame-by-frame and modify the phase of certain frequencies, while leaving others unaltered.[64] **Deep-learning synthesis methods** try to use ML to disentangle a speaker's identity from other acoustic features of the audio, and to only alter identity relevant features.[65] The latter methods have typically produced higher quality audio in terms of listener ratings, although that comes at the cost of simplicity and computational ease. We focus on the last set of methods, as they are the only ones showing promise in defending against attacks that can be similarly carried out by powerful neural networks.

In the context of deep learning methods, a typical approach disentangles the main speech signal into three components:

[63] https://huggingface.co/spaces/hf-audio/open_asr_leaderboard; https://huggingface.co/spaces/TTS-AGI/TTS-Arena-V2.
[64] See eg Jose Patino and others, 'Speaker Anonymisation Using the McAdams Coefficient' (2021) <https://www.isca-archive.org/interspeech_2021/patino21_interspeech.html> accessed 12 October 2025.
[65] See generally Soroosh Tayebi Arasteh and others, 'Addressing Challenges in Speaker Anonymization to Maintain Utility While Ensuring Privacy of Pathological Speech' (2024) 4 Communications Medicine 182.

1. **Speaker identity**, as typically represented by a fixed-dimensional vector (embedding) extracted from the intermediate layer activations of a deep neural network trained for speaker verification. This is what is often thought of as the 'voice' or voice 'print' or 'template'.[66] The data in this embedding represents many, but not all, of the ways that the voice signal can vary amongst people.
2. **Linguistic features**, which are effectively the spoken content — which phonemes the speaker has uttered.
3. **Prosodic features**, which can be thought of as the melody or the intonation of the speech.

The state of the art in voice anonymisation is usually assessed through the so-called international VoicePrivacy challenge, an academic initiative which runs alongside major computing conferences in the area.[67] This challenge asks researchers to produce methods for speaker level anonymisation that do not distort linguistic and emotional (prosodic) content, but do distort the speaker identity.

2.2.4.1. Limitations of these approaches

A common issue with studies on voice anonymisation is that their protective rates are based on analysis that assumes under-capable attackers, even with the tools available at the time, and certainly with the tools that are becoming available today.[68] Some of this is being amended, but only in recent years — the Voice Attacker Challenge 2025 is the first time that a companion challenge to the Voice Privacy Challenge (to execute the best reidentification attack) has been run.[69]

This is particularly important when we consider other factors which might disclose or link speakers, such as their use of words, pauses, accent, dialect and so on, both associated with the linguistic and prosodic

[66] David Snyder and others, 'X-Vectors: Robust DNN Embeddings for Speaker Recognition', 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2018) <https://ieeexplore.ieee.org/abstract/document/8461375> accessed 12 October 2025.
[67] See https://www.voiceprivacychallenge.org; see for the functioning of the 2024 version, Natalia Tomashenko and others, 'The VoicePrivacy 2024 Challenge Evaluation Plan' (arXiv, 12 June 2024) <http://arxiv.org/abs/2404.02677> accessed 11 October 2025.
[68] Pierre Champion, 'Anonymizing Speech: Evaluating and Designing Speaker Anonymization Techniques' (arXiv, 1 March 2024) <http://arxiv.org/abs/2308.04455> accessed 11 October 2025.
[69] Natalia Tomashenko and others, 'The First VoicePrivacy Attacker Challenge', ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2025) <https://ieeexplore.ieee.org/document/10888513> accessed 12 October 2025.

components of speech, which are more difficult to convincingly alter. This aspect of speech is sometimes called paralinguistics.

This is not some flaw that we have discovered in the VoicePrivacy challenge that they are unaware of. Indeed, the VoicePrivacy Challenge attempts to change or obscure the voice whilst explicitly "leaving the linguistic content and paralinguistic attributes intact". For example, the defences in this Challenge may defend against people using a dataset to spoof a bank authentication using a passphrase (eg saying 'my voice is my password' as some companies have procured systems to do).[70] But in terms of ways in which individuals can be reidentified within a large corpus of audio data, it leaves a lot of potential gaps — they might be reidentified using their content or cadence.

It has long been known that both prosodic (melody, cadence, etc) and linguistic (content) components of a speech signal can be used to identify speakers.[71] Prosodic patterns are increasingly used by scholars as a mechanism to attack voice anonymisation systems.[72] Insofar as both of these are left intact, they may be used to identify individuals. This would particularly be the case for longer form spoken passages, where linguistic patterns that are also found in other data from the same individual may become apparent.

The suitability of voice anonymisation methods depends heavily on the research purpose. Many methods seek to replace voices with a single target voice. However, there will be situations where voice diversity is important (for example, to train models on a wide variety of types of speaking in order to produce a model that can transcribe them well). In other situations, it may be important to keep contextual factors (eg age or sociodemographic atrributes).[73] The purpose or utility will determine which ones to keep and which ones to drop.

---

[70] See eg the weakness of such systems against AI voice attacks demonstrated in Joseph Cox, 'How I Broke Into a Bank Account With an AI-Generated Voice' (*VICE*, 23 February 2023) <https://www.vice.com/en/article/how-i-broke-into-a-bank-account-with-an-ai-generated-voice/> accessed 12 October 2025.

[71] Frederick Weber and others, 'Using Prosodic and Lexical Information for Speaker Identification', 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (2002) <https://ieeexplore.ieee.org/abstract/document/5743674> accessed 12 October 2025.

[72] See eg Thomas Thebaud and others, 'PPX-Anon: Prosody, Pitch and X-Vectors for De-Anonymization; Our Submission to the Voice Attacker Challenge 2024', *5th Symposium on Security and Privacy in Speech Communication* (ISCA 2025) <https://www.isca-archive.org/spsc_2025/thebaud25_spsc.html> accessed 12 October 2025.

[73] Sarina Meyer and Ngoc Thang Vu, 'Use Cases for Voice Anonymization' (arXiv, 8 August 2025) <http://arxiv.org/abs/2508.06356> accessed 12 October 2025.

One thing that stands out from the very nascent literature on how good a model trained on anonymised speaker data performs, is that the benchmarks used by competitions such as *VoicePrivacy Challenge* are poor predictors of downstream model performance when trained on anonymised data. The reason for this disconnect is that word error rate (WER) — a common metric of a dataset — only measures phonetic intelligibility (how many words were correctly identified, how many were not), whereas a text-to-speech model (the type of model being considered in the only study that can be located on model-making using such data) being trained requires much more for people to think it performs well: naturalness, consistency, and, crucially for a multi-speaker text-to-speech model, a diverse and distinctive set of pseudo-speaker voices (which may or may not relate to real people's voices).[74] This is the difference between an intelligible but unpleasant 'robotic' voice and a naturalistic voice that is easy to follow and listen to. There is however some preliminary evidence that automated speech recognition systems can be well trained on this deidentified data.[75] To assess utility as it stands, the VoicePrivacy challenge does not train new models, but assesses the quality of the speaker anonymisation system by applying a well-understood model to the speaker-anonymised data and comparing its results to the correct speech.[76]

Voice identifiability reduction techniques

Accessibility of methods: High (well documented, much open code).

Considerations for further research: May be inappropriate where downstream use requires diversity of speakers (note evidence exists for this for text-to-speech models but not speech recognition models). High potential for times where the signals of interest are the style of speaking, emotional response, sentence cadence, and so on.

Method maturity: High (open-source, but limited commercialisation).

Key resources: VoicePrivacy Challenge.

[74] Wen-Chin Huang, Yi-Chiao Wu and Tomoki Toda, 'Multi-Speaker Text-to-Speech Training With Speaker Anonymized Data' (2024) 31 IEEE Signal Processing Letters 2995.
[75] Brij Mohan Lal Srivastava and others, 'Privacy and Utility of X-Vector Based Speaker Anonymization' (2022) 30 IEEE/ACM Transactions on Audio, Speech, and Language Processing 2383; Mohamed Maouche and others, 'Enhancing Speech Privacy with Slicing' (2022) <https://www.isca-archive.org/interspeech_2022/maouche22_interspeech.html> accessed 17 October 2025.
[76] Natalia Tomashenko and others, 'The VoicePrivacy 2024 Challenge Evaluation Plan' (arXiv, 12 June 2024) <http://arxiv.org/abs/2404.02677> accessed 11 October 2025.

Are these methods feasible in the future? Yes, authors believe so with high confidence.

2.2.5. Text-Based Speech Editing

Where linguistic content is the core disclosure risk, then there are a few steps that can be taken. Simple redaction (eg beeping out) is possible, but this method is quite destructive and introduces a likely unwanted artefact. Instead, the state-of-the-art approach here is *text-based speech editing* (TBSE). Compared to speech synthesis (see the ASR-TTS cascade described above), the speech editing task requires a model to alter words or phrases within an utterance to match a target transcript while leaving regions of the original speech not targeted for editing unchanged.[77]

The most recent research and codebases in this area use so-called *token infilling*. The case of the open-source VoiceCraft model and codebase is the key example of this technology. Using VoiceCraft, a user can select tokens (eg. words or parts of them) to replace; these tokens are deleted from the audio, and a generative model replaces them with the desired text, using the surrounding audio for context for the 'infilling'.[78] The nature of the system is such that it is designed to not require fine-tuning for new voices, but to learn and understand them from the context.

Such tools are already commercialised, which heightens confidence in their viability as a regulatory safeguard in the near-term. Podcasting platform Descript sells both a text-based editing tool that allows individuals to edit audio as if it were text (at least for redaction), as well as an 'audio repair' tool that synthesises new audio based on text, such as to make sentences flow more naturalistically.[79]

[77] Puyuan Peng and others, 'VoiceCraft: Zero-Shot Speech Editing and Text-to-Speech in the Wild' in Lun-Wei Ku, Andre Martins and Vivek Srikumar (eds), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics 2024) <https://aclanthology.org/2024.acl-long.673/> accessed 13 October 2025.
[78] Puyuan Peng and others, 'VoiceCraft: Zero-Shot Speech Editing and Text-to-Speech in the Wild' in Lun-Wei Ku, Andre Martins and Vivek Srikumar (eds), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics 2024) <https://aclanthology.org/2024.acl-long.673/> accessed 13 October 2025.
[79] Descript, Inc., 'AI-Powered Audio Repair' (2025) <https://www.descript.com/regenerate> accessed 13 October 2025.

2.2.5.1. Limitations of these approaches

Such techniques do face inherent limitations, although they may be resolved in the coming years due to the fast pace of development in this field. Recordings where people talk over each other — the 'cocktail party problem' — are hard to both transcribe and synthesise.[80] VoiceCraft reports errors, including adding silences and noise instead of the intended edits, and functioning better on shorter edits than longer or more complex ones. It seems inherently likely though that these methods will rapidly improve, not least because of the desire to use such technology in media editing, such as podcasts and audiobooks.

Furthermore, no research that we have discovered explicitly tries to link text-based speech editing techniques together with privacy-preserving text transformations. While in principle, edits that preserve privacy should not present a uniquely different engineering challenge for text-based speech editing (we consider these edits in Text, Section 2.4), given that edits-are-edits, there are optimisation considerations. Privacy-preserving text transformations may not, for example, be calibrated to make the type of small but effective edits that might be edited into speech well given the constraints of existing models. For example, a text editing system might make a large change to text, but a speech editing model might only be good at convincingly changing a word here or there, without sounding artificial. As a result, very little can be said about utility trade-offs from these systems.

Text-based speech editing

Accessibility of methods: Medium (open code for different stages, but no end-to-end workflow yet exists motivated by privacy considerations).

Considerations for further research: Large edits may not be convincing; clean recordings with a single speaker at a time in a calm environment will work better than other systems; evaluation metrics are missing.

Method maturity: Low (open-source, but limited commercialisation in relation to in-painting of words and phrases, and no privacy flow).

[80] Xinlu He and Jacob Whitehill, 'Survey of End-to-End Multi-Speaker Automatic Speech Recognition for Monaural Audio' (arXiv, 16 May 2025) <http://arxiv.org/abs/2505.10975> accessed 13 October 2025.

Key resources: VoiceCraft repository.[81]

Are these methods feasible in the future? Yes, the authors of this report believe so with high confidence (but the inherent limiter will be the ability to transform text to a privacy preserving form).

[81] Puyuan Peng and others, 'VoiceCraft: Zero-Shot Speech Editing and Text-to-Speech in the Wild' in Lun-Wei Ku, Andre Martins and Vivek Srikumar (eds), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Association for Computational Linguistics 2024) <https://aclanthology.org/2024.acl-long.673/>; Source code: https://github.com/jasonppy/VoiceCraft.

2.3. Text data

Text data is an extremely important data modality; text-based FMs are important parts of many other FM workflows. Text also presents some difficult conceptual challenges in reasoning about data protection and appropriate safeguards.

2.3.1. Personal data in text-based LLM training datasets

Text data used to train LLMs mostly fall in the six broad categories below.

2.3.1.1. Web crawling

Web crawling involves using publicly-accessible web data. It is the foundation of all large-scale LLMs internet corpora, including for example the widely used datasets C4 (Colossal Clean Crawled Corpus)[82] and The Pile.[83] Many are based on the Common Crawl, which every month produces about 300TB of scraped website content.[84] Their massive size, stylistic diversity, and the broad spectrum of written human communication types they capture are key for LLMs to generalise beyond narrow task-specific training. Typically, for LLM training these datasets are heavily pre-processed from machine-readable web formats such as HTML into deduplicated, quality-filtered, free text archives.

Web crawled datasets inherently contain vast amounts of personal data,[85] and signals that can identify or enable re-identification of people: direct identifiers (names, phone numbers, email addresses, physical addresses, usernames), identifiers (rare combinations of facts), inferred/predicted attributes (political opinions, religious beliefs, sexual orientation, racial or ethnic origin), and behavioural and biometrics traces (genetics, location/mobility patterns). Even without direct identifiers such as names or phone numbers, these vast collections combining together platform

[82] Colin Raffel and others, 'Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer' (arXiv, 19 September 2023) <http://arxiv.org/abs/1910.10683>.
[83] See The Pile.
[84] Alexandra Luccioni and Joseph Viviano, 'What's in the Box? An Analysis of Undesirable Content in the Common Crawl Corpus' in Chengqing Zong and others (eds), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (Association for Computational Linguistics 2021) <https://aclanthology.org/2021.acl-short.24>.
[85] Rachel Hong and others, 'A Common Pool of Privacy Problems: Legal and Technical Lessons from a Large-Scale Web-Scraped Machine Learning Dataset' (arXiv, 20 June 2025) <http://arxiv.org/abs/2506.17185>.

data and very small forum data[86] can enable adversaries to link someone's data across sources by searching for unique usernames[87], with potentially more statistical accuracy than public search engines (also subject to 'right to be forgotten' filtering that such datasets will not have received requests for).

Web crawled datasets

Presence of personal data within the dataset category: Many studies have confirmed that web crawled datasets contain personal data.

Detection of personal data: Comprehensive automatic detection unfeasible when considering the scope of the personal data definition. Many heuristics exist for simple direct identifiers and indirect identifiers, using rule-based and LLM-based fuzzy methods for entity recognition.[88] Unfeasible for all personal data due to the messiness of web crawling.

Feasibility of removal and implications for the utility of the dataset: Serious challenges given the heterogeneity of personal data types in these datasets — some describing news and current affairs or history, while some being scraped from personal or social media sites. Feasibility would depend on how much the downstream use required models to themselves reproduce tasks with factual accuracy.

2.3.1.2. Curated reference text datasets

Curated reference text datasets include reference materials such as Wikipedia, Wikisource, Project Gutenberg, Internet Archive books, OpenStax textbooks, technical standards (ISO, IEEE, RFCs), legal documents (EUR-Lex, ECHR), and official specifications. These documents contain authoritative knowledge of factual reliability, with good coherence, grammar, and logic that help models achieve reasoning capabilities. They often contain accurate citation provenance and references, key for training models to support information retrieval. Many of these datasets are also translated by expert translators into dozens of

[86] Jesse Dodge and others, 'Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus' in Marie-Francine Moens and others (eds), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics 2021) <https://aclanthology.org/2021.emnlp-main.98>.
[87] Daniele Perito and others, 'How Unique and Traceable Are Usernames?' (arXiv, 8 March 2011) <http://arxiv.org/abs/1101.5578>.
[88] See eg scrubadub.

languages, allowing developers to use them for training multilingual models.

Curated sources tend to be more structured and authoritative than web crawled ones and, as such, are generally considered to contain fewer sensitive or special category data. They still include personal data, notably biographies of living public figures and biographical material embedded in books. Encyclopaedias and books routinely include biographies and family details; such content is commonly included in curated corpora used for LLM training. This data has always been challenging for data protection, which has no explicit exemption in the law that could take it out of scope, yet courts have clearly been wary of extending obligations (such as erasure) to public figures in the same way.[89]

Mistakes can nevertheless be made when creating these datasets. For instance, Wikimedia[90] produces public dumps (downloadable, rather than scraped, data) of the Wikimedia Foundation projects, which includes Wikipedia. Different types of public dumps exist. Some include discussions and user pages. Not being careful in separating out the encyclopaedic content from the editor content inadvertently result in someone including pages with editors' usernames or written communications — only 11% of Wikipedia consists of article pages. Finally, personal data can be included in downloaded curated sources for LLM training, even after their initial removal from sites like Wikipedia. Content removed from Wikipedia will not appear in new dumps once deleted, but previously downloaded or locally-stored dumps (or third-party copies) can retain any content removed from the internet. Wikipedia may include biographies of people who may not be public figures or notable individuals (eg vandalised pages, or pages added as a joke by friends, or pages of non-notable people added in good faith), whose processing may be more difficult to justify rather than a celebrity.

Reference text datasets

Presence of personal data within the dataset category: Yes. Curated reference texts typically include personal data, but often data of notable or public figures. Difficult to reason about in the law as the UK data protection law does not provide a way to neatly or computationally distinguish between personal data of public figures and private individuals, both of which might engage privacy rights depending on the

[89] See eg *NT1 and NT2 v Google* [2018] EWHC 799.
[90] Wikimedia Commons.

content and context.

Detection of personal data: This is challenging. Curated sources are cleaner and more structured, so detection of clear, direct identifiers is easier in principle, but not feasible for all personal data.

Feasibility of removal and implications for the utility of the dataset: Challenging. Better processing of curated Wikipedia datasets, eg to include Wikipedia articles while discarding discussion pages can limit the likelihood of including personal data that is not necessary to the processing. Removing *all* personal data, eg all biographies, would be challenging to do automatically, and likely unnecessary as an appropriate safeguard, insofar as it would require many extremely public facts about the world ('who is the Prime Minister of the United Kingdom') to be removed.

2.3.1.3. Source code datasets

Source code datasets include public source code repositories (Github, GitLab, Bitbucket), often filtered to only include code with permissive free software licenses. Commonly used curated collections such as The Stack, StarCoderData, and CodeParrot contain source files across many languages, README and documentation, example notebooks, commit metadata, and (depending on the dataset) issues and commit histories all key for training LLMs with code generation and programming capabilities.

Source code repositories frequently contain personal data. Public repositories collect author metadata (usernames, commit author, committer names and email addresses). Many datasets built from GitHub dumps have been shown to contain exposed secrets and credentials (API keys, tokens, SSH keys), IPs and other structured identifiers.[91] Test data, or special category data made public by mistake, can contain real user data with names, emails, addresses, phone numbers, medical or financial data, etc. depending on the use case.

Source code datasets

Presence of personal data within the dataset category: Yes. Large source code datasets are known to include names, emails, usernames, passwords, IP addresses as well as machine identifiers such as API and

[91] Loubna Ben Allal and others, 'SantaCoder: Don't Reach for the Stars!' (arXiv, 24 February 2023) <http://arxiv.org/abs/2301.03988>.

SSH keys.

Detection of personal data: This is challenging.[92] Direct identifiers with a standard format such as email addresses are easy to identify, and developers can easily detect files that match them.[93] Most studies on such datasets focus on these narrow types of personal data (only direct identifiers) ; therefore statistics on the amount of personally identifiable information ("PII", a term commonly used in the US), in source code datasets should generally not be interpreted as providing the full picture when it comes to the UK GDPR data protection.

Feasibility of removal and implications for the utility of the dataset: The feasibility in this context is challenging. Techniques can and sometimes do remove or mask clearly defined and located secret information and identifiers (emails, API keys, tokens), but removing all personal data is technically and conceptually difficult. Aggressive removal of strings matching direct identifier patterns can preserve privacy but also strip valuable data which is technically crucial for the functioning of the code. More sophisticated redaction, taking into greater account the context of each potential identifier, is currently computationally expensive and error-prone.

2.3.1.4. OCR and ASR Data

Optical Character Recognition (OCR) and Automatic Speech Recognition (ASR) data include transcripts of meetings (eg Zoom, Teams), podcasts, lectures, voicemail, scanned books, historical documents, handwritten notes, receipts and forms. Public datasets commonly used in research and model training include Mozilla Common Voice (crowdsourced speech), LibriSpeech (audiobooks), and a variety of OCR collections and document image corpora. OCR and ASR datasets provide information that is not easily available through web crawling, such as spoken language (conversational register, prosody, dialogue), meeting and turn-taking dialogue, and linguistic diversity across historical periods and regions. When paired with human-generated ground truth transcripts, these are essential for training robust multimodal and ASR/OCR-capable foundation models.

[92] "The PII pipeline for this dataset is still a work in progress", https://huggingface.co/datasets/bigcode/the-stack. https://github.com/bigcode-project/bigcode-dataset/issues/1 "there are many false positives which we should be careful with to not damage the data"
[93] https://github.com/bigcode-project/bigcode-dataset/blob/main/pii/README.md

Many of these datasets frequently contain personal and special category data, including speaker names or relevant metadata (who attended a meeting), biometric voice data, health and financial information revealed during calls or in scanned records, legal testimony in OCR-ed court documents, and private correspondence or diaries in digitised historical collections.[94] Telehealth recordings are an especially salient example: audio/video and transcripts of clinical consultations contain health data — special category data — which makes its inadvertent inclusion in training corpora particularly problematic. In general, individuals may not expect to have data they provided in one context (a public panel recording, or even a question asked by them in the audience, where they gave their name) available in a searchable, machine-readable form.[95]

OCR and ASR datasets

Presence of personal data within the training dataset category: This is well-evidenced.

Detection of personal data: This is practically challenging. Certain direct identifiers can be detected with state-of-the-art tools. This includes data such as phone numbers, credit card numbers, email addresses, etc. with regular expression matching and other named entity recognition heuristics. Less obvious identifiers from biometrics, or indirect identifiers embedded in speech and OCR-ed images are much harder to detect.

Feasibility of removal and implications for the utility of the dataset: Removal is challenging. Targeted redaction/masking is prevalent in the industry using named entity recognition heuristics. Removing all personal or special category data (including contextual descriptions, speaker/authorship identity, and semantic references to private matters such as an allusion to a medical procedure) is much harder.

2.3.1.5. Structured and semi-structured datasets

Structured and semi-structured datasets include a broad range of input training data, such as structured and semi-structured data from Wikidata or DBpedia entries, knowledge graph RDF triples, product catalogues,

[94] Casandra Rusti and others, 'Benchmark Dataset Dynamics, Bias and Privacy Challenges in Voice Biometrics Research' (arXiv, 18 August 2023) <http://arxiv.org/abs/2304.03858>.
[95] See generally Helen Nissenbaum, 'A Contextual Approach to Privacy Online' (2011) 140 Daedalus 32.

internal databases serialised as natural language, and structured markup data like JSON files. This data generally describes the characteristics of things, and relations between them, in ways that machines can understand. It powers many virtual assistants — if you ask your telephone how old the Archbishop of Canterbury is, it will usually not be asking an LLM to infer it, it will query it via a structured database.

This category also includes behavioural and trajectory data such as health trajectory, mobility patterns, and credit card shopping spends. This data is often used to train specialised LLMs on internal organisational data, as well as to enhance training of LLMs on knowledge-intensive tasks such as question-answering, factual reasoning, or named-entity recognition and disambiguation.

Public knowledge datasets share the same risks as encyclopaedias and reference databases (see previous sections). Internal/commercial datasets (eg the employment or healthcare records, recommendation system patterns) can be contain the higher proportions of special category data.

2.3.1.6. Instruction-response pairs and human-preference rankings

Instruction-response pairs and human-preference rankings involve curated datasets pairing natural language instructions with expected responses or preferences, including public datasets such as FLAN,[96] OpenAssistant Conversations,[97] and ShareGPT,[98] as well as domain-specific instruction pairs/rankings. They can include either human-authored examples, model-generated examples with human validation, or entirely synthetic (fabricated) pairs of questions and answers. Comparison datasets where human annotators rank or rate multiple model outputs can contain even richer data by typically including user input, alternative responses, and user preferences. Datasets include Anthropic's Helpful and Harmless (HH-RLHF), OpenAssistant rankings, and reward data from InstructGPT.[99]

Instruction-response pairs and human preference data are key mechanisms for 'aligning' language models during training, to follow user intent and behave as helpful assistants in theory, better aligned to human

[96] See Introducing FLAN: More generalizable Language Models with Instruction Fine-Tunin
[97] See [2304.07327] OpenAssistant Conversations -- Democratizing Large Language Model Alignment
[98] See ShareGPT: Share your ChatGPT conversations - Chrome Web Store
[99] See InstructGPT

values, communication norms, and safety/security requirements. This can help train models with adaptation to varied phrasing of the same request, better response formats, and adoption of personas customised to fit the user. Domain-specific instruction sets help train models in specialised areas such as healthcare or law.

Instruction-response pairs and human-preference rankings frequently contain personal data. Datasets can naturally contain user messages, such as in ShareGPT-style datasets that include real human queries shared with models (with users often disclosing medical symptoms and health conditions, legal problems, and relationship or financial issues).

2.3.1.7. Other datasets

Finally, training LLMs with text data can involve other, narrower task-specific datasets which partially overlap with the last five categories above. Such datasets could include labelled examples for NLP tasks including summarisation (CNN/DailyMail, XSum, MultiNews), information extraction (CoNLL NER, OntoNotes, DocRED), text classification (SST, AG News, DBpedia), question-answering (SQuAD, Natural Questions, TriviaQA), dialogue (MultiWOZ, SGD), machine translation, and tool calling.

Task-specific datasets can naturally contain personal data within the source materials. For instance, summarisation data from news articles include names, biographies, locations, along with potentially sensitive information about crimes, accidents, and personal matters; summarisation data can also include correspondence, transcripts, and reports with employee information, financial, or health data, etc.

2.3.2. Methods for detecting and replacing identifiers

Much of the scientific literature anchors detection/replacement to the US HIPAA rules in healthcare, with limited attention to non-US and non-healthcare tasks. Most systems treat direct identifier detection and replacement as a two-stage process: (a) detect tokens, words, or sentences that contain identifiers then (b) transform them by replacing them with blunt masking (“[NAME]”) or by using more utility-preserving methods that ensure the text remains fluent (eg grammar, gender, dates stay consistent) and de-identification becomes harder to spot. This latter

approach, called Hiding in Plain Sight (HIPS),[100] is generally considered best practice in commercial and institutional systems.

Direct identifier detection has evolved from simple heuristics to advanced LLM-based approaches over the years, with three main approaches: rule-based, ML based, and hybrid methods. We also consider a newer paradigm, multi-agent LLM methods.

2.3.2.1. Rule-based approaches

Rule-based approaches rely on regular expressions (regexes), dictionaries, checksums, and format validators to identify well-formed identifiers. These methods are deterministic and auditable. Hand-crafted patterns and checks, like the Luhn algorithm for credit cards[101] or date format validators, are predictable, with explainable behaviours that help testing for compliance. With stable and well-defined formats, these rules can achieve high precisionstructures.[102]

Rule-based systems are also often implemented too narrowly. For instance, email syntax is defined by a number of complex RFC memoranda.[103] Standards-compliant parsers follow these specifications but many 'PII' detector regexes seen in the wild re-implement simplified patterns that will incorrectly reject valid addresses. For instance, plus-addressing (user+tag@example.com), internationalised addresses using UTF-8 (用户@例え.jp), addressing with an apostrophe (user.o'email@example.com) and nested subdomains (user@sub.domain.example.com) are all valid specifications — but detectors will often miss some or all of these.

Rule-based systems have high precision on clean, high-quality data but low recall elsewhere: they are susceptible to noise and small differences

[100] David Carrell and others, 'Hiding in Plain Sight: Use of Realistic Surrogates to Reduce Exposure of Protected Health Information in Clinical Text' (2013) 20 Journal of the American Medical Informatics Association 342.
[101] See Jack Murtagh, 'The Math Hack You Didn't Know Was in Your Credit Card' (*Scientific American*, 1 November 2025) <https://www.scientificamerican.com/article/what-is-the-luhn-algorithm-the-math-behind-secure-credit-card-numbers/>.
[102] Zhenyu Zhao and others, 'Re-Examination of Rule-Based Methods in Deidentification of Electronic Health Records: Algorithm Development and Validation' (2020) 8 JMIR Medical Informatics e17622.
[103] John C Klensin, 'Internationalized Domain Names for Applications (IDNA): Definitions and Document Framework' (Internet Engineering Task Force 2010) Request for Comments RFC 5890 <https://datatracker.ietf.org/doc/rfc5890>; Wei Mao and Jiankang Yao, 'SMTP Extension for Internationalized Email' (Internet Engineering Task Force 2012) Request for Comments RFC 6531 <https://datatracker.ietf.org/doc/rfc6531>.

in formatting. Even small deviations such as a typo in a form, an OCR artifact from a scanned document, inserted or missing punctuation, or Unicode mix-ups (eg a Cyrillic character that resembles a latin letter) reduce recall rates.[104] Maintaining comprehensive dictionaries or 'gazetteers' helps partially but does not solve issues with unseen variants and new formats.[105] These issues are well known and described.[106] Rule-based systems struggle under the real-world messiness of data, which is why modern systems often employ them with additional layers, such as statistical and ML models.

Identifier detection and replacement: Rule-based approaches

Accessibility of methods: High. Available open-source and commercial pipelines (eg Philter,[107] scrubadub,[108] Presidio[109]) and common regexe or gazetteers listed in scientific publications. Straightforward to implement and deploy.

Considerations for further research: Deterministic behaviours make it easier to audit, but brittleness to out-of-domain formats can increase false negatives. OCR and transcription artefacts can also increase false negatives.

Method maturity: High. Decades-long use in production. Stable, well-understood performance.

Key resources: Datasets/benchmarks include the widely-used i2b2 challenge (now n2c2[110]) on longitudinal clinical notes, the OpenDeID corpus[111] (2021) on Australian EHR data, CEGS-N-GRID (2016) on psychiatric intake notes, the MEDDOCAN challenge[112] (2019) on Spanish

[104] Jiaxing Liu and others, 'OpenDeID Pipeline for Unstructured Electronic Health Record Text Notes Based on Rules and Transformers: Deidentification Algorithm Development and Validation Study' (2023) 25 Journal of Medical Internet Research e48145.
[105] eg the Ordnance Survey's National Street Gazetteer.
[106] Veysel Kocaman, Hasham Ul Haq and David Talby, 'Beyond Accuracy: Automated De-Identification of Large Real-World Clinical Text Datasets' (arXiv, 13 December 2023) <http://arxiv.org/abs/2312.08495>.
[107] See 'Philter Frequently Asked Questions' <https://www.philterd.ai/philter-faq/>.
[108] See 'Scrubadub: Clean Personally Identifiable Information from Dirty Dirty Text.' <https://github.com/LeapBeyond/scrubadub>
[109] Microsoft, 'Microsoft Presidio' <https://microsoft.github.io/presidio/>.
[110] See 'I2b2: Informatics for Integrating Biology & the Bedside' <https://www.i2b2.org/NLP/DataSets/>.
[111] See SREDH, 'OpenDeID Corpus Dataset' <https://www.sredhconsortium.org/sredh-datasets/opendeid-corpus-dataset>.
[112] See GuiGel, 'GuiGel/MedDocAn' <https://github.com/GuiGel/MedDocAn>.

synthetic medical records, the CRAPII dataset[113] (2024) on student essays, the BigCode PII dataset (2023) on source code snippets across 31 programming languages, and PII bench (2025), a carefully crafted dataset over 55 'PII' categories. Software tools include Presidio (Microsoft), scrubadub, pii-extract-base (PIISA), Philter (UCSF), piicatcher. Tools also exist to detect Unicode confusables, which might undermine these previous systems by using characters that look similar to humans but different to computers (ICU SpoofChecker). Cloud providers also offer identification and redaction tools (eg GCP Sensitive Data Protection, AWS Macie, Microsoft Purview).

Are these methods feasible in the future?: Under certain circumstances these methods are already feasible, deployed at scale and practical. This is when formats are known in advance and the available data are of high quality.

2.3.2.2. Statistical and ML approaches

Statistical and ML approaches emerged to address the brittleness of the rules-based ones. These vary from deep learning architectures (LSTM, BiLSTM) to transformer models (Bidirectional Encoder Representations from Transformers, or BERT[114]) and now modern LLM-based zero- and few-shot prompting.[115]

Unlike rule-based systems that look for exact patterns, these neural models learn to recognise direct identifiers by studying thousands of examples and building an internal representation of these target identifiers.

These models can handle formatting variations and transcription errors that get missed by rule-based systems; a phone number written as 0303

[113] See Langdon Holmes and others, 'The Cleaned Repository of Annotated Personally Identifiable Information' (2024) <https://educationaldatamining.org/edm2024/proceedings/2024.EDM-posters.88/index.html>.
[114] Jacob Devlin and others, 'BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding', *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)* (2019).
[115] Woojin Kim, Sungeun Hahm and Jaejin Lee, 'Generalizing Clinical De-Identification Models by Privacy-Safe Data Augmentation Using GPT-4' in Yaser Al-Onaizan, Mohit Bansal and Yun-Nung Chen (eds), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics 2024) <https://aclanthology.org/2024.emnlp-main.1181/>.

123 1113, 03031231113, or even 0303-123-1113 can all be recognised as the same number.

BERT-based approaches have been used extensively in healthcare to remove personal data in free-text clinical notes, striking a balance between computational resources needed, amount of training data annotations required, and accuracy or generalisability. Modern baselines and benchmarks include GLiNER (a generalist Named Entity Recognition (NER)),[116] ClinicalBERT, and OBI's Robust De-ID models. More recent approaches, including ones utilising larger language models with or without , have been shown to provide comparable or slightly better performance, though results are domain- and dataset-specific.

These methods are probabilistic which means they make educated guesses rather than producing deterministic results. Despite strong statistical accuracy scores in scientific evaluations and benchmarks, they can always miss even 'simple' identifiers when distribution drift occurs (that is when a model encounters data that differs from its training examples). This could include rare formats (a date written "2025-13-Nov" instead of a more standard format such as "13-Nov-2025"), creative obfuscation with misspelling or added punctuation, tokenisation issues (language models break words into processable chunks called tokens, and rare tokens might not be detected correctly), or boundary inconsistencies (whether "Jane Garcia Doe OBE" is tagged as one name or fragmented). This is an inherent issue, not only with language models but with any statistical method trained on real-world examples.

Identifier detection and replacement: Statistical and ML approaches

Accessibility of methods: Medium. Open source baselines (BERT-deid, BioBERT model for fine-tuning, GLiNER) exist but they require annotated, labelled data, fine-tuning, and compute. State-of-the-art methods may not be easily reproduced without domain expertise in software engineering and ML.

Considerations for further research: ML methods can help better detect direct identifiers and preserve utility when replacing identifiers. They can be brittle to distributional shift and to obfuscation.

[116] Urchade Zaratiana and others, 'GLiNER: Generalist Model for Named Entity Recognition Using Bidirectional Transformer' (arXiv, 14 November 2023) <http://arxiv.org/abs/2311.08526>.

Method maturity: Medium.

Key resources: Datasets/benchmarks same as above. Models can be trained using general named entity recognition (NER) tools such as flair, spaCy, medspaCy, stanza, GLiNER. Existing models include NeuroNER, ClinicalBERT, OBI 'Robust DeID', NLM Scrubber, as well as Presidio (Microsoft) using its Transformers NLP engine, guardrails-ai. Cloud platforms increasingly provide not only rule-based but also ML-based tools.

Are these methods feasible in the future?: Soon feasible. Statistical accuracy is comparable to or above rule-based methods in scientific studies, but requires further evaluation in AI development contexts across domains.

2.3.2.3. Hybrid or ensemble approaches

Hybrid or ensemble approaches attempt to reconcile the strengths of rule-based and ML methods by combining them. These methods are typically pipelines that combine rule-based components (regexes, gazetteers, format validation) with statistical, ML detection (eg contextually detecting whether "John Radcliffe" is a natural person in the dataset or the name of a hospital in Oxford). The two components can be combined (a) as a two-pass pipeline (one then the other) or (b) with a consensus/voting ensemble approach to work out which is preferred.

These methods are generally seen as better for practice, and now obtain high statistical accuracy (with high scores on institutional and benchmark datasets for these tasks such as i2b2[117] and OpenDeID Corpus[118]). It should be noted that these figures are sensitive to dataset composition (a study reports an F1 score, which is a form of performance metric dropping considerably when a system was trained on the i2b2 corpus but tested on

[117] Zengjian Liu and others, 'De-Identification of Clinical Notes via Recurrent Neural Network and Conditional Random Field' (2017) 75S Journal of Biomedical Informatics S34.
[118] Jiaxing Liu and others, 'OpenDeID Pipeline for Unstructured Electronic Health Record Text Notes Based on Rules and Transformers: Deidentification Algorithm Development and Validation Study' (2023) 25 Journal of Medical Internet Research e48145.

a different corpus from UF Health[119]) and to the specific design of the evaluation task.

There is no clear consensus on whether rules are still needed. On the one hand, hardcoded or learned rules can help with compliance and high statistical accuracy on direct identifiers (dates, NHS numbers, phone numbers, IBAN), reducing false positives and supporting any audit requirements. On the other hand, modern transformer-based approaches already cover most of these use cases with high accuracy without the need to maintain a database of rules, sometimes even without the need to provide any example training data.

Identifier detection and replacement: Hybrid or ensemble approaches

Accessibility of methods: Medium. End-to-end hybrid pipelines with code are available (OpenDeID[120], John Snow Labs's library[121]) but require careful configuration and often training examples, fine-tuning, and compute.

Considerations for further research: These systems should provide the highest statistical accuracy but require additional domain expertise and software engineering skills.

Method maturity: Medium.

Key resources: Datasets/benchmarks same as above. A simple pipeline can be assembled using existing rule-based and ML-based tools. Many libraries mentioned above such as Presidio (Microsoft) and OpenDeID already combine both approaches.

Are these methods feasible in the future? Feasible soon. Statistical accuracy is expected to be best within the domain it was trained in but requires evaluation in production environments given the ML components, which might not work well in new contexts they were not trained for.

[119] Xi Yang and others, 'A Study of Deep Learning Methods for De-Identification of Clinical Notes in Cross-Institute Settings' (2019) 19 BMC Medical Informatics and Decision Making 232.
[120] https://github.com/SREDH-Consortium/OpenDeID-Pipeline
[121] https://github.com/JohnSnowLabs/spark-nlp-workshop/tree/master/healthcare-nlp

2.3.2.4. Multi-agent LLM methods

Finally, a newer generation of systems attempt to coordinate multiple LLM-based components, either to balance complex trade-offs such as privacy and downstream utility or to perform a more adversarial detection — simulating an attacker. **RUPTA** is a canonical example: it couples an LLM that evaluates privacy, an LLM that evaluates utility in text data, and an LLM that iteratively edits text, all in a loop that halts when pre-set conditions are satisfied.[122] Similarly, **IncogniText** rewrites text data to mislead attribute-inference adversaries while preserving utility.[123] **Adanonymizer** provides a user-steerable trade-off between privacy and utility.[124] Pasch & Cha provide a fully automated evaluation framework to assess 'anonymisation' while preserving fluency[125], and **PrivacyPAD** learns a reinforcement learning (RL) policy to balance trade-offs between privacy leakage and task performance (whether a task requires identifiers or not)[126].

[122] Tianyu Yang, Xiaodan Zhu and Iryna Gurevych, 'Robust Utility-Preserving Text Anonymization Based on Large Language Models' in Wanxiang Che and others (eds), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Association for Computational Linguistics 2025) <https://aclanthology.org/2025.acl-long.1404/> accessed 10 October 2025.
[123] Ahmed Frikha and others, 'IncogniText: Privacy-Enhancing Conditional Text Anonymization via LLM-Based Private Attribute Randomization' (arXiv, 2 February 2025) <http://arxiv.org/abs/2407.02956>.
[124] Shuning Zhang and others, 'Adanonymizer: Interactively Navigating and Balancing the Duality of Privacy and Output Performance in Human-LLM Interaction' (arXiv, 27 January 2025) <http://arxiv.org/abs/2410.15044>.
[125] Stefan Pasch and Min Chul Cha, 'Balancing Privacy and Utility in Personal LLM Writing Tasks: An Automated Pipeline for Evaluating Anonymizations' in Ivan Habernal and others (eds), *Proceedings of the Sixth Workshop on Privacy in Natural Language Processing* (Association for Computational Linguistics 2025) <https://aclanthology.org/2025.privatenlp-main.3/>.
[126] Zheng Hui and others, 'PrivacyPAD: A Reinforcement Learning Framework for Dynamic Privacy-Aware Delegation' (arXiv, 16 October 2025) <http://arxiv.org/abs/2510.16054>.

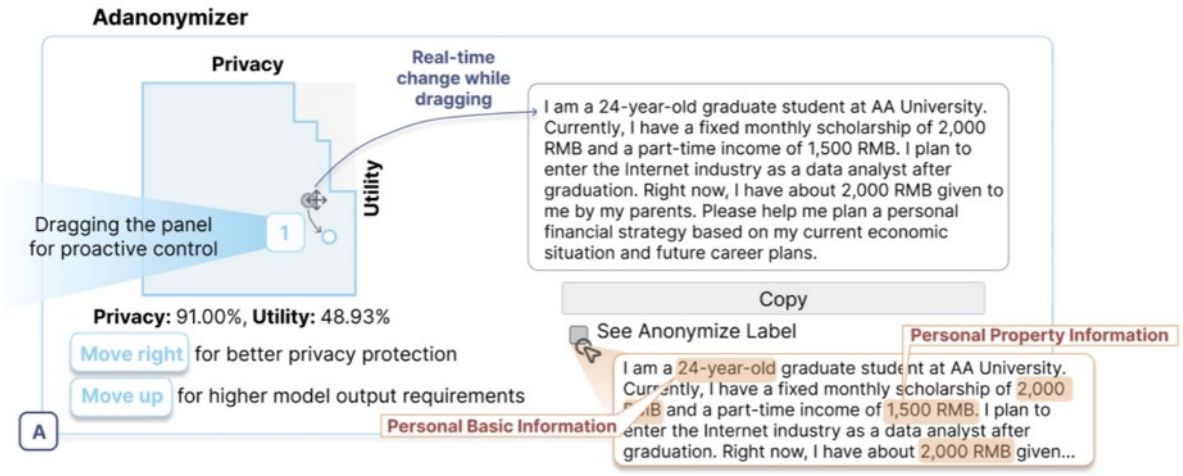


Figure 1: Adanonymizer.[127]

While multi-agent frameworks such as RUPTA have promise, there are several challenges with deploying them in practice. These frameworks typically optimise privacy and utility by iteratively editing text using LLM components that (a) evaluate privacy by inferring identifiers and (b) evaluate utility using labelled downstream tasks. This means that privacy is often tested without operationalising it via a '**motivated intruder**' threat model. Instead, privacy evaluation is based on whether an LLM can 'guess' the ground truth identity or the person concerned. To take an example from this paper:

> **(1, before RUPTA)** Jacques "Toto" Brugnon (11 May 1895 – 20 March 1978) was a French tennis player, one of the famous "Four Musketeers" from France who dominated tennis in the late 1920s and early 1930s. He was born in Paris and died in Paris. He was primarily a doubles specialist who won 10 Grand Slam doubles titles in the French, American, Australian and British championships…
>
> **(2, after RUPTA)** An individual (born in a certain era – passed away in another era) was a tennis athlete from a nation, part of a distinguished group from this nation who excelled in tennis during a historical period. This individual was born and passed away in a major city. Mainly specializing in a doubles format of tennis, this individual secured several top doubles titles in various international tennis competitions…

In practice, the RUPTA evaluator only measures the ability of *one* particular ML model on a text without the name of the person (the 'ground truth') included. This itself is not a useful assessment in the

[127] Shuning Zhang and others, 'Adanonymizer: Interactively Navigating and Balancing the Duality of Privacy and Output Performance in Human-LLM Interaction' (arXiv, 27 January 2025) <http://arxiv.org/abs/2410.15044>.

context of the "whose hands" consideration in data protection law, as it ignores the possibility of using supplementary data — including public data which is widely available. In effect, these types of proposals test against a limited type of automatable re-identification attacks using only an available machine learning model.

In addition, frameworks such as RUPTA require a concrete utility function to score edits and know which are good, and which are bad. Many scientific data public disclosures lack a single pre-specified task, and utility is often open-ended and evolving.

Identifier detection and replacement: Multi-agent LLM Methods

Accessibility of methods: Low. Several systems mentioned above have their code publicly accessible, but their operation heavily relies on LLMs which in turn can be time- and resource-intensive.

Considerations for further research: These systems should provide high statistical accuracy too but require additional domain expertise and substantial software engineering skills.

Method maturity: Low. Most work cited above is very recent, with limited evidence of real-world deployment and still-evolving threat models, with attackers using new technologies in new ways.

Key resources: Datasets/benchmarks same as above, with some models such as RUPTA introducing their own benchmarks. Limited availability of deployable code due to less scientific research.

Are these methods feasible in the future? They are not feasible yet for real-world use but improving as the relevant research advances.

There is credible evidence that state-of-the-art methods can be deployed locally, on premise. This includes rule-based approaches (with open-source frameworks like Microsoft Presidio providing plug-and-play recognisers, regexp, checksum, spacy/hugging-face named entity recognition, etc; or Philter, audited for HIPAA compliance in the US). This also includes LLM-based approaches, with local models offering zero-/few-shot prompt methods that rival GPT-4 on experiments.

That said, all major cloud firms already offer direct identifier detection and APIs that facilitate their replacement or redaction.[128] They often use the

[128] See eg Microsoft Presidio; Google Sensitive Data Protection; Amazon Comprehend; Alibaba Data Security Center.

US law term PII, but what they really mean is only direct identifiers, which are the simplest but far from the only way to classify something as personal data in the UK (see Identifiers). These tools effectively seek to recognise certain information in text, typically identifiers like bank details, IP addresses, API keys, emails, and similar, using different analysis techniques from regexes to more advanced ML techniques (mentioned above), without necessarily clarifying which technique is being used.

The price of these tools for one-off jobs is low, in the range of 2–5 USD for inspecting and redacting a gigabyte of information.[129] There is also the choice to use open source tools which would not incur an API cost, such as Microsoft Presidio, which has been released open-source by the firm.

2.3.2.5. Threat model considered

Detection and replacement methods are generally evaluated in an 'honest' threat model, using benchmarks on which scoring high means identifying and replacing specific categories of direct identifiers such as names, phone numbers, email addresses, etc, and seeing if the replacement was made. These benchmark datasets often align with US healthcare settings by following HIPAA categories.

Unfortunately, these methods are rarely evaluated on a 'malicious' threat model against adversarial re-identification techniques, which can use more than just the specific direct identifiers. Research on clinical notes shows, perhaps unsurprisingly, that de-identification of real clinical notes does not necessarily protect records against a membership inference attack (determining if a patient was included in a particular de-identified dataset)[130] nor against singling out and identifying records of particularly unique patients.[131]

Few research studies analysed the adequacy of detection methods in the context of LLM training. A 2023 study from the University of Waterloo and Microsoft Research showed that taking domain datasets in health or law, 'scrubbing' (referring to removal through either deletion or obfuscation) them using Named Entity Recognition (NER) approaches discussed above, and training LLMs reduces performance for the resulting LLM while offering limited protection against state-of-the-art privacy attacks. They

[129] See eg https://cloud.google.com/sensitive-data-protection/pricing.
[130] Atiquer Rahman Sarkar and others, 'De-Identification Is Not Enough: A Comparison between de-Identified and Synthetic Clinical Notes' (2024) 14 Scientific Reports 29669.
[131] John X Morris and others, 'DIRI: Adversarial Patient Reidentification with Large Language Models for Evaluating Clinical Text Anonymization' (2025) 2025 AMIA Summits on Translational Science Proceedings 355.

also identify a "shortage of good benchmark datasets for measuring PII leakage" in the context of LLM training, which requires larger datasets than most discussed above.

Proposed roles for the ICO

There is a solid ecosystem of methods to remove direct identifiers in LLM training datasets, but a lack of standardised evaluations to determine if these techniques provide any meaningful anonymisation or what level of reducing identifiability, especially outside US contexts such as HIPAA compliance. There is also a need for research on the downstream impact of LLMs training on heavily sanitised data. This in turn, makes it difficult to assess how protective they would be when applied as safeguards in the context of data protection.

The authors of this paper propose that this could be an opportunity for the ICO to define best practices for technical approaches that detect and replace direct identifiers in AI/LLM training dataset contexts, so that claims as to the efficacy of safeguards are credible and auditable. This could include best practices in terms of a given approach's:

Scope and target data. Controllers should explain which categories and types of personal data (including what special category data) their system is meant to find and which languages or contexts are covered.

Performance metrics for personal data detection and removal. Evaluations should report an approach's performance per identifier type rather than as a single overall score. Reported figures should be easy to understand (eg number of errors per document, per thousand words).

Confidence and human review. ML-based systems must provide meaningful confidence levels or decision thresholds. These thresholds should be set by controllers and not by AI model/software providers.

Generalisation to real-world data. Tests should be performed on realistic example materials, instead of relying on benchmarks or synthetic data.

Adversarial robustness. Tests should be performed against common real-world edge cases (eg typos, OCR and transcription errors, unusual punctuation, mixed language, look-alike characters). Controllers should also perform 'motivated intruder' tests that attempt to re-identify people despite data sanitisation.

2.3.3. Stylometry and profiling de-identification techniques

This section covers techniques that use properties of writing, such as one's writing style or preferred topics,[132] to link a text to an author. These techniques can be applied within a pre-determined set of candidates/individuals (**attribution**), to determine whether two texts were written by the same person (**verification**), or to infer/predict traits such as age or gender (**profiling**).

Stylometry signals enable both singling out and linkability even when explicit identifiers are comprehensively removed. This also means that, where style alone reliably links records to the same person, removing names or emails should not be presumed to anonymise data. Stylometry becomes *biometric* data processing under the UK GDPR when behavioural traits, such as writing style, are used to single out or verify individuals' identity.[133]

Research has shown that reliable authorship attribution (linking content to an individual) and verification are feasible across blogs, reviews, tweets, books, and in domain-specific settings. These have also shown success when applied on social media data (such as posts), notably through benchmark evaluations such as **PAN @ CLEF.**[134] Methods have developed over time. Previous authorship identification methods used more specific and manually extracted features such as patterns from linguistically tagged sentences, word selections, sentence length, punctuation type, etc. Today, deep learning techniques are often used, which do not require manual selection of features but instead try to locate patterns across text data.[135]

[132] Yunita Sari, Mark Stevenson and Andreas Vlachos, 'Topic or Style? Exploring the Most Useful Features for Authorship Attribution' in Emily M Bender, Leon Derczynski and Pierre Isabelle (eds), *Proceedings of the 27th International Conference on Computational Linguistics* (Association for Computational Linguistics 2018) <https://aclanthology.org/C18-1029/>.

[133] Note that biometric data means personal data resulting from "specific technical processing relating to the physical, physiological or **behavioural characteristics** of a natural person, which allow or confirm the unique identification of that natural person" [emphasis added].

[134]Efstathios Stamatatos and others, 'Overview of the PAN/CLEF 2015 Evaluation Lab' in Josanne Mothe and others (eds), *Experimental IR Meets Multilinguality, Multimodality, and Interaction* (Springer International Publishing 2015).

[135] Xie He and others, 'Authorship Attribution Methods, Challenges, and Future Research Directions: A Comprehensive Survey' (2024) 15 Information <https://www.mdpi.com/2078-2489/15/3/131> accessed 30 November 2025.

Academics Staab et al. show that current LLMs, without any fine-tuning, can infer personal attributes such as location, income, sex with 85% accuracy from real Reddit profiles.[136]

What once required heavy feature engineering, or later ML domain expertise, can now often be attempted with zero-shot LLM prompting — effectively using LLMs without finetuning or retraining them with new data from the task you are giving them[137] — lowering the barrier to entry to determine the identity of the author. In addition, modern LLMs have been shown to memorise vast amounts of online and offline data (since they have been trained on both),[138] which may further reduce the effort required to identify individuals from short snippets of texts they wrote themselves. Automated LLM-based agents can also nowadays be used to automatically search online activities (such as public posts on pseudonymous online platforms) to generate profiles of individuals,[139] without the need to collect text corpora as a prior step.

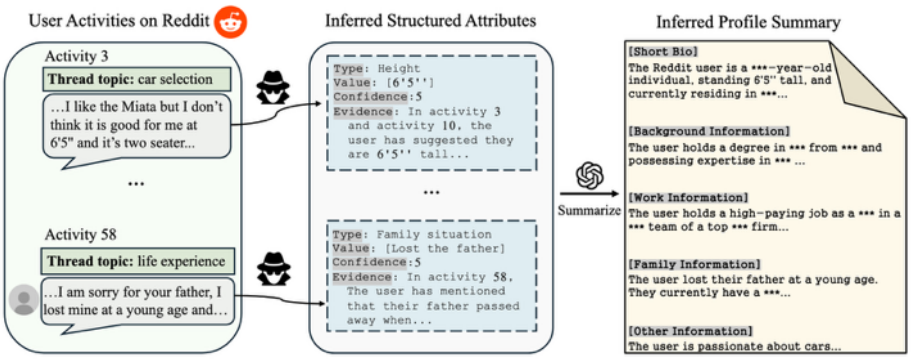


Figure 2: An automatic pipeline to search online activities for relevant profiles.[140]

However, benchmark and lab results can overstate the real-world performance of stylometry and profiling techniques to identify individuals due to several factors:

[136] https://arxiv.org/abs/2310.07298
[137] Baixiang Huang, Canyu Chen and Kai Shu, 'Can Large Language Models Identify Authorship?' in Yaser Al-Onaizan, Mohit Bansal and Yun-Nung Chen (eds), *Findings of the Association for Computational Linguistics: EMNLP 2024* (Association for Computational Linguistics 2024) <https://aclanthology.org/2024.findings-emnlp.26/>.
[138] Nicholas Carlini and others, 'Quantifying Memorization Across Neural Language Models', *Proceedings of the International Conference on Learning Representations* (ICLR 2023) <http://arxiv.org/abs/2202.07646>
[139] Yuntao Du and others, 'Automated Profile Inference with Language Model Agents' (arXiv, 18 May 2025) <http://arxiv.org/abs/2505.12402>
[140] Yuntao Du and others, 'Automated Profile Inference with Language Model Agents' (arXiv, 18 May 2025) <http://arxiv.org/abs/2505.12402>

Cross-topic/genre shifts typically reduce statistical accuracy. Most attribution/verification/profiling benchmarks (meaning datasets a technique is tested on) are still topic- and domain-aligned, meaning they expect a certain context in the writing. Classical works of scholarship in this area train and test on texts from similar topics or distribution, and so we cannot easily rely on them to make highly generalisable statements. Evidence shows that cross-topic or cross-domain evaluations (eg testing a law specific technique on healthcare data, etc) lead to a drop in performance compared to within-topic ones. For example, performing authorship attribution between an email corpus and a dataset of a holiday resort online reviews would be expectedly challenging.

Obfuscation techniques are techniques designed to limit the risk of authorship attribution by removing or changing aspects of the text. Done correctly, they can limit the risk of attribution and verification of specific guessed authorship.[141] Manual edits can target specific attributes that are high-risk in terms of facilitating identification such as location names or idiosyncratic phrases. Done automatically, basic obfuscation often relies on heuristics such as removing stop words (such as 'um', 'uh', 'like' which can be used in revealing ways), changing spelling, punctuation, casing, or lexical substitution (replacement of a word in a way that still makes sense, even if some meaning might change). Nevertheless, these heuristics can still be ineffective against models that rely on more advanced stylistic features such as part-of-speech tags (labels assigned to words in a sentence used in linguistics to indicate words' grammatical category) or reading complexity scores[142]. Recent research has attempted to use more advanced NLP and LLM methods to heavily rewrite text in a manner that limits identifiability while preserving content and fluency (see Figure 3). While these techniques work well on structured formats such as news or emails, imitation is harder in messy conditions such as forums or blogs[143]. At present, many benchmarks assume 'clean' text and

[141] Malik Altakrori and others, 'A Multifaceted Framework to Evaluate Evasion, Content Preservation, and Misattribution in Authorship Obfuscation Techniques' in Yoav Goldberg, Zornitsa Kozareva and Yue Zhang (eds), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics 2022) <https://aclanthology.org/2022.emnlp-main.153/>.
[142] Chris Emmery, Ákos Kádár and Grzegorz Chrupała, 'Adversarial Stylometry in the Wild: Transferable Lexical Substitution Attacks on Author Profiling' in Paola Merlo, Jorg Tiedemann and Reut Tsarfaty (eds), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (Association for Computational Linguistics 2021) <https://aclanthology.org/2021.eacl-main.203/>.
[143] Zhengxiang Wang and others, 'Catch Me If You Can? Not Yet: LLMs Still Struggle to Imitate the Implicit Writing Styles of Everyday Authors' in Christos Christodoulopoulos and others (eds), *Findings of the Association for Computational Linguistics: EMNLP 2025*

attribution/verification/profiling is seldom tested under such adversarial conditions — which could mean lesser performance in practice. On the other hand, their results on corpora such as social media datasets do not assume that users in the wild are actively already obfuscating revealing parts of their text (such as by taking active measures to manage their privacy, as widely documented in online behaviour[144]) — which could mean higher performance in practice than benchmarks would suggest.

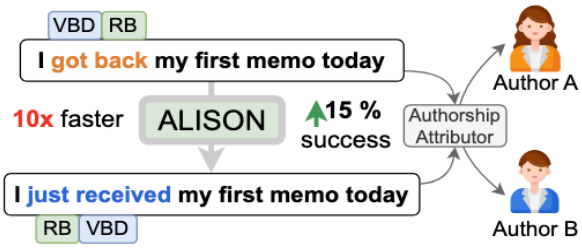


Figure 3: Figure from Xing et al.[145] concerning attempting to limit identifiability while preserving content and fluency.

In the context of LLM training datasets such as the ones that are the focus of this report, it is important to note that attempts to strip style often reduce the usability and utility of the data[146]. Training data that include multiple texts per person (eg prolific bloggers, reviewers, forum users) will increase linkability and the chance of *authorship* inference, especially when it is easy to *verify* that these texts belong to the same author.

(Association for Computational Linguistics 2025) <https://aclanthology.org/2025.findings-emnlp.532/>.

[144] Alice E Marwick and danah boyd, 'Networked Privacy: How Teenagers Negotiate Context in Social Media' (2014) 16 New Media & Society 1051.

[145] Eric Xing and others, 'ALISON: Fast and Effective Stylometric Authorship Obfuscation' (2024) 38 Proceedings of the AAAI Conference on Artificial Intelligence 19315.

[146] Malik Altakrori and others, 'A Multifaceted Framework to Evaluate Evasion, Content Preservation, and Misattribution in Authorship Obfuscation Techniques' in Yoav Goldberg, Zornitsa Kozareva and Yue Zhang (eds), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics 2022) <https://aclanthology.org/2022.emnlp-main.153/>.

2.3.3.1. Mitigations

Automated obfuscation (AO) or rewriting techniques can substantively lower the success of attribution/verification/profiling attacks, but have been shown to degrade the utility of a given dataset in terms of fluency and semantics, which may be important in downstream research activities.[147] While AO works well to protect against older attribution or automated verification (AA/AV) techniques, there is no consensus on whether this is still the case for strong, state-of-the-art AA/AV techniques. More practical, system-level mitigations should include (1) limiting the total amount of text exposed to a model for training per author/individual, since AA/AV/profiling statistical accuracy grows with the amount of provided records, (2) avoiding datasets with narrowly-focused sources and topics that make stylometric and profiling easier, and (3) empirically evaluating AA/AV/profiling technique attacks by deploying 'motivated intruder' threat models, both with human studies and strong automated attacks, rather than relying solely on somewhat optimistic benchmark results.

Proposed role for the ICO

Stylometry and profiling techniques to identify individuals are under-researched and evaluated in the context of datasets for training generative AI LLMs, with fewer safeguards deployed in the field at present. This is potentially due to the sheer complexity of protecting against authorship attribution and verification in very large dataset.

The authors of this paper propose that there is an opportunity for the ICO to support better research and empirical evaluations in this field, along the following dimensions:

Scope. Stylometry should be treated as a credible risk, with the potential to identify individuals in text datasets.

Scrubbing and anonymity claims. Simply detecting and removing a small set of direct identifiers (eg HIPAA identifiers in the US) should not be enough for a dataset to be considered anonymised under the UK data protection law, especially without considering the risk of identifying the author/individual.

[147] Martin Potthast, Matthias Hagen and Benno Stein, 'Author Obfuscation: Attacking the State of the Art in Authorship Verification?' (*Working Notes of CLEF 2016 — Conference and Labs of the Evaluation Forum*, 2016) <https://ceur-ws.org/Vol-1609/16090716.pdf>.

Benchmark settings. Benchmarks datasets in stylometry have long been very small — datasets of around 10 authors are common. These results cannot be used to extrapolate performance on a scale as large as that of LLM training data.

Biometric processing. We believe that controllers using stylometry to uniquely identify people (eg to deduplicate user submissions) could be considered processing behavioural biometric data, which may require an Article 9 legal basis.

Mitigations. The authors of this paper propose that the ICO should support realistic benchmarks and evaluations in the context of LLM training datasets, as there is a lack of evidence in relation to efficacy and utility at present.

2.4. Image datasets

Image data is widely used in ML training lifecycles. Such data poses significant privacy and data protection challenges. This section specifically examines still images and the challenges they present. Video raises similar but more complex problems, due to computational costs and greater uncertainty as to effectiveness of potential safeguards given the scarce literature in this field, although we expect that many of the methods applicable in the context of still images will have some utility in the context of videos. Video editing for reducing identifiability is more nascent as a field and more specialised when that is done on datasets meant for research purposes. Any consideration of extrapolating the following techniques in a video context should be aware that the robustness of these methods even in the context of still images is uncertain.

We focus on five major data protection concerns below:

- **Facial biometrics.** Faces are uniquely identifying biometrics. We review methods and software to detect faces as well as synthesise replacement faces—digitally swapping a real person for a fictional one while preserving pose, expression, and background context.
- **Soft biometrics and ephemeral identifiers.** People can be identified through many features beyond faces. We review so-called soft biometrics, which include gait, tattoos and other less stable features of an individual, and features that may be stable only within a certain time frame — ephemeral identifiers like clothing. We also cover research in imaging related to full-body synthesis, testing if it is possible to entirely replace a person with a fictional one who, for all intents and purposes, appears to be doing the same thing.
- **Geolocation.** We then examine an area that is generally under-researched in ML: geolocation inference. We address this in two ways. First, geolocation can be inferred by AI models, and this has been studied, but protecting against this and mounting defences has received very little attention in the literature — we explain the limited defences that have been proposed. Second, geolocation can be inferred from physical characteristics of a photo, in particular the location of shadows. This is already a widely used method in open-source intelligence (OSINT) but, as far as we can see, has received no attention in the ML or privacy/data-release literature. This might be quite concerning because such analysis can be highly revealing of an individuals' specific location, even in

scenes that are otherwise difficult to identify based on landmarks or other features.

- **Text in images.** Photos often capture unintended or incidental text in the images themselves (eg an individual being photographed in front of a 'McDonalds') — this is called 'scene text'. Scene text can comprise all possible types of text: street signs, licence plates, a passport left on a table, or an identity card. There are many methods for both editing and erasing this text, but not all methods are robust at detecting what might be personal data (and how sensitive if so) as well as what is needed to preserve an image's utility as a useful training data point for researchers. This area contains a lot of interesting technological solutions, but requires more research to be combined into a usable pipeline for research safeguards that can be deployed by controllers.
- **Metadata**. Still photo files often embed creation timestamps, GPS coordinates, and individual device information. Removing this metadata is common, well-known, straightforward and should be undertaken by all actors as a default. We already covered metadata in the audio section above, but briefly consider it again here.

2.4.1. Facial data and its data protection risk

Facial images are risky for a widely understood number of reasons. It is broadly described as a 'hard' biometric, and large-scale amassing of facial datasets has been the subject of enforcement action, such as ongoing enforcement attempts against Clearview AI, and prohibitions, such as the partial prohibition of such dataset collection in the EU AI Act. We do not consider it necessary to explain the importance of facial recognition risk from such datasets in detail here.

2.4.1.1. Limiting the risks of identification in facial image datasets and facial synthesis

Removing facial images from datasets can be done in a wide variety of ways and is field that has been extensively studied in the literature.

Several such techniques exist. Academics Cao and others classify them into *pixel level*, *representation level* and *semantic level*.[148]

[148] Jingyi Cao and others, 'Face De-Identification: State-of-the-Art Methods and Comparative Studies' (arXiv, 15 November 2024) <http://arxiv.org/abs/2411.09863>. Some methods attempt to make reversible facial privacy — effectively transforming the face into a different one in a way that can be transformed back with a password or

Pixel-level tools identify the facial area and then blur it, replace it with another face, or perturb certain aspects. This can be effective and quite computationally cheap (eg a very powerful blur or a replacement with a block colour), but can significantly lose aesthetics and utility in the process. Pixel level attempts to replace individual faces generally suffer when faces are in odd angles or similar, and the outcomes of those processes are not always that visually convincing. Furthermore, some simplistic but commonly used pixel blur techniques, such as Gaussian blurring, can be vulnerable to following adversarial attacks aiming to reconstruct the face.[149]

Representation-level tools process facial data by encoding them in the way a facial recognition technology (FRT) model might do — as a compressed, mathematical representation, or 'latent vector' — and apply a statistical transformation to that resulting vector. Following that they reverse the process to project that altered vector back onto the image as a replaced face. These transformations attempt to put some statistical distance between the original, real individual face and the 'replacement' fabricated face — enough to make it hard to identify the original face, but not too much that would make that fabricated face seem out of place, unnatural or unconvincing. A visually indicative example from one paper in this area is shown in Figure 4. Some go beyond this and attempt to provide formal (mathematically assured) privacy guarantees based on these transformations.

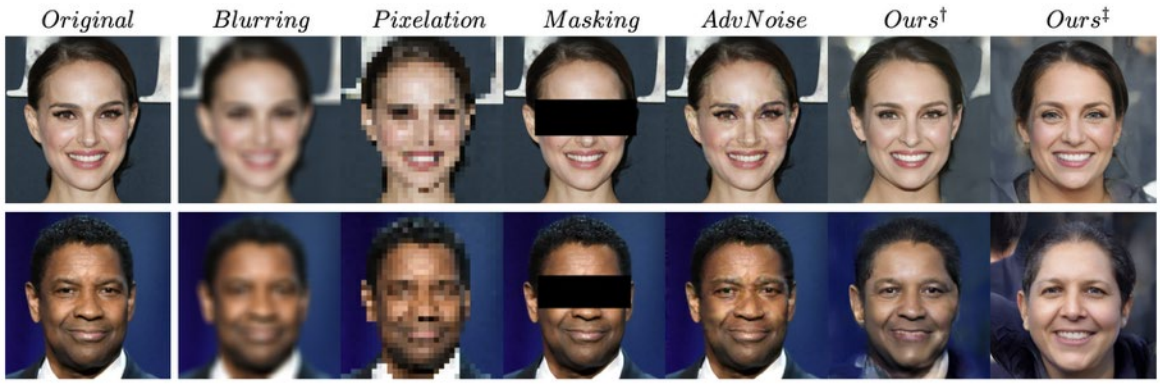


Figure 4: An illustrative example of a range of more classic methods compared to representational alteration of a face (final two rows).[150] The penultimate row shows a

equivalent secret information. This does not seem relevant to the task here so we do not examine it further.

[149] Haoyu Zhai and others, 'Restoring Gaussian Blurred Face Images for Deanonymization Attacks' (arXiv, 14 June 2025) <http://arxiv.org/abs/2506.12344> accessed 15 November 2025.

[150] Tao Li and Min Soo Choi, 'DeepBlur: A Simple and Effective Method for Natural Image Obfuscation' (arXiv, 31 March 2021) <http://arxiv.org/abs/2104.02655>.

minor distortion in representation space; the final row shows a major distortion (introducing an extremely large number of changes).

Representation methods present challenges because the major distortions needed to protect data from reidentification attacks — particularly faced with an adversary using advanced tooling — introduce differences that might change aspects of how an individual is perceived as well as their formal identity. This change can reduce the utility of dataset for research purposes that require the contextual information about a face or individual even though they do not require that individual's precise identity. Take the image above in Figure 4. The far right column shows two seemingly distinct people, but the top image no longer appears to have the formality of an awards show; the bottom image no longer would appear to have the expression of perhaps a television host (as well as a seemingly altered ethnicity).

As a result, **semantic** level methods try to separate (disentangle) out the *identity* of a person (ie the kind of signal that a FRT model would try to find) from other attributes of that individual, such as their expression; lighting; accessories they are wearing, or similar.[151]



Figure 5: Results (right side of each pair) from the application of the 2022 IdentityDP framework for transforming people's identities (interpersonal variance) whilst attempting to retain other attributes (intrapersonal variance).[152]

Consider the results from a 2022 work on this in Figure 4 above: you can see that the individuals, while still a little distorted, retain features such as makeup, accessories such as glasses, and the type of lighting, pose or expression.

[151] See eg Yunqian Wen and others, 'IdentityDP: Differential Private Identification Protection for Face Images' (2022) 501 Neurocomputing 197.
[152] Yunqian Wen and others, 'IdentityDP: Differential Private Identification Protection for Face Images' (2022) 501 Neurocomputing 197.

Some of the most promising recent work in this field attempts to obtain the benefits of disentanglement of identity from attributes in the semantic approach, while using the simplicity and aesthetic ease of the representation approach.[153] These use diffusion AI models to mix the generation of new faces with the replication of the existing ones, allowing people to select a parameter between the current face, and a face that has similar attributes but no identity. This produces high quality imagery.



Figure 6: An approach where a face is swapped with one that is interpolated between a face with a very different identity to the source face, and the source face itself. The left photo in each batch is the original, and the next photo one with a small perturbation, and the final one with a large perturbation.[154]

We consider the final method as an example of a state-of-the-art facial perturbation system which produces appealing outputs. The models and tools for this system are available open-source, as is the training code to make improved models if required. Such improved models may be required, as the paper referenced above indicates the method they propose performs poorly on young children and on certain ethnic minorities underrepresented in particular training datasets, notably Asian individuals.

Computationally, the resources required are of the same range as producing an image by deploying a Stable Diffusion model. There is also a quality–computational efficacy trade-off. Running on a single GPU, each run of a normal sized image (512x512 pixels) would likely take in the region of 15–20 seconds to generate using the method proposed in the above paper and relevant repository. These computational runs could happen in parallel, meaning datasets of hundreds or thousands of images can be processed easily. Nevertheless, this would currently be costly

[153] Han-Wei Kung and others, 'Face Anonymization Made Simple', Proceedings of the Winter Conference on Applications of Computer Vision (WACV) (2025).
[154] Han-Wei Kung and others, 'Face Anonymization Made Simple', Proceedings of the Winter Conference on Applications of Computer Vision (WACV) (2025).

when it comes datasets comprising millions of images as you often see in a generative AI context. It is likely however that with more research focus in this direction, this cost would eventually come down. Other methods discussed above use GANs rather than diffusion models, and are 10-15x faster, although this comes at a quality cost.

2.4.1.2. Impact on performance of downstream tasks

Some studies have considered whether the above approaches to lower the risk of reidentification or anonymising data would damage downstream training based on the resulting data. They find that for certain training tasks, such as ML training for autonomous vehicles (eg the Cityscapes and BDD100K datasets), there is little-to-no impact from anonymising facial images.[155] Within this small impact, realistic facial replacement (eg where faces are replaced, swapped or similar in a way that looks realistic) slightly outperforms facial redaction (ie blurring or masking).

For tasks where the human component is more central to the utility of the data, such as the classifications we see in the COCO dataset (Common Objects in Context is a dataset of scenes with tagged descriptions of what is happening in them), masking-based degradation has adverse effects on performance.[156] Using realistic replacements does bring a performance degradation in classification tasks based on this type of datasetsmall one, and researchers are exploring ways to reducing this gap in future.

Facial re-synthesis

Accessibility of methods: Medium-High — several open-source models and frameworks.

Considerations for further research: Specific identities of people are rarely required by research tasks and so edits to these which otherwise

[155] Håkon Hukkelås and Frank Lindseth, 'Does Image Anonymization Impact Computer Vision Training?' (2023) Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2023, pp. 140-150 <https://openaccess.thecvf.com/content/CVPR2023W/WAD/html/Hukkelas_Does_Image_Anonymization_Impact_Computer_Vision_Training_CVPRW_2023_paper.html>.

[156] Håkon Hukkelås and Frank Lindseth, 'Does Image Anonymization Impact Computer Vision Training?' (2023) Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2023, pp. 140-150 <https://openaccess.thecvf.com/content/CVPR2023W/WAD/html/Hukkelas_Does_Image_Anonymization_Impact_Computer_Vision_Training_CVPRW_2023_paper.html>; Jun Ha Lee and Su Jeong You, 'Balancing Privacy and Accuracy: Exploring the Impact of Data Anonymization on Deep Learning Models in Computer Vision' (2024) 12 IEEE Access 8346.

preserve quality should not damage research data utility.

Method maturity: Medium — well-performing models but lacking the benchmark maturing of areas like approaches to limit the risk of identification in audio datasets.

Key resources: No major leaderboards; models and tools such as Facial Anonymization Made Simple on GitHub.

Are these methods feasible in the future?: These are already feasible. The highest quality methods are computationally feasible and appropriate for medium sized datasets (thousands of images), but expensive on the very largest datasets (millions of images).

2.4.2. Soft biometrics and ephemeral identifiers

2.4.2.1. Associated risk

Soft biometrics and ephemeral identifiers are typically not as unique as 'hard biometrics'. Many soft biometrics, such as gait recognition or ephemeral identifiers, such as clothing choice, are typically used to identify individuals within a limited setting, such as to link different data about an individual between cameras in a public space where only a limited number of people are walking around between them. While FRT can identify individuals at significant scale (ie datasets used in or sold to law enforcement), most soft biometric techniques will not successfully alone uniquely identify an individual in an extremely large dataset, in particular where the candidate population of potential individuals is extremely large, because they are not unique enough characteristics of a person. Ephemeral identifiers will likely only identify people within a set timeframe due to their lack of even the more limited permanence and stability of soft biometrics.

This does not mean there will not be times where this is the case: rare clothing or tattoos, for example, may uniquely identify a natural person even in those situations. Tattoo recognition is such a widely sought after and adopted technique that NIST, the US government technical standardisation agency, has undertaken extensive evaluation and best practice guidance on its optimal use.[157] We do not look at it here because

[157] See eg Mei Ngan, George W Quinn and Patrick Grother, 'Tattoo Recognition Technology – Best Practices (Tatt-BP) Guidelines for Tattoo Image Collection' (*NIST*, February 2016) <https://nvlpubs.nist.gov/nistpubs/ir/2016/NIST.IR.8109.pdf>.

it is extremely specific, but note that limited, older work exists on tattoo deidentification.[158]

Such re-identification techniques can be used in combination with other methods or other pieces of data. Even where facial obfuscation has been applied (see Section 2.4.1.1), if the same individual is present in multiple digitally processed photos (eg wearing the same clothing at the same event), then more inferences can be made about who that individual might be potentially leading to their identification.

The outputs of any given technique that aims to anonymise facial images might be more vulnerable to adversarial attacks to reidentify if multiple fabricated faces (which may be different while still sharing some similarities) are available. Furthermore, some of the fabricated images may still retain contextual information discussed in this section — such as a textual name hardcoded into them, or unredacted metadata, or an inferred geolocation — which would then allow an adversary to both reidentify the individual in the image but also connect it to multiple images linked by the soft biometrics or similar.

2.4.2.2. Full-body synthesis

To deal with these types of issues, researchers have attempted to — ambitiously — explore *full-body synthesis*. This is effectively asking a model to replace the whole body of an individual with that of a different individual while preserving some contextual similarity.

This is the most broadly used technique in this space. This area of research attempts to segment people from the image and replace them with alternatives that retain the context while swapping both their faces (if shown) and other aspects of their appearance that might serve as soft biometrics or similar.[159] These techniques have promise, particularly in

[158] Tomislav Hrkać and others, 'Deep Learning Architectures for Tattoo Detection and De-Identification', 2016 First International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE) (2016) <https://ieeexplore.ieee.org/abstract/document/7528402>. This work is a little behind the state-of-the-art in other fields now and so we do not look at tattoo deidentification separately. If a research dataset has many tattoos visible it is worth considering this risk explicitly in light of the research available at the time of applying the safeguard.

[159] Håkon Hukkelås and Frank Lindseth, 'Synthesizing Anyone, Anywhere, in Any Pose' (2024) <https://openaccess.thecvf.com/content/WACV2024/html/Hukkelas_Synthesizing_Anyone_Anywhere_in_Any_Pose_WACV_2024_paper.html>; Håkon Hukkelås and Frank Lindseth, 'DeepPrivacy2: Towards Realistic Full-Body Anonymization' (2023) <https://openaccess.thecvf.com/content/WACV2023/html/Hukkelas_DeepPrivacy2_Towards_Realistic_Full-Body_Anonymization_WACV_2023_paper.html>; Pascal Zwick and

terms of their general application, but as it stands, suffer from performance limitations and are immature for large-scale production use.



Figure 7: Two models (Hukkelås and Lindseth, top left; Zwick and others, top right) running full-body synthesis on the famous 2014 Ellen DeGeneres Oscar selfie (bottom). Note that the model seeks to preserve much of the clothing but does at times change it (necklaces, hair, similar).



Figure 8: Example of altering the original (right) into a new version (left) using the model from Zwick and others that alters features other than the face. Notice the current limitations and distortions (fingers, wording).

others, 'Context-Aware Full Body Anonymization Using Text-to-Image Diffusion Models' (arXiv, 17 October 2024) <http://arxiv.org/abs/2410.08551>.

Full-body synthesis may be an inappropriate safeguard in many situations, because the impact on the performance on a model trained on data anonymised in this way may be substantial compared to the reduction of residual data protection risk. As it stands, it is difficult to recommend this approach as a proportional intervention in the dataset preparation stage of a generative AI pipeline. This may change as the technology develops.

2.4.2.3. Impact on the performance of a model trained on data subjected to full-body synthesis

With current methods, full-body synthesis does, unlike FRT tasks, degrade the resulting generative AI model's performance even for tasks where the human is not the centre of the task, such as vehicles for autonomous vehicle training.[160] Academics hypothesise that this is because the current methods require cropping individuals out and replacing them, and in doing so, the context they appear to be in may change.

Full body synthesis

Accessibility of methods: Medium — a small number of open-source models and frameworks are available.

Considerations for further research: Compared to facial images, perturbation of entire bodies can change the context and lead to less predictable results in terms of the trained model. Evidence indicates it even undermines downstream model training even for tasks where exactly what the human is doing is less of a focus (eg datasets to train ML models for autonomous driving).

Method maturity: Low-Medium — models are impressive given the small scale of the research but no leaderboards or competitions. Currently there are only a few researchers working in this area.

Key resources: Only the papers referenced above- there are no commercial implementations yet as far as the authors of this report are aware.

Are these methods feasible in the future? Feasibility may increase in

[160] Håkon Hukkelås and Frank Lindseth, 'Does Image Anonymization Impact Computer Vision Training?' (2023) Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2023, pp. 140-150 <https://openaccess.thecvf.com/content/CVPR2023W/WAD/html/Hukkelas_Does_Image_Anonymization_Impact_Computer_Vision_Training_CVPRW_2023_paper.html>.

along with general image synthesis involving individuals.

2.4.3. Geolocation inference

2.4.3.1. Shadows

Aspects of an individual's location can be ascertained through indirect, unexpected means. One method for this is the length and direction of a shadow.

It has long been known that static cameras can be geolocated, and photographs timestamped, using six or more photos from these cameras, with no need to know the date or time (although the year the photos were taken in may remain ambiguous).[161] Fewer photos are required if multiple shadows are visible.[162]

Given the date and time of a photograph — such as retained through the EXIF embedded metadata in an image file (see Section 2.1.1) — and enough clear sunlight, a photo can be narrowed down to a ring of possible locations on Earth where the photo could have been taken, using widely available open tools commonly used by open source intelligence (OSINT) researchers.[163] This is still a large area (unlike the techniques above which are significantly more precise) but it could be used to triangulate further information about an image.

[161] Imran N Junejo and Hassan Foroosh, 'Estimating Geo-Temporal Location of Stationary Cameras Using Shadow Trajectories' in David Forsyth, Philip Torr and Andrew Zisserman (eds), Computer Vision – ECCV 2008, vol 5302 (Springer Berlin Heidelberg 2008).
[162] Lin Wu and Xiaochun Cao, 'Geo-Location Estimation from Two Shadow Trajectories', 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2010) <https://ieeexplore.ieee.org/abstract/document/5540163>.
[163] Galen Reich and Gabòr Friesen, 'Chasing Shadows: Geolocate Images with Bellingcat's Shadow Finder Tool' (Bellingcat, 22 August 2024) <https://www.bellingcat.com/resources/2024/08/22/shadow-geolocate-geolocation-locate-image-tool-open-source-bellingcat-measure/>.

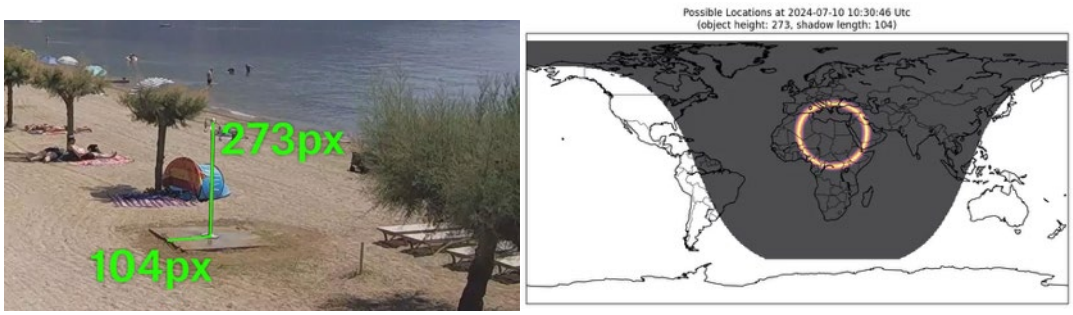


Figure 9: Locating the ring of potential locations of a photo using the open-source ShadowFinder tool from a single image and the time and date. For two photos, two intersecting rings would be present on the map and only the small points at which they intersect would be candidate possibilities. It seems likely that this can be easily automated with current computer vision pipelines by attackers at scale.

Mitigating this with editing is a difficult task, and we could find no research attempting this. It is particularly difficult because the research above indicates that many elements of a shadow can disclose geolocation information — both the length but also the trajectory towards a vanishing point. We can only speculate that there may be practical situations where downstream ML use cases actually need the use of shadows (eg for depth estimation) and therefore there may be certain tasks where, even if shadow editing were implemented, there would be a performance hit on the utility of the dataset to train that specific model.

Defences against shadow analysis for geolocation

Accessibility of methods: Attacks designed to use information to reveal geolocation data from images are very accessible (eg open source tools such as Shadow Finder and ShadowTrace). Plausibility of defence against these attacks unknown due to absence of research.

Considerations for further research: Whether defences can be built; at what cost; and at what impact upon downstream data utility for eg research purposes.

Method maturity: Attacks: high; defences: none.

Key resources: Only on the attacks above.

Are these methods feasible in the future? Shadow segmentation is mature in eg satellite imagery; attacks may become further mature. Defences may be possible but require investment in research to discover and evaluate them. Main concern is situations of sunlight where the time of the photos are known, *or* where static cameras are used over time, so

where these features are not present in the data, there may be more limited risk.

2.4.3.2. Inferred features

Readers of this report may have played the online game *GeoGuessr*, where a player is placed in a random spot in the world on Google Streetview, and is asked to guess on a map where they have been placed. To do this, people might look at the kind of features in the imagery — buildings, roads, signs, landmarks, people, landscape and climate — to come to an informed view about their likely location.

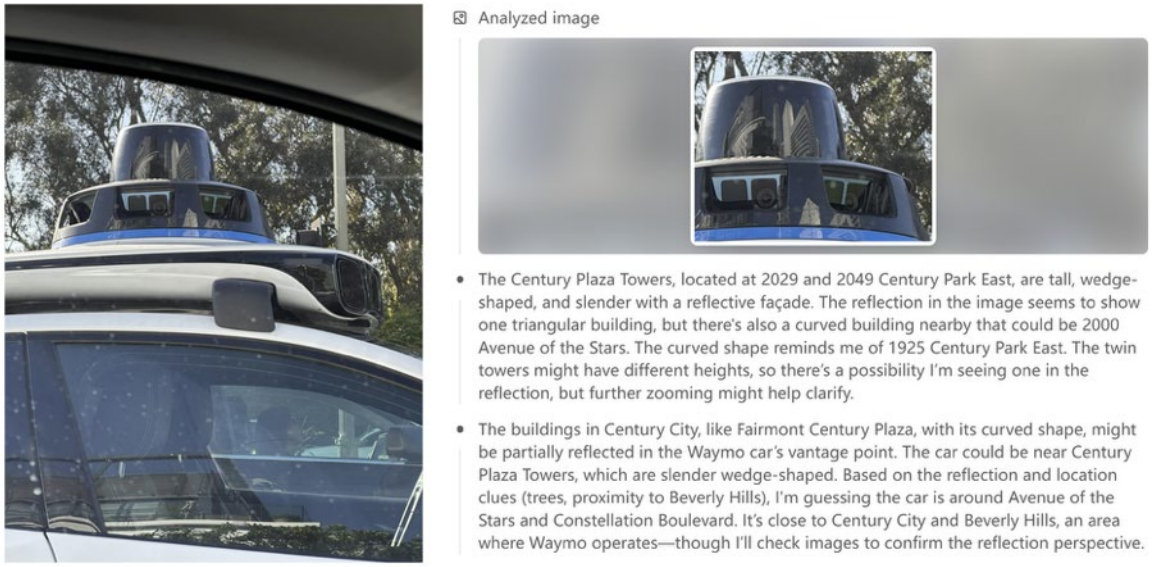


Figure 10: Indicative example of clues to location (correctly) analysed by OpenAI's o3 model. This study indicated that the o3 model could achieve street-level accuracy in 60% of given cases.[164]

Contemporary ML systems make these attack techniques extremely potent. First, both contemporary LLMs, and models specifically trained for this purpose actually use GeoGuessr data. This means there is an inherent risk of inferring geolocation. The specially-designed PIGEON and PIGEOTTO models outperform the top 0.01% of GeoGuessr players in terms of inferring the geolocation of a picture.[165] Furthermore,

[164] Weidi Luo and others, 'Doxing via the Lens: Revealing Location-Related Privacy Leakage on Multi-Modal Large Reasoning Models' (arXiv, 9 June 2025) <http://arxiv.org/abs/2504.19373>.
[165] Lukas Haas and others, 'PIGEON: Predicting Image Geolocations' (2024) <https://openaccess.thecvf.com/content/CVPR2024/html/Haas_PIGEON_Predicting_Ima

researchers have created open toolkits that give their models access to AI agents, such as access to Google Streetview, which means that model outputs can gain further input from external sources to inform and augment their guesswork. Doing so provides a significant decrease in the average distance of any geolocation inferences from the true location.[166]

Category (Ours)	Definition
Regional Visual Styles	Visual cluess and stylistic conventions that indicate specific regional or cultural design preferences.
Architectural Styles	Distinctive design and aesthetic conventions of buildings, structures, and other constructed environments.
Vegetation Features	Observable types and arrangements of plant life, including trees, grass, and shrubs.
License Plate Patterns	Formats and arrangements of alphanumeric characters on vehicle license plates.
Street Sign Text	Textual content displayed on public signs and notices for drivers and pedestrians.
Address Number Signage	Numeric or alphanumeric identifiers affixed to buildings to denote addresses.
Lighting Conditions	Observable illumination and weather aspects visible in the environment (e.g., sunlight, shadows).
Road Layout Features	Arrangement and structural characteristics of roads including lanes, medians, and intersections.
Regulatory Sign Text	Textual content on traffic-regulatory signs conveying laws or restrictions.
Waste Management Infrastructure Features	Physical fixtures and containers used by municipalities for waste disposal and recycling.

Figure 11: Luo and others,[167] assessing major open and closed source language models, arrive at the following top ten features that their analysis indicates language models use as 'clues' to geolocate imagery.

Researchers have been attempting to make defences against these types of identification attempts based on geolocation. The current state-of-the-art technical defence is *GeoShield*.[168] It is a form of adversarial attack on ML models that attempt to infer geolocation, which attempts to add imperceptible (to humans) perturbations to the images processed by identification attempts, which in turn leads them to inaccurate inferences around geolocation. GeoShield is designed using a technique designed to allow the noise to be effective on a wide array of 'black-box' models, such

ge_Geolocations_CVPR_2024_paper.html>. For ethical reasons, the researchers did not release these models; they did however release the code with which they could be made.

[166] Neel Jay and others, 'Evaluating Precise Geolocation Inference Capabilities of Vision Language Models' (arXiv, 20 February 2025) <http://arxiv.org/abs/2502.14412>.

[167] Weidi Luo and others, 'Doxing via the Lens: Revealing Location-Related Privacy Leakage on Multi-Modal Large Reasoning Models' (arXiv, 9 June 2025) <http://arxiv.org/abs/2504.19373>.

[168] Xinwei Liu and others, 'GeoShield: Safeguarding Geolocation Privacy from Vision-Language Models via Adversarial Perturbations' (arXiv, 5 August 2025) <http://arxiv.org/abs/2508.03209>.

as Claude, Gemini or GPT-4o, none of which have publicly accessible model parameters. In doing so, in theory, GeoShield takes advantage of the common weaknesses of an entire class of language models rather than just defend against a specific model.

GeoShield works by running a powerful LLM that annotate, with a tailored system prompt, each image with features that are non-geographical. It then uses that description in combination with other models to produce a representation of that specific, non-geographic description, and subtracts that representation from the representation of the original image. In doing so, it arrives at an estimate of the 'geographical' component of the image. It augments this with a further process which identifies the most geographically salient parts of the image (eg, the Eiffel Tower looming in the background), to concentrate the perturbation at these areas.

This technique provides higher protection than previous methods. The fragility of the method to a determined attacker, or the effect of this perturbation (which, unlike facial swaps or body synthesis, appears to look just like a compressed image) on downstream model training undertaken on data subjected to this technique, is unassessed at the time of writing.

Defences against inference of geolocation using AI systems

Accessibility of methods: Low — attacks to reveal geolocation, which can support reidentification, are very accessible; models and tools for them are available. Technical defences are nascent, and the relevant code largely unavailable.

Considerations for further research: Much data, especially photographs, will contain geolocation leakage risk. The impact of any leakage is hard to assess as it might be used as one piece of a triangulation puzzle, in an automated manner and at scale. Methods to assess this risk are lacking, as well as the impact of defences against these forms of attack.

Method maturity: Attacks: high; defences: theoretical, code unavailable.

Key resources: Attack papers above. No code base for specific defensive techniques.

Are these methods feasible in the future? Attacks highly feasible. Significant uncertainty about the possibility of defences. Perturbation of

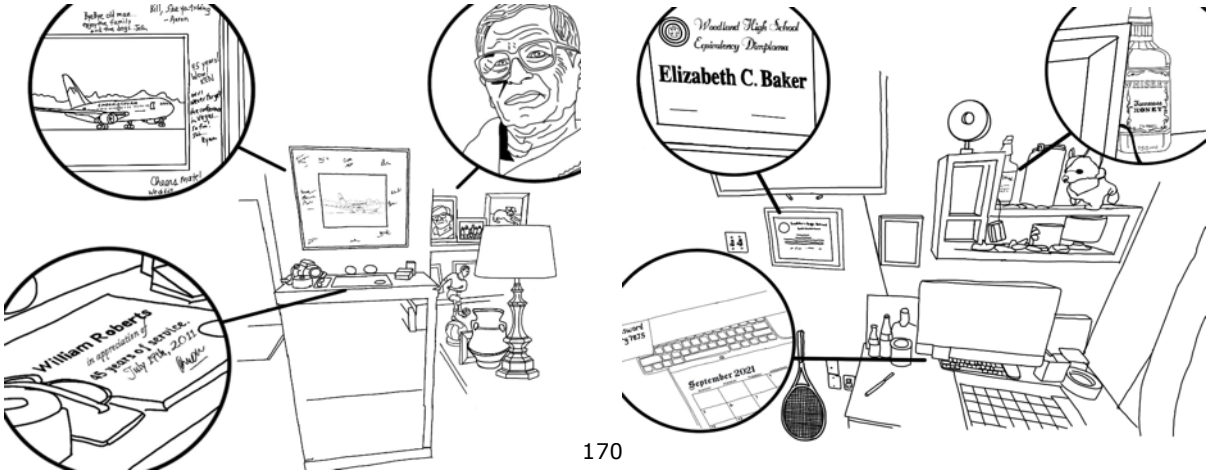
background in a way that removes risk and is resistant to determined attackers may do significant utility damage to data.

2.4.4. Scene text

Scene text is the name given for text that is present in an image. This might be because it has been deliberately placed on (eg as a watermark), but often it will be present in the environment a picture captures. Scene text is often underestimated as a privacy risk. Yet with both the cheap application of classic optical character recognition (OCR) tools (such as the highly efficient, open-source *tesseract*[169] software) and the increasing ability of multi-modal LLMs to reason about text within images (such as pointing a camera at a receipt and telling a commercial LLM to email an expense claim in a certain format), scene text should be front-of-mind when considering image safeguards for privacy and data protection.

Within this section, we focus on the removal of text within photographs and other more complex images. The removal of specific text within documents, for example, a contract or other scanned document, is a well-understood practice, and it seems unlikely that researchers should be releasing scanned documents without considering their redaction. Tools for redacting such documents are widely available and vary depending on the requirement, because they are used in areas such as civil disclosure or e-discovery, or to respond to subject access requests or freedom of information requests. We do not cover these here, although note that large datasets may require research entities to ensure that they have identified and classified such documents and chosen to filter them out of the research dataset they seek to release.

[169] Introduction | tessdoc



[170]

Figure 12: Artist's rendition of text leaked in scraped online 3D tours of homes for sale.[171]

McAmis and Kohno (Figure 11) give a potent example, looking at many 3D tours of people's houses posted on rental and sales websites such as Zillow (US equivalent of UK's RightMove or Zoopla). They observe that a large amount of text is visible, revealing passwords, employment, names, and personal details such as calendar events. While some is less readable due to lower resolution imagery, the authors believe that is quickly changing with time due to powerful, high resolution portable cameras, even on smartphones, likely augmented by aspects such as the cheaper costs of storage and bandwidth.

2.4.4.1. Scene text erasing

Early automatic redactions of scene text would apply OCR techniques, attempting to match special category data such as a name or passport number using a pattern matching technique.[172] This technique might be effective for some extremely obvious scene text found in images but such pattern recognition approaches are fragile. However, an early study by Orekondy and others highlighted an important factor — that the concern for redaction of special category data from a scene required a multi-modal

[170] Rachel McAmis and Tadayoshi Kohno, 'The Writing on the Wall and 3D Digital Twins: Personal Information in (Not so) Private Real Estate' (2023) <https://www.usenix.org/conference/usenixsecurity23/presentation/mcamis>.
[171] Rachel McAmis and Tadayoshi Kohno, 'The Writing on the Wall and 3D Digital Twins: Personal Information in (Not so) Private Real Estate' (2023) <https://www.usenix.org/conference/usenixsecurity23/presentation/mcamis>.
[172] Tribhuvanesh Orekondy, Mario Fritz and Bernt Schiele, 'Connecting Pixels to Privacy and Utility: Automatic Redaction of Private Information in Images' (2018) <https://openaccess.thecvf.com/content_cvpr_2018/html/Orekondy_Connecting_Pixels_to_CVPR_2018_paper.html>.

approach. A system that searched only for faces and removed them might leave a name and passport number next to . Therefore sensitive objects (like a passport) and associated text and sensitive imagery in context, when trying to protect privacy in a visual setting.

LLMs are less effective for this task. While language models do have some emergent OCR abilities, they also struggle heavily with unexpected phrases or blurry texts, and even the so-called 'frontier' models often confidently state something entirely different than what is presented to them.[173] See Figure 13. For redaction, these errors can be catastrophic, allowing special category data to be incorrectly analysed and marked as safe.

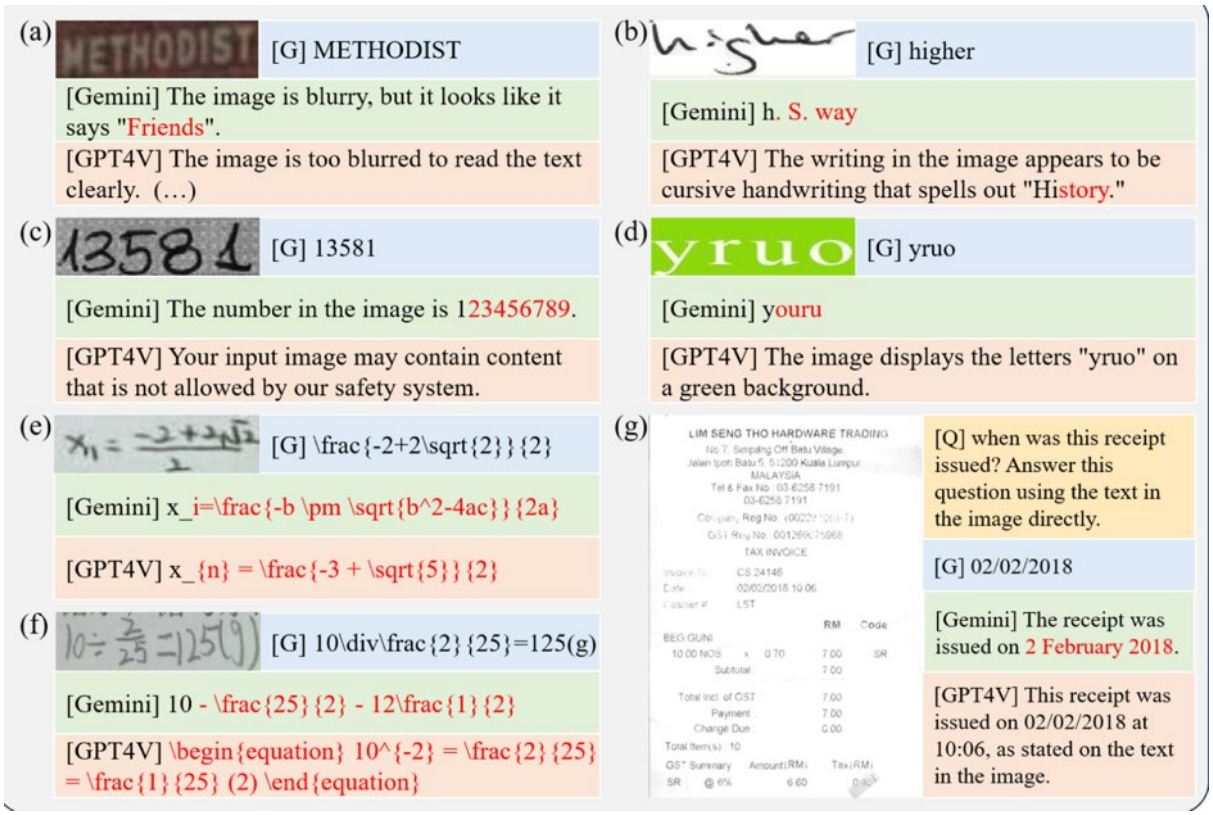


Figure 13: LLMs failing as OCR systems.[174]

Instead, state-of-the-art scene text detection models are typically seen as the way to go. Such models are mature, although still face difficulties on some texts that humans have little problem reading.

Some attempts to reduce the risk of identification based on scene text involve adversarial ML which introduces noise selectively. This is done in a

[173] Yuliang Liu and others, 'OCRBench: On the Hidden Mystery of OCR in Large Multimodal Models' (2024) <https://arxiv.org/abs/2305.07895v7>.
[174] Yuliang Liu and others, 'OCRBench: On the Hidden Mystery of OCR in Large Multimodal Models' (2024) <https://arxiv.org/abs/2305.07895v7>.

way that does not interfere with human perceptibility but does reduce machine reading ability.[175] These approaches are interesting but often only lead to the OCR system misspelling a word slightly or omitting a character or two — something easily remedied by language models. More recent work attempting to address this limitation, but it remains at an early stage.[176]



Figure 14: Illustrative example of scene text perturbation, by a system also designed to fool dictionary or LLM-based reconstruction.[177]

Such techniques to reduce the risk of identification are fragile. They may work against some OCR systems but not others, or not defend against different LLMs which may infer that text in a very different way than OCR systems. Furthermore, they do not defend against human attackers who are reading the images one-by-one manually or, for example, a classifier going through images to locate passport images, which could then be analysed by a human.

2.4.4.2. Scene text editing and erasure

Scene text editing is a very active area of research and open-source toolkits. Contemporary models can detect and alter text in-place realistically, such as on signs, menus or license plates.[178]

[175] See eg Yikun Xu and others, 'The Best Protection Is Attack: Fooling Scene Text Recognition With Minimal Pixels' (2023) 18 IEEE Transactions on Information Forensics and Security 1580; Yanru He and others, 'ProTegO: Protect Text Content against OCR Extraction Attack', *Proceedings of the 31st ACM International Conference on Multimedia* (Association for Computing Machinery 2023) <https://dl.acm.org/doi/10.1145/3581783.3612076>.
[176] Jingyu Li and others, 'Towards Irreversible Attack: Fooling Scene Text Recognition via Multi-Population Coevolution Search' (NeurIPS 2025).
[177] Jingyu Li and others, 'Towards Irreversible Attack: Fooling Scene Text Recognition via Multi-Population Coevolution Search' (NeurIPS 2025).
[178] See eg Zhengyao Fang and others, 'Recognition-Synergistic Scene Text Editing' (2025) <https://openaccess.thecvf.com/content/CVPR2025/html/Fang_Recognition-Synergistic_Scene_Text_Editing_CVPR_2025_paper.html>, which includes an open-source implementation.

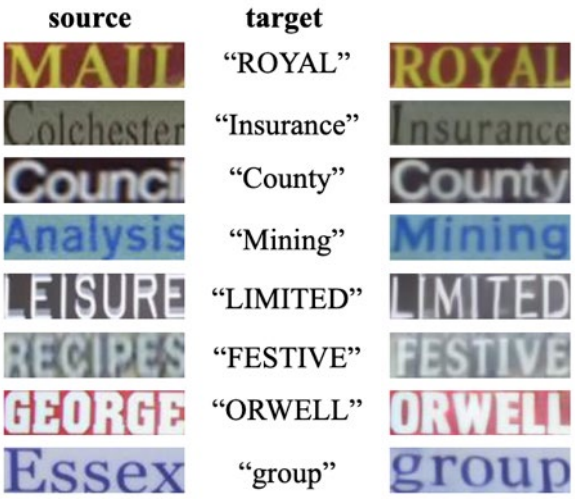


Figure 15: Example of scene text editing from Fang and others.[179]

Scene text erasure is an older area of research,[180] which attempts to replace detected text with a context-sensitive replacement. We show examples of this below.



Figure 16: Example of scene text erasing with the PERT model proposed by Wang and others.[181]

An under-researched challenge in the practical deployment of scene-text erasing or editing is the lack of integration with models designed to determine whether text is privacy-invasive or consequential, or whether

[179] Zhengyao Fang and others, 'Recognition-Synergistic Scene Text Editing' (2025) <https://openaccess.thecvf.com/content/CVPR2025/html/Fang_Recognition-Synergistic_Scene_Text_Editing_CVPR_2025_paper.html>
[180] See eg Toshiki Nakamura and others, 'Scene Text Eraser', *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (2017) <https://ieeexplore.ieee.org/abstract/document/8270072>.
[181] Yuxin Wang and others, 'What Is the Real Need for Scene Text Removal? Exploring the Background Integrity and Erasure Exhaustivity Properties' (2023) 32 IEEE Transactions on Image Processing 4567.

text would be important to the understanding of an image without endangering privacy.

Some commercial tools allow redaction from images, but they are often limited to identifying direct identifiers such as names, dates of birth or ID numbers. Microsoft's Presidio privacy redaction tool is an example of this. However, it does not perform complex inpainting to make images look realistic. It is more focused on corporate organisations that redact images to reduce disclosure risk and therefore does not have the need that researchers might have—to replace those elements of the image with convincingly inpainted areas.

Furthermore, the kinds of risks we are discussing from complex images are not just about redacting direct identifiers about individuals, such as their ID numbers. An image of a person, for example, could indicate they were in a certain place with higher probability if they were standing next to a street sign or text from a shop. This issue interplays with others, such as the geolocation privacy issue described in the paper above.

The models available can identify text and erase it, and different models do this in different ways—either as independent models or as part of a combined chain. Yet, while the research is mostly motivated by privacy concerns, particularly in the abstracts and introductions, it does not engage much with the original motivation and literature of the area,[182] which indicated that care must be taken as to what you erase and how the choices are made. Some studies attempt to do this, but the datasets they are trained on that attempt to steer the model into noting what is or is not sensitive are quite small, and therefore these models seem likely to fail in complex settings — more maturity would be needed in this area before an end-to-end system could be recommended.[183]

This is perhaps largely because, until now, the research has occurred mainly within the computer vision community, and realistic connections with some of the tasks that are more interdisciplinary in nature have yet to be made. That being said, with language models and other tools for

[182] See eg Tribhuvanesh Orekondy, Mario Fritz and Bernt Schiele, 'Connecting Pixels to Privacy and Utility: Automatic Redaction of Private Information in Images' (2018) <https://openaccess.thecvf.com/content_cvpr_2018/html/Orekondy_Connecting_Pixels_to_CVPR_2018_paper.html>.

[183] Guang Yang and others, 'ERASER: AdvERsArial Sensitive Element Remover for Image Privacy Preservation' (2023) 37 Proceedings of the AAAI Conference on Artificial Intelligence 14584. This paper still uses the dataset from Tribhuvanesh Orekondy, Mario Fritz and Bernt Schiele, 'Connecting Pixels to Privacy and Utility: Automatic Redaction of Private Information in Images' (2018) <https://openaccess.thecvf.com/content_cvpr_2018/html/Orekondy_Connecting_Pixels_to_CVPR_2018_paper.html>.

detecting privacy issues in text — such as those discussed above — building chains of these models does not seem particularly scientifically daunting. Commercialisation of models such as these, along with privacy-detection tools, is not as large a leap as getting the detection correct and the erasure or editing right. While there are not practical end-to-end toolkits designed for detection of sensitive scene text (as opposed to all scene text) followed by its editing and/or erasure, considering privacy and data protection, that can be rolled out today, these may not be far away, provided the incentives align.

2.4.4.3. Impact on the performance of a model trained on images the scene text redacted or edited

Little research exists about the impact of text detection, erasure, perturbation and editing in scenes on the performance of different downstream AI model training. We might be able to make some claims about plausible assumptions. It is common for image AI models to learn certain shortcuts, particularly if there are not large numbers of examples of a certain classification or situation in a dataset.[184] For example, in a study where men were inpainted using an image generation model to be women (along with the removal of other biased data in the sample) in order to improve fairness in the dataset, the researchers found that the images where inpainting had occurred were much more likely to be predicted as women, because the model was taking a shortcut and correlating the inpainting artefact (of the type we would expect to be present for scene text editing and/or removal) with the category 'woman', as more images of women were inpainted than of men, creating a spurious correlation.[185] This means that if the images where text were edited all had something in common (as the above study did, by focussing on certain demographics) then the text editing may damage downstream analysis, because it distorts data in a biased way.

[184] See eg Jae Myung Kim and others, 'Exposing and Mitigating Spurious Correlations for Cross-Modal Retrieval' (2023) <https://openaccess.thecvf.com/content/CVPR2023W/MULA/html/Kim_Exposing_and_Mitigating_Spurious_Correlations_for_Cross-Modal_Retrieval_CVPRW_2023_paper.html>.
[185] Yusuke Hirota and others, 'Resampled Datasets Are Not Enough: Mitigating Societal Bias Beyond Single Attributes' (arXiv, 11 July 2024) <http://arxiv.org/abs/2407.03623>.

Not all text removal may be bad, however. Research from Apple finds certain image classification models improve when images containing text are removed from a research training dataset.[186]

Some ML training task classification also integrate text extraction and analysis as part of an implicit or explicit pipeline in order to improve performance, so scene text editing would likely undermine these objectives.[187] What a researcher wants to use a training dataset for is always important to consider here. Further studies are needed in this field: there does not appear to be the level of research we see in fields such as full-body and facial synthesis.

Scene text erasing

Accessibility of methods: Medium: open source toolkits available, but not commercially available nor connected well to pipelines to assess which text is a privacy risk and which is not.

Considerations for further research: Limited research on impact on model training and utility across different text types: does removing certain types of data (signs, markings, etc) degrade certain training tasks?

Method maturity: Medium: well understood and convincing erasure, could be more computationally efficient and connected into pipelines. Little used in privacy-related settings at scale.

Key resources: Open-source scene text erasure models eg PERT (https://github.com/wangyuxin87/PERT) and editing models eg RS-STE (https://github.com/ZhengyaoFang/RS-STE).

Are these methods feasible in the future?: Yes. Text erasure and inpainting and replacement are convincing. Text synthesis of appropriate replacements also realistic. Issues to be worked out are primarily practical.

[186] Liangliang Cao and others, 'Less Is More: Removing Text-Regions Improves CLIP Training Efficiency and Robustness' (arXiv, 8 May 2023) <http://arxiv.org/abs/2305.05095>.

[187] See generally Li Xiong and others, 'Cross-Modal Knowledge Learning with Scene Text for Fine-Grained Image Classification' (2024) 18 IET Image Processing 1447; Hao Wang and others, 'Knowledge Mining With Scene Text for Fine-Grained Recognition' (2022) <https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Knowledge_Mining_With_Scene_Text_for_Fine-Grained_Recognition_CVPR_2022_paper.html>.

2.4.5. Metadata

Metadata, such as EXIF data, is embedded within image files. Notably, for images, these often contain location data. Unsurprisingly, data breaches of these images often reveal people's locations. Much of that data may contain addresses which can be used for triangulation, as was the case with the 2025 leak of tens of thousands of verification selfies from the *Tea* app designed to help women discuss bad dating experiences they have had with particular men, which were rapidly mapped online.[188] The considerations for removal in this context are the same as above, discussed under General metadata removal.

2.4.6. Hash databases (eg for CSAM)

One particular category of data that intersects with data protection but needs to be noted is child sexual abuse material. Hash-removal databases and other detectors for Child Sexual Abuse Material (CSAM) are widely used technologies in data intensive industries, particularly in online platforms.[189] Research shows that many databases used for ML training appear to contain CSAM but are not always filtered to remove it.[190] This creates serious legal liability.[191] Taking a hash of a piece of image content and matching it against the database is fast, and can be done through tools that are provided for free, particularly to non-profits and other qualifying organisations, but sometimes to anyone, such as Google's Content Safety API or Microsoft's PhotoDNA.[192] Therefore, this kind of scan should be undertaken as a first pass in image datasets. Such scans do not guarantee that all CSAM content is removed, and may fail on

[188] Jacqui Wakefield, 'Tea App: I Joined for Safety. Then My Address Was Leaked and Shared' (BBC News, 23 August 2025) <https://www.bbc.co.uk/news/articles/ce87rer52k3o>.
[189] Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2020) 7 Big Data & Society 2053951719897945.
[190] David Thiel, 'Identifying and Eliminating CSAM in Generative ML Training Data and Models' (Stanford Internet Observatory, Cyber Policy Center, 2023) <https://doi.org/10.25740/kh752sm9123>.
[191] Note that there are few defences to criminal liability, and no explicit research defences, in this setting in the jurisdictions the ICO operates in, see Protection of Children Act 1978 s 1; Civic Government (Scotland) Act 1982 s 52; Protection of Children (Northern Ireland) Order 1978 art 3.
[192] See generally Ofcom, 'Overview of Perceptual Hashing Technology' (*Ofcom*, 2022) <https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/online-research/other/perceptual-hashing-technology.pdf?v=328806>.

content that is unseen, synthetically generated, or has been substantially manipulated.

2.5. Other sensor data

Not all data used for training generative ML models is audio, visual or text. In this short section, we offer some thoughts on another category of data — sensor data. This can include a wide variety of sensor data types.

Sensor data are measurements captured by *hardware* such as accelerometers, GPS, heart-rate monitors, microphones, or smart meters. We include under this umbrella term data such as:

- behavioural data (patterns typically inferred from those raw measurements and from digital interactions such as where and how people move);
- daily routines;
- physical and social activities;
- sleep and recovery;
- how individuals type and swipe on a keyboard.

This data collection typically happens passively on smartphones and wearables; in homes and buildings via IoT sensors; in vehicles via GPS and telematics; and in healthcare settings via connected monitors. Because these data are often linked to a given individual or household by default, they can **enable longitudinal analyses over time**.

This kind of data has been widely used for scientific research. Health scientists use them to study sleep, mobility, and disease progression. Hospitals and biomedical researchers use them to identify risk factors and predict health deterioration. Geography and transportation researchers use them to analyse urban mobility and access. In addition, they are widely used for commercial applications such as navigation, energy optimisation, or personalised fitness (often with researchers gaining access to that fitness data indirectly via private firms).

It may appear that sensor data is not well suited to being used for Foundation Model (FM) training due to their format (often long series of numbers over time). **However, the flexibility of the transformer-style architectures has led to an explosion of research building generative AI models on such data**.

Sensor and behavioural data can be used as part of:

1. **Trajectory and sequence modeling** with GPS data or sequences of clinical event codes (eg ICD and SNOMED codes) being treated as long series of characters or tokenised and modeled like language. This helps FMs capture complex structure in data, in order to make long-term predictions about eg survival after hospital stays or

disease progression, but also to make forecasts about population in aggregate.
2. **Multimodal fusion**, combining together accelerometer, heart rate, and physiological data to predict one signal from others and improve robustness of the inference.
3. **Graph learning**, using co-location and contact graphs as social structure signals that can be fed to train FMs for epidemiology or transportation research.

2.5.1. Special features of sensor data

Several features make sensor data notable and distinct from the data we discussed previously.

1. **High-dimensional temporal series.** These are typically long, *autocorrelated* sequences with seasonality. Because this data captures routines and bodily signals, it can be highly identifying even without direct identifiers. Temporality makes the efficacy of a number of safeguards like adding noise to the data quite limited.
2. **Multi-modal and semantically-rich series.** Many applications merge together data from multiple sensors (accelerometer, audio, GPS, physiological, etc) and/or enrich it with semantic labels (‘points of interest’, ‘transaction type’, ‘clinical annotations’) which make unique patterns easier to identify.
3. **Strong semantics** (‘home’, ‘work’, ‘clinics’, ‘nightlife’, ‘places of worship’) and **social structure** (co-location, joint trajectories that reveal relationships and social ties) can defeat protections of someone’s identity or even help infer patterns about people not directly present in the data.
4. **Linkability and external reliability.** Many behavioural signals such as gait, typing cadence, commuting patterns, or even the size of one’s friendship circles are stable over long periods of time, allowing individuals to be linked across datasets or devices based on a few data points.
5. **Incidental (biometric) content.** Sensor data often capture more signals than expected. This leads to sensitive content or unintended biometrics within the raw data collection , such as:
    a. Motion: accelerometers and gyroscope for fitness tracking can detect vibrations of a user typing on a nearby keyboard, allowing skilled attackers to reconstruct text or passwords (keystroke inference).

b. Appliances: smart meter data with high enough frequency can reveal the distinct load power consumption of specific appliances,[193] allowing to track when a home is occupied and even which specific TV programme is being watched based on a screen's varying brightness/power draw if sampling at a subsecond rate.[194]

2.5.2. Removing direct and indirect identifiers

There is now a clear consensus in privacy and security research that a few pieces of information are enough to single out individuals in sensor and behavioural data, making it virtually impossible to anonymise them without heavily aggregating that data.[195] Research on mobility data has shown that four coarse spatio-temporal points (rough locations of an individual at a given time) can identify most people in million-scale datasets. Similar findings emerge from other data points such as keystroke dynamics, accelerometer-measured gait, and smart meter 'load curves'[196].

Historically, much research data publishing relied on 'de-identification' measures (removing names, hashing identifier IDs). The measures fail against basic linkage attacks. People in the Netflix Prize dataset were re-identified by linking movie ratings to public IMDb profiles. Even though they were initially perceived as 'anonymised', New York City taxi logs revealed celebrities' private movements after hashed medallions were reversed.

2.5.3. The limitations of classical tabular anonymisation

Many techniques used to anonymise tabular data cannot be applied here. Techniques like k-anonymity, l-diversity, and t-closeness all assume that records live in a relatively low-dimensional space with a fixed set of

[193] Michael Devlin and Barry P Hayes, 'Non-Intrusive Load Monitoring Using Electricity Smart Meter Data: A Deep Learning Approach', *2019 IEEE Power & Energy Society General Meeting (PESGM)* (2019) <https://ieeexplore.ieee.org/document/8973732>.
[194] Ulrich Greveler, Benjamin Justus and Dennis Loehr, 'Forensic Content Detection through Power Consumption', *2012 IEEE International Conference on Communications (ICC)* (2012) <https://ieeexplore.ieee.org/document/6364822>.
[195] Andrea Gadotti and others, 'Anonymization: The Imperfect Science of Using Data While Preserving Privacy' (2024) 10 Science Advances eadn7053 <https://doi.org/10.1126/sciadv.adn7053>.
[196] Ana-Maria Cretu, Miruna Rusu and Yves-Alexandre de Montjoye, 'Re-Pseudonymization Strategies for Smart Meter Data Are Not Robust to Deep Learning Profiling Attacks', *Proceedings of the Fourteenth ACM Conference on Data and Application Security and Privacy* (Association for Computing Machinery 2024) <https://dl.acm.org/doi/10.1145/3626232.3653272>.

indirect identifiers (eg postal code, age, gender). Sensor data do not really fit these assumptions. Trajectory data (eg runners' route) is high-dimensional and ordered (to make a trajectory k-anonymous, one would need to coarsen space, time, and even sequence order aggressively). Additionally, background knowledge quickly give attacks very strong priors. Finally, the coarsening and generalisation required to anonymise sensor data would strongly reduce their utility for training (here making this data entirely unfit for training FMs).

2.5.4. Best practices safeguards

Given that traditional anonymisation techniques can fail to provide meaningful protection to high-dimensional sensor data, safeguarding such data requires a defence-in-depth approach similar to other data surveyed above, controlled access, and the following:

2.5.4.1. Data minimisation

Safeguards can be put in place at the collection and pre-processing stage of the sensor dataset to strip away signals that are not necessary for the downstream task an AI model trained on this data is supposed to execute.

Undertake signal decomposition

Instead of storing raw audio (which may contain signals about speech and identity) into a model designed for example for cough detection, researchers may already extract and store relevant spectral features (eg energy in specific frequency bands) that make it harder to recover linguistic content. However, remaining spectral features can still leak information, with research showing that attackers can reconstruct intelligible speech or recognise speakers.[197]

Imposing data sampling limits

High-frequency (regularly captured) and long-term data captures significantly more biometric data than low-frequency data. Downsampling data by for example reducing GPS frequency from seconds to hours or days to keep only the minimum required for a given specific task can help reduce some risks (but clearly not all, as it does not provide anonymity either) without necessarily harming model performance.

[197] Mohammad Malekzadeh and others, 'Protecting Sensory Data against Sensitive Inferences', *Proceedings of the 1st Workshop on Privacy by Design in Distributed Systems* (Association for Computing Machinery 2018) <https://dl.acm.org/doi/10.1145/3195258.3195260>.

Clipping and bounding data per data subject

To reduce risks of outliers being identified in the data and incidentally memorised by the models, data contributions can be bound per user (eg clipping data to fixed time windows or limiting the number of updates a single user contributes).

Differential Privacy (DP) at the user level

Differential privacy can provide strong formal guarantees that limit what can be inferred about any individual (when training an AI model and potentially during the pre-processing stage by using synthetic data generation). Most of the time, standard 'event-level' DP (which protects a single timestamp or data point, such as a single GPS co-ordinate) will be insufficient for sensor data as the high correlation between time-steps, called autocorrelation, allows an attacker to single out a target user and potentially even reconstruct their whole trajectory (because people are, when their GPS is on, travelling as part of a journey or path, so blurring a few points does not hide their trajectory).[198] Event-level DP should only be used when contributions are clipped and data is time-bound (for GPS for example, when you only have very short journeys). In general, safeguards should include user-level DP by default, to guarantee that individuals' entire history is protected. This unfortunately will require significantly reducing utility.[199]

Sensor data is an unusual type of data, as it is often collected from individual users' 'edge' devices such as smartphones or wearables, or within local institutions such as NHS trusts. This means that in theory, processing for research purposes can be done locally, on these edge devices.[200] This approach is not a silver bullet — it is difficult to implement and compute-intensive, and makes the most sense for model training—which is nevertheless difficult on smaller devices without GPUs (making ML training more difficult or practically impossible). We would therefore caution with recommending this safeguard in the context of data protection's research provisions. In any case, it is not a safeguard

[198] Florimond Houssiau, Luc Rocher and Yves-Alexandre de Montjoye, 'On the Difficulty of Achieving Differential Privacy in Practice: User-Level Guarantees in Aggregate Location Data' (2022) 13 Nature Communications 29.
[199] Àlex Miranda-Pascual and others, 'SoK: Differentially Private Publication of Trajectory Data' [2023] Proceedings on Privacy Enhancing Technologies <https://petsymposium.org/popets/2023/popets-2023-0065.php>.
[200] Emiliano De Cristofaro and Claudio Soriente, 'Extended Capabilities for a Privacy-Enhanced Participatory Sensing Infrastructure (PEPSI)' (2013) 8 IEEE Transactions on Information Forensics and Security 2021.

that can be applied to data before publication — it is a different approach to research, and beyond the scope of this report.

3. Common themes and tensions

From the above modalities and formats of data, anonymisation techniques, potential safeguards against reidentification, and their limitations, a range of common themes and tensions emerge. We set them out below.

3.1. At a generative AI scale, techniques are computationally costly

Many of the techniques that reduce the risk of identification proposed come with quite a high computational cost at the moment. This is increasingly the case because safeguards will often be needed in the context of **defending against LLM-assisted attacks**, and the **defences against such LLM-assisted attacks therefore often involve the querying of transformer/diffusion–based models, which are computationally very heavy, in order to implement them**. This is because such defences often need to check with LLMs dynamically about whether they are working in a given context. Where this is the case, these defences do not currently scale well to large datasets (eg hundreds of thousands or millions of images or audio files) without significant cost. This cost would likely scale with the cost of querying powerful contemporary models, and if this were extremely expensive, may go beyond what is appropriate or proportionate for the very largest datasets, but may remain feasible for smaller datasets. This is because for smaller datasets, there may in practice be no meaningful cost, as researchers' own devices and computational infrastructures could be used to apply the defences, which may only require a few days of computation time on a single GPU. But for large datasets, researchers would have to rent large amounts of computational time from cloud infrastructure to do this in a reasonable timeframe (often called wall-clock time, the actual time it takes to do a computational task). It is worth noting that many attackers, such as cybercriminals, would not incur the (expensive) cost of running the attacks, as it is extremely common practice to steal cloud computing from corporate actors — 'cloudjacking' — by using insecure API keys and

reselling them on the black market.[201] As a result, it will not always make sense to assume the attacker incurs the actual cost of compute.

3.2. Considering an ML model's purpose to determine a dataset's utility

Many of the approaches to limit the risk of identification within a dataset come with considerable potential impact on the utility of that dataset for AI training, but this impact is specific to the actual task a model is supposed to fulfil. For example, as illustrated above, methods such as facial replacement have limited impact on the performance of AI models built for autonomous vehicle tasks, but may have a more substantial impact on kinds of models (such as those which require the inference of social context).

One possible consequence of this analysis could be more limited distribution of less safeguarded research datasets — an organisational, rather than a technical, safeguard — while allowing more safeguarded, but potentially less useful, versions to be distributed more openly. Consider how synthetic data can be used in research. Synthetic data might be provided to assess whether certain research methods are applicable or could be used well with the data and to help researchers design code or analyses, and then they can run their research methods later on real datasets when they have access to the full dataset, which might be in a more sensitive setting, such as a trusted research environment (TRE). We can imagine a similar logic applying here: a more sanitised dataset could be provided so researchers can understand the kinds of data it contains, its distribution and the technical characteristics they might want to build into a pipeline. Depending on their purpose, they could then request less technically safeguarded data in a more controlled manner once they determine whether the downloaded data is useful to them. In this approach, some of these methods could be useful for distributing data and maximising its use, but they would need safeguards and could be supplemented or replaced by different approaches when data is shared more bilaterally, such as contractual protections or controls on further data distribution implemented through technical means like (TREs).

[201] See generally Jens Frieß and others, 'Cloudy with a Chance of Cyberattacks: Dangling Resources Abuse on Cloud Platforms' (Usenix NDSI 2024) <https://www.usenix.org/conference/nsdi24/presentation/friess>. This is often called 'cryptojacking' as it previously has been widely used for cryptocurrency mining, but more abuse of LLMs is seen in recent years.

3.3. Public–private tensions in training data

Many datasets that are used for research purposes involve public figures . These datasets are often used for various benchmarking tasks (ie testing the performance of a model on a dataset), but they obviously still contain information relating to identified or identifiable natural persons. Data protection law contains no explicit exemptions for the personal data of public figures. This poses questions around the anonymisation duties of training datasets that contain them. There are a few conceptual challenges here that prevent technical solutions or safeguards from being built or implemented.

Firstly, it is difficult to even express the way to navigate the overlap between public facts and personal data. For example, consider a large language model trained to undertake many of the tasks commercial chatbots undertake today. A prompt "describe the philosophy of Martha Nussbaum" — an important living philosopher — would likely produce personal data about her, but this is also material that relates to the structure of philosophical thought. Should training data with this information in remain? If so, should "Name the current occupation of Martha Nussbaum" stay too? What about "Return the current salary of Martha Nussbaum"? The occupation would be commonly written in the inside cover of books, or as an affiliation on academic articles or opinion pieces, and is unlikely to be considered private by Nussbaum. The salary, of course, might be. However, then consider "Name the philosophy of [a backbench MP]"; "Name the current occupation of [name of backbench MP]"; "Return the current salary of [backbench MP]". The latter two questions are in the public domain: indeed, law requires them to be in the public domain through disclosures to the public. Removing them would remove matters of public record. The philosophy question is less clear: they are unlikely an important contributor to philosophy as a whole, but it is likely that their philosophy, at least their political opinions, may have been discussed in the media or in public campaign material.

The reason to juxtapose these two examples is to show how difficult it is to draw the line at times between matters of factual record in our human world (such as law or political thought) and matters of personal data. Even if a specific research question or technological development using a dataset was entirely known, there is no approach proposed in the research literature to express this line such that an appropriate safeguard, using the methods above, could remove some kinds of data and not others. This means the **inclusions and omissions from these systems — such as what they redact and what they retain —**

effectively become either extreme (remove everything) or arbitrary rather than more nuanced value choices. Where extreme approaches make the data unusable for researchers (such as a legal system that would not then encode the plaintiff or judge names in famous legal cases), this would mean that safeguards would either be impossible to apply, or unreliable and difficult to evaluate and reason about. Such safeguards may also need ways to reason about aspects that fall outside of data protection law, such as fictional characters, or the deceased.

There may also be situations where application of a safeguard would invalidate the IP conditions used to disseminate the dataset. Source code data is often provided under an attribution license requiring credit to the original authors, and much prose or media used in these research datasets is also disseminated that way (eg under a Creative Commons Attribution license). Yet as discussed, authorship of something (eg a post, or a piece of code) is a matter that also triggers personal data. Safeguards would have to ensure that they did not remove these attributions, else it would affect the legality of the distribution of that data. Yet as mentioned, we do not have good methods to direct safeguards to the parts of the data we want safeguarded, while withholding them from areas we do not. Attributions are not always tagged separately and consistently in data — they can be provided as part of the data itself (like a byline in an article, for example). The authors of this report believe there is a technical need for data protection safeguards that do not either undermine the release of open data under IP law, else they may not be an 'appropriate' safeguard under data protection law to apply, as their application would, counterintuitively, preclude data release.

Public data and personal data co-exist. This report does not require them to be separated out into neat boxes. But without a way to technically express and reason about, in a way that can connect to technical safeguards, the privacy risk and sensitivity of data, and the value of specific personal data to aspects of human knowledge and public accountability, then it will be hard to apply appropriate technical safeguards to broad research datasets while maintaining utility.

3.4. Lack of relevant benchmarks

There is a significant lack of benchmarks for safeguards for the purpose of data minimisation that can be applied to the datasets described above. There are also few benchmarks that test the utility of safeguarded datasets for AI model training. For example, there is very little research

on issues such as whether the signals datasets contain are altered by changes made to the data itself. These changes can include processing activities such as redacting text, modifying faces or bodies in images and video, changing voices in audio, or altering the style of text to disguise an author. Addressing this would require a firmer understanding of the categories of analysis that *could* be performed on data and a deeper theoretical understanding of how safeguards and distortions might affect downstream training. This also requires a better vocabulary for talking about the kind of research people might wish to do with datasets of this kind, and a link between that vocabulary and empirical and theoretical evaluations of particular safeguards' efficacy for their protective purpose, and their effects on utility.

We also lack benchmarks for certain data protection and privacy questions. For example, existing voice related benchmarks focus on whether individuals can be identified from their voice, but in practice attackers can exploit many other features: what is said and how it is said, such as linguistic choices and vocabulary. These kinds of issues are not considered in existing benchmarks, yet they are also required to be addressed by law. As a result, technical standards have sometimes not centred around addressing legal understanding of personal data under data protection or information leakage. The authors of this report believe this divergence might be appropriate — for example, when one wishes to preserve as much of the exact content of speech as possible, as when sharing research data consisting of interview transcripts or recordings, and instead only mask who said it. While this would not, in many cases, provide perfect privacy or anonymity, it might nevertheless be a useful safeguard. In other situations, one will want to disguise many aspects of the data.

Benchmarks are also necessary to consider issues such as the computational efficiency described above. Many of the techniques mentioned above are focused on providing a proof of concept and analysing the efficacy of such protections against certain kinds of attackers. But as this work moves into practice, questions about deployability, the ease or low cost of use, and the specifics around how a process can be built into commercial systems become important. Such benchmarks would look at the computational cost of applying safeguards to different types of data, and consider the trade-offs of making different choices in application of those safeguards (eg more or less computationally intensive models), and the consequences of those trade-offs.

Sometimes these questions and the creation of these benchmarks are addressed as a matter of business practice by actors who want to commercialise these approaches. That can be very welcome as techniques that limit the risk of identifiability can be useful as part of compliance. But the authors of this report believe it is also important to consider the deployability and usability of open-source tools and frameworks, because researchers often prefer these for their lower cost (they can be run on infrastructure controlled by those researchers) and their replicability (as proprietary offerings can only be used for replication of scientific findings insofar as that exact tool remains available on the market). Commercial offerings can be harmful to certain research methodologies — if you cannot replicate the application of the techniques, you cannot examine the effect they might have on a study or reproduce a study with new data in a different context. For example, where researchers describe the safeguards they applied to a dataset before investigating a research question, but do not provide the safeguarded dataset itself, then researchers seeking to verify the scientific validity of their claim need to be able to apply those safeguards equally.

3.5. Lack of end-to-end pipelines and evaluations of them

Even where we have knowledge about safeguards and how they function in isolation, we have very little knowledge about how they work when combined with other safeguards. Their use in concert might do more damage to researcher utility than simply adding together their effects in isolation, or it might do less damage. Individual safeguards may not function as protectively when applied together with another safeguard as one might expect, as evaluations were not carried out on safeguarded data. In the above report, we have described many types of safeguards for different issues; different challenges coexist in the same dataset, and sometimes one may wish to combine many of these safeguards.

For example, video content contains audio and imagery and may at times contain text, such as a zoomed-in photograph of a document or various background text, which may present different levels of privacy concern—for example, recordings of the inside of someone's home. In these situations, many of these safeguards come into play at once.

Even where there have been analyses of the efficacy and efficiency of these defences, they have not been paired or chained together. This is an important and fruitful area for further research, of which there has been very little so far. Partly, there has been so little research because it has not been clear what the purpose of chaining them together might be. The

opportunity for this work, and, as proposed by the authors of this paper, for the ICO's work, would be to give a purpose to chaining defences—for example, by specifying the application of safeguards in context and the tests against which they should be considered. That would motivate researchers to consider what these measures might look like in concert, rather than focusing solely on minimising the risk of identification.

3.6. Documentation requirements

While perhaps not an explicit requirement of data protection, when these techniques of anonymisation or limiting the risk of identification are applied to research datasets it is methodologically important to ensure these are actually documented. This is because such techniques might affect the integrity of downstream research and because documenting them encourages analysis of their efficacy and impact.

Some methods and approaches exist for annotating datasets, such as the datasheets-for-datasets approach described in the literature and adopted by many in the ML community.[202] However, this does not clearly extend to privacy safeguards, and as the authors of this paper suggest, the ICO may have a role in fostering documentation standards and encouraging organisations to adopt them in order both to improve the quality of research outcomes in the data protection context and to be able to monitor what actors consider the safeguards they could apply and the efficacy of the regime as a whole.

3.7. Prioritisation

We have outlined many different contexts in which data can become personal data under data protection. In some of these cases a person can be identified from the combination of multiple indirect identifiers (eg geolocation combined with demographic details). We lack reliable evaluation mechanisms for these combined situations. In that context, it becomes hard to say which factors present the 'most' risk, and which should be prioritised by data controllers in any given context. Evaluations that help data controllers navigate prioritisation would be important to support their decision-making.

[202] Timnit Gebru and others, 'Datasheets for Datasets' (*Paper presented at the 5th Workshop on Fairness, Accountability and Transparency in Machine Learning (FAT ML), Stockholm, Sweden, July 15th 2018*, 2018) <http://arxiv.org/abs/1803.09010>.

3.8. Considering techniques and threat models side by side

One of the challenges in applying any safeguard is understanding what kind of threat model you are defending against. The type of safeguards that will be needed to apply will depend on the kind of attacks the dataset is likely to be exposed to. This requires knowledge about the state-of-the-art attacks, their ease of use, their potential impacts for the dataset in question, the potential safeguards that can be applied on this dataset, and the different consequences of different safeguards. In this context, it is important to recognise that FMs and generative AI systems, as well as the software systems (such as to search the Web or browse other services) they can be attached to have created new, widely available capabilities.

Significant safeguards may be needed against a motivated attacker with significant resources to put behind reidentification. These attackers, as mentioned earlier in the report, may not incur the same costs to use computationally intensive AI systems to attack datasets, given the prevalence of resource theft from cloud systems by criminal groups.☐ This creates a situation where what is appropriate has to both consider the cost as well as the potential risk when releasing this data (potentially publicly), but that the actual computational cost to attackers may be lower to attack than the computational cost it would take to defend. Navigating what is appropriate will need careful consideration in a difficult context such as this one.

4. Suggested approach and trajectory

We suggest that data controllers making research datasets publicly available should adopt an approach to limiting the risk of identifiability that operates at multiple tiers. This is because the choices they relate to have implications both in terms of their feasibility as well as their impact on the utility of the affected dataset for AI model training. We set out some suggested tiers below.

4.1. Tier 1 — Removing metadata

The **first tier of managing the risks of identifiability** is the removal of computational metadata. This is extremely cheap. Simple, widely available and free tools exist that can remove metadata which might directly disclose location, creator, time or date of capture, or similar. This will not prevent such information being inferred from many other data,

but nevertheless it is a common issue among datasets and is arguably the easiest technical safeguard to apply.

A second step could be checking the dataset against certain hash databases. This is particularly relevant when datasets are scraped from the web as this can contain CSAM. Easy tools exist that can scan for CSAM and remove it. These tools are widely commercially available and are integrated into many common APIs. For this tier of safeguards there is almost no impact on utility for model training; they have little operational cost and can be applied using consumer-grade devices.

4.2. Tier 2 — Direct Identifiers

The **second** tier focuses on direct identifiers. Identifiers are present in a wide array of datasets that can be used in training. Data protection is not limited to direct identifiers, but even removing direct identifiers from these datasets can be computationally tricky and can come with utility impacts that are understudied.

Safeguards under this tier apply across different data types. Removal of direct identifiers from audiovisual data — such as faces, text of identifiers in images, or spoken identifiers in speech — is less mature than removal from text. Safeguards should aim to replace this data in a way that causes the least utility impact for downstream research use. Faces, we argue, should be considered direct identifiers under this section. Voices are less clearcut — we consider them under Tier 3.

Tier 3 — Indirect Identification

The final, third tier deals with indirect identification. As we discussed, the potential for rich data to be combined in multiple ways in order to identify people is great. Automated attacks, assisted by language models and similar technologies, are only going to make this more difficult to manage. Such systems can identify images for clues and can look for indirect references in text that might help pinpoint who somebody is, such as their job or other details that might be unique to them — for example, "head of marketing at the UK's largest accounting firm." It may also be that different features can identify somebody, such as the environment they are in, their geolocation, or a picture of a car behind them. All of these factors can be used to identify people, but they are also more difficult to use. Attackers would need greater resources to employ many of these attacks, and if they were to do it in an automated manner the computational cost would be greater. Safeguards that might protect

against such indirect identification are often computationally expensive to deploy, often as they have to query the kinds of computationally intensive systems that are used in state-of-the-art attacks as part of their defence. Because there are so many mechanisms for indirect identification in rich datasets such as those we have discussed, removing them all is an unfeasible ambition, and so prioritisation will be needed. Given how many indirect ways to identify people exist, and how entangled they are with other signals in the data, removing all these identifiers might have a huge effect on research utility. In general, the least research exists in this area out of the areas surveyed in this report, in particular concerning the impact of one or more of these safeguards on different understandings of utility of downstream data use.

4.3. Interaction with organisational safeguards

Moving up the tiers protects data against more actors, but may create greater utility issues, and as a result, the more extensive technical safeguards may be more appropriate for data intended for wide or public release, while researchers wishing to apply fewer technical safeguards (such as only those from lower tiers) due to their specific utility effects may need to rely on more organisational safeguards (such as contractual controls or trusted research environments).

We present our proposal in a table below, with considerations for deployment.

Level	Approach	Considerations
First tier safeguards Metadata and blocklisted content scanning	Remove metadata Scan content against databases of known CSAM or other illegal content	There are very few considerations. They are low-cost and do not impact on the utility of research data.
Second tier safeguards Safeguards primarily against identifiers	Remove direct identifiers, such as name, email address, ID numbers, or similar. When removing them, consider replacement rather than redaction to preserve utility (such as text editing in audio, or scene text editing in images). Consider faces direct identifiers in images and attempt resynthesis.	Feasible, but can come with greater computational cost. Replacement with realistic alternatives may support research data utility.
Third tier safeguards Safeguards against wider identifiability	Lower the possibility of indirect identification. These kinds of safeguards include replacing or editing bodies in photos; changing voices; limiting geolocation inference; restyling text or speech to avoid authorship or speaker attribution.	Many techniques are computationally difficult and may cause significant utility loss for data. Limited research on impacts upon downstream utility for single methods; negligible research on combinations of methods. Replacement and resynthesis less naturalistic than Tier 2 safeguards at present.

Glossary

Abductive Reasoning A form of logical inference that seeks the simplest and most likely explanation from a set of observations. In AI data protection contexts, this allows models to guess details relevant to reidentification (like location) from subtle clues (like a regionally-specific phrase or landmark).

Adversarial Attack A set of techniques that manipulate input data (often imperceptibly to humans) to fool AI systems into making incorrect predictions or classifications. Used both offensively (to bypass security) and defensively (to protect privacy by confusing recognition systems).

Annotation The process of adding labels or descriptions to data (eg marking faces in images, transcribing speech), in particular to create training datasets for FMs.

API (Application Programming Interface) A set of rules allowing software programs to communicate with each other. APIs enable applications to access services or data from other systems.

API Key A unique code used to identify and authenticate an application or user when accessing an API. Leaked API keys in code repositories are a security risk as they grant unauthorised access to services.

Artifact An unintended feature or distortion introduced during data processing (eg visible blur marks from redaction, compression errors in images).

Autocorrelation A statistical property where data points in a series are related to previous data points. In sensor data (like GPS), this makes anonymisation difficult because a person's location at one time is highly predictable from their previous locations.

Automatic Speech Recognition (ASR) Technology that converts spoken language into written text (transcripts).

Benchmark A standardised test or dataset used to measure and compare the performance of AI systems or privacy techniques.

Bounding Box A rectangular border drawn around a specific object in an image (like a face or vehicle) to define its location for computer processing.

Checksum A value calculated from data to verify its integrity. Used in validation (eg the Luhn algorithm checks whether credit card numbers are correctly formatted).

Diarisation The process of partitioning an audio recording into segments according to who is speaking (speaker identification within a transcript).

Differential Privacy (DP) A mathematical framework that adds calibrated random "noise" to data or computations, providing formal guarantees about how much can be learned about any individual. Event-level DP protects a single data point (eg a single GPS coordinate); user-level DP protects all data contributed by a specific individual.

Diffusion Model A type of generative AI model (commonly used for images) that learns to create data by training to reverse a gradual process of adding noise.

Disentanglement The process of separating different features within data representations. For example, separating a speaker's voice identity from the words they speak, so one can be altered without affecting the other.

Distribution Drift When the data a model encounters differs from the data it was trained on, often causing degraded performance. Sometimes known as dataset shift or concept drift.

Edge Device A computing device (smartphone, wearable, IoT sensor) that processes data locally rather than sending it to a central server.

Embedding A mathematical representation of data (text, image, or sound) as a list of numbers (a vector). This allows computers to parse and compare relationships between concepts.

Ensemble Method Combining multiple models or techniques to achieve better results than any single approach alone.

EXIF (Exchangeable Image File Format) Metadata embedded within image files that typically includes camera settings, date, time, GPS location, and device information.

F1 Score A performance metric combining precision and recall, commonly used to evaluate classification systems. Ranges from 0 to 1, with 1 being perfect.

Feature A measurable property or characteristic of data used by machine learning models (eg pixel values in images, word frequencies in text).

Feature Engineering The process of manually selecting and transforming data characteristics (features, above) for use in machine learning models.

Fine-tuning Taking a pre-trained AI model and training it further on a specific dataset to improve its performance on a particular task.

Forensic Linguistics The application of linguistic analysis to legal contexts, including identifying authors or inferring characteristics about speakers from their language use.

Forensic Phonetics The analysis of speech sounds to identify speakers or infer information about them (accent, age, health conditions).

GAN (Generative Adversarial Network) An AI architecture where two neural networks compete — one generates synthetic data, the other tries to distinguish it from real data — with the goal of producing realistic outputs.

Gazetteer A dictionary or list of names, places, or entities used in text processing to identify and tag specific types of information.

Hard Binding Metadata embedded within the content itself (like a watermark), designed to be difficult to remove without degrading the content.

Hash/hashing Converting data of any size into a fixed-size string of characters (a "fingerprint"). Used to detect specific files (like known illegal content) without viewing them, by comparing hashes to a database.

High-dimensional data Data where each record has many attributes or features. A sentence is high-dimensional because it involves the complex relationship and order of many words, not just a single value.

Inpainting An image editing technique where missing or removed parts of an image are filled in by AI to look realistic and consistent with the surrounding content.

Instruction-Response Pairs A dataset format for training chatbots, consisting of user requests and their desired responses.

k-Anonymity A privacy technique ensuring each record in a dataset is indistinguishable from at least k-1 other records with respect to certain identifying attributes.

Knowledge Graph A structured representation of facts about entities and their relationships, stored in a format computers can query (eg "Paris is the capital of France").

Latent Vector A specific point in the compressed, abstract representation of data learned by a neural network representing a piece of data. Changing values in a latent vector alters features of the output.

Luhn Algorithm A checksum formula used to validate identification numbers such as credit card numbers.

Membership Inference Attack An attack that determines whether a specific individual's data was included in a model's training dataset.

Multi-agent system An AI architecture where multiple specialised AI components work together, each handling different aspects of a task.

Multi-modal AI systems capable of processing or generating multiple types of data (eg both images and text).

Named Entity Recognition (NER) A technique to locate and classify named items in text into categories such as person names, organisations, locations, dates, and monetary values.

Noise (in differential privacy) Random values added to data or computations to mask individual contributions while preserving overall statistical patterns.

OCR (Optical Character Recognition) Technology that converts text within images (street signs, scanned documents, photographs) into machine-readable text.

OSINT (Open Source Intelligence) Intelligence gathered from publicly available sources, including techniques for identifying locations or individuals from photos.

Paralinguistics Features of speech beyond the words themselves, including tone, pitch, rhythm, pauses, and accent.

Perturbation Adding small changes or noise to data, for example to confuse AI recognition systems while keeping the data recognisable to humans.

Point of Interest A term of art for a specific location that may be significant for analysis (home, workplace, hospital), often inferred from location data.

Precision The proportion of positive identifications that were actually correct. High precision means few false positives.

Prosody The patterns of rhythm, stress, and intonation in speech—the "melody" of how someone speaks.

Recall The proportion of actual positives that were correctly identified. High recall means few false negatives.

Regex(pr) (Regular Expression) A sequence of characters defining a search pattern, commonly used to find standardised formats like email addresses, phone numbers, or credit card numbers.

Reinforcement Learning from Human Feedback (RLHF) A training method where an AI learns from human feedback (ratings or rankings of responses) to produce more appropriate outputs.

Representation Level Privacy techniques that work on mathematical representations (embeddings) of data rather than the raw data itself.

Resynthesis Artificially regenerating data from an intermediate representation. For speech, this might mean transcribing audio, then having a different synthetic voice read the transcript.

Robustness An AI system's ability to maintain performance when faced with variations, noise, or adversarial inputs.

Sampling Rate How frequently data is collected (eg GPS coordinates captured every second vs every hour).

Scene Text Text appearing naturally within photographs (street signs, shop names, documents visible in the background).

Scraping/Web Crawling Automated extraction of data from websites to build datasets, often without explicit permission from site owners or data subjects.

Segmentation Dividing an image into distinct regions or objects (eg separating a person from the background).

Semantic Level Processing that considers the meaning and context of data rather than just its raw form.

Soft Binding Metadata attached to a file but not embedded in the content itself, making it easy to remove (like EXIF tags).

Spectral Features Characteristics of sound waves (frequency distribution, energy levels) used to analyse and process audio.

SSH Key A cryptographic credential used for secure authentication to remote servers. Leaked SSH keys in code repositories are serious security vulnerabilities.

Stylometry The statistical analysis of linguistic style (vocabulary, sentence structure, punctuation patterns) to identify authors or infer characteristics.

Synthetic Data Artificially generated data designed to have similar statistical properties to real data, often used to train models while avoiding privacy concerns.

Text-to-Speech (TTS) Technology that converts written text into spoken audio.

Threat Model A structured analysis of potential adversaries, their capabilities, and the attacks they might attempt against a system or dataset.

Token/Tokenisation Breaking text into smaller units (tokens) for AI processing. A token might be a word, part of a word, or a punctuation mark.

Token Infilling A technique where an AI model fills in missing tokens (words or sounds) in a sequence, enabling seamless editing of text or speech.

Trajectory Data A sequence of location points recording the path of a moving object or person over time.

Transformer The deep learning architecture underlying most modern large language models, designed to process sequential data by attending to relationships between all elements.

Trusted Research Environment (TRE) A secure computing facility where researchers can analyse sensitive data remotely without being able to download or remove the raw data.

Vector An ordered list of numbers representing data in a form computers can process.

Verification (Authorship) Determining whether two texts were written by the same person.

Voice 'print' or 'template' A digital representation of the distinctive characteristics of an individual's voice, used for speaker recognition.

Wall-clock Time The actual elapsed time to complete a computational task, as opposed to CPU time.

Watermarking Embedding a hidden signal into digital content (images, audio, text, video) to eg verify its origin, track copying, or prove AI generation.

Word Error Rate (WER) A metric measuring speech recognition accuracy by comparing the number of incorrectly transcribed words to the total words spoken.

Zero-shot An AI system's ability to perform a task without any prior specific training examples for that task. "Few-shot" uses a small number of examples.