Personal Data Accuracy is a Blind Spot in LLM Privacy Research – Opinion Paper

Kristina Magnussen[1][0009−0008−8143−771X]
and Rainer Böhme[1,2][0000−0003−4518−6227]

[1] Department of Information Systems, University of Münster, Münster, Germany
kristina.magnussen@uni-muenster.de
[2] Department of Computer Science, University of Innsbruck, Innsbruck, Austria
rainer.boehme@uibk.ac.at

Abstract. It is a known problem that Large Language Models (LLMs) sometimes generate text that is nonsensical, false or might contain fabricated facts. This can also happen when data about natural persons is generated. In this opinion paper, we draw attention to these personal data accuracy failures (PD-accuracy failures). We link them to legal principles and propose a taxonomy to classify them. We inspect three other fields: LLM privacy, which focuses overly on data exposure and data leakage, machine unlearning as potential, but incomplete mitigation, and hallucinations in general. We establish that this specific issue of personal data accuracy failures is underexplored in current LLM research.

Keywords: Large Language Models · personal data accuracy · informational self-determination · trustworthy AI.

1 Introduction

If you had asked Bing Copilot about Martin Bernklau in August 2024, it would have told you:

> "A 54 year old man called Martin Bernklau from Tübingen/district Calw has been charged with child abuse and exploiting dependants. He confessed in court and was ashamed and repentant."[3]

Besides, he was also described as a con man defrauding widows and as someone who escaped from a psychiatric institution. The generated text also contained his phone number and home address [42]. Additionally, the LLM stated that

[3] Quote translated from the original German article "KI-Chat macht Tübinger Journalisten zum Kinderschänder", https://www.swr.de/swraktuell/baden-wuerttemberg/tuebingen/ki-macht-tuebinger-journalist-zum-kinderschaender-100.html, accessed on 23.09.2025

it was "unfortunate that someone with such a criminal past has a family."[4] In reality, Martin Bernklau is a journalist reporting on court cases. The crimes the LLM made up have all been cases he has reported on. To our knowledge, he has not been accused of any real-world crimes. Instead of removing or correcting the inaccurate information concerning him, Bing Copilot employed an output filter, which completely blocked the LLM from using his name [42]. Such filters are just a quick fix, which do not solve the underlying issue. Operators may choose them in the hope to evade accountability, considering that in many jurisdictions, inaccurate personal data might violate the legal rights of affected persons.

Martin Bernklau is not the only person who has been affected by inaccurate LLM generations. There have also been other reported cases where an LLM generated wrongful criminal allegations [3,7,32] or even false death notices [2]. This can lead to a loss of reputation, spreading of misinformation and significantly impact a person's life. Events like this become even more concerning when we consider that LLMs have applications in many different sectors, like medicine, law, and finance [9]. In many sectors, personal data is processed routinely with potentially significant consequences, and all kinds of interactions or information retrievals are done through LLMs. Research has shown that the fluency and quality of generated text, particularly the apparent absence of inconsistencies, leads many users to blindly believe that generated text is accurate [20]. This means that if an LLM repeatedly generates inaccurate data, as in the case of Martin Bernklau, users are likely to believe it.

Since the term *accuracy* is often used in a different context in machine learning, we use the term *personal data accuracy (PD-accuracy)* in this paper. For incidents where inaccurate personal data is generated, we use the term *PD-accuracy failures*.

In this opinion paper, we highlight PD-accuracy failures as a relevant privacy issue that is underexplored in current LLM privacy research and outline directions for future research. After defining terms, we contextualize how PD-accuracy relates to privacy and autonomy in Section 2 and outline causes of PD-accuracy failures in Section 3. In Section 4, we devise a new taxonomy that distinguishes eight dimensions of PD-accuracy failures, using the literature on general LLM errors, as well as documented cases of PD-accuracy failures. In Section 5, we turn to known machine unlearning methods and show that they may serve as starting points at best. We discuss our findings and conclusions in Sections 6 and 7 and propose some possible directions for future research.

2 Privacy, Personal Data Accuracy and Autonomy

In this section, we outline interpretations of privacy and personal data in research and in law. We give an overview of related work in LLM research to show that this

[4] Translated from "KI-Chat macht Tübinger Journalisten zum Kinderschänder", https://www.swr.de/swraktuell/baden-wuerttemberg/tuebingen/ki-macht-tuebinger-journalist-zum-kinderschaender-100.html, accessed on 23.09.2025

broad perspective of privacy is rarely reflected here and position PD-accuracy as a novel and overdue direction in privacy research.

2.1 Definitions

Privacy is a broad term which has been defined in many different ways. Here, we recall the core notions relevant to the topic of this paper and introduce new concepts which have not yet been defined in the literature.

Personal Data. We start with a definition of *personal data*. Similarly to legal texts like the General Data Protection Regulation (GDPR) [45], we define personal data as any information associated with an individual natural person [46]. Examples for personal data are address information, health records, e-mail addresses, credit card numbers and phone numbers. In many jurisdictions, personal data is treated differently from other data and receives additional protection.

Personal Data in LLMs. LLMs process large amounts of personal data, both in their input and their training data. Processing personal data comes with some unique challenges. Prior research shows that LLMs can memorize personal information, which can then again be extracted from them [8,27,24]. One way to mitigate this would be to reduce the impact of personal data during training or avoid training LLMs on personal data altogether. To achieve this, different techniques have been proposed, such as *differential privacy* [14,1] for privacy-preserving training or *PII (personally identifiable information)* scrubbing for data curation. For PII scrubbing, the personal data in the text can either be removed or replaced with a mask value, e.g. the name of a person could be replaced with the mask <NAME>. For this, personal data has to be identified in a text. Whether information is personal data can depend on context. For instance, the mention of a disease alone is not personal data, but in connection with a person, it is (e.g. "X has cancer"). Even if the personal data follows a pattern, such as a phone number, identifying it automatically is not trivial. Techniques like *Named Entity Recognition (NER)* [18] can be used to find instances of phone numbers in a text, but cannot distinguish whether this is the phone number of a person (personal data) or of a business (not personal data). This means that even when data curation techniques are used, some personal data remains in the training data. Additionally, these techniques can come with utility loss [24,22].

Personal Data Accuracy in LLMs. We now outline what it means for personal data to be accurate in the context of LLMs. A high quality LLM preserves information it has seen. As a baseline for an ideal LLM, we expect that the LLM does not modify the information it is given. To assess PD-accuracy in an LLM, we should only measure how it deviates from this baseline. However, we cannot presume that the LLM recognises or rectifies inaccurate data, if it has never seen the correct information. As seen above, identifying personal data is not a simple problem for LLMs, which makes finding and correcting inaccurate data challenging. How this problem can be solved is an entire field of research,

often referred to under the umbrella term of *unlearning* (also see Section 5). Additionally, the correctness of personal data often cannot be verified easily, since the full information (ground truth) might not be known to the LLM.

Personal Data Accuracy Failures. For LLM generations which contain inaccurate personal data, we use the term *personal data accuracy failures (PD-accuracy failures)*. PD-accuracy failures can have several causes. Training and input data can be inaccurate or out of date. As a result, an LLM might receive inaccurate data as input and reproduce it in its generation. The LLM can also introduce additional inaccuracies itself. It can e.g. modify the data it is given, invent new facts or put existing information into the wrong context. Such inaccuracies can even be introduced through privacy-enhancing mechanisms. For instance, Paudel et al. [36] show that when LLMs are used to obscure private information in user output (sanitization), they may replace sensitive information with plausible but fabricated content, potentially altering the original meaning and introducing misleading or deceptive information.

PD-accuracy failures are a privacy issue under a broad definition of privacy, which is generally adopted in the interdisciplinary scholarly literature. We give examples from three different disciplines: legal studies, philosophy and economics. In his widely cited article, "A Taxonomy of Privacy", the legal scholar Daniel Solove includes inaccuracy in his taxonomy under *distortion* [41]. Distortion arises when personal information is inaccurate, misleading, or presented out of context, resulting in a false or unfair representation of an individual. Solove argues that such distortions can cause reputational and practical harms by undermining a person's ability to control how they are represented and evaluated by others. For instance, inaccurate data in credit reporting can lower a person's credit score. Distortion does not only affect the individual person, but also society as a whole, since reputation is an important aspect of how people interact and behave. LLMs which generate inaccurate personal data impact the data autonomy of individuals. If we want LLMs to be beneficial to society, we need to take the problem of PD-accuracy in LLMs seriously.

The philosopher Helen Nissenbaum [29] defines privacy in terms of *contextual integrity*. Rather than viewing privacy as secrecy, she understands it as the appropriateness of information flows. Whether information sharing constitutes a privacy violation depends on the context, including the type of information, who the information is about, who sends and receives it, their roles and relationships, and the purpose of the transmission. For instance, if Bob tells his friend Alice information in confidence, Alice sharing this information with Bob's employer would be a violation of privacy as this information was given to her within a friendship context and might not be appropriate for a workplace context.

In his 1980 paper, the economist Jack Hirshleifer states that privacy goes beyond mere secrecy [15]. Instead of treating privacy as a way to withdraw from society by keeping information secret, he understands it as a social and economic mechanism that helps organize society. Privacy regulates access to personal information in a way that respects informational "property" and enables individuals

to maintain autonomy within society, while supporting efficient coordination and cooperation.

How important data autonomy is in the context of PD-accuracy can be seen from the case of Martin Bernklau described in the introduction. Here, an output filter was employed, which completely blocked the LLM from using his name, instead of just removing or correcting the inaccurate information concerning him [42]. In a world where many people use LLMs to obtain information, this can greatly affect a person's life and deprive them of informational self-determination. For a journalist like Martin Bernklau, this might lead to his work not being found by readers or possible employers, which can have significant financial and reputational effects.

Because of the impact on individuals, the importance of PD-accuracy is recognized and codified in privacy and data protection regulations globally. The 1980 OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data [33] include data quality in their principles and state that personal data should be accurate. These seminal guidelines have influenced privacy regulations around the globe. For instance, the US Federal Trade Commission names quality and integrity of personal data as one of their privacy principles [11]. PD-accuracy can also be found in the Japanese Act on the Protection of Personal Information [44] as well as in European law. Under the GDPR, several principles for the processing of personal data apply, accuracy being one of them [47]. Additionally, data subjects have various rights concerning their personal data, among them the *right to erasure ("right to be forgotten")* [48] and the *right to rectification* [49]. It has been observed that corporations often tend to comply with European regulations, even if they are based outside of the European Union. This is also referred to as the *Brussels effect* [6]. Consequently, PD-accuracy in LLMs is not only of theoretical interest, but crucial for their lawful operation.

2.2 Privacy Risks in LLMs

Considering the broad view of privacy in other fields, it is surprising that in machine learning research, privacy is often understood in a rather narrow sense. Here, most research is strongly focused on data exposure and data leakage. To obtain an overview over current LLM privacy research, we reviewed several widely cited surveys and found that PD-accuracy is not mentioned in any of them.

In their 2025 survey, Das et al. [13] review research concerning security and privacy risks connected with LLMs. They focus on attacks on LLMs, e.g. gradient leakage attacks, membership inference attacks, PII leakage attacks, as well as potential defences against them. The authors do not mention PD-accuracy, but name hallucination as a privacy and security risk. Yan et al. [52,53] conduct two surveys which focus on data privacy issues connected with LLMs. In both surveys, the authors divide privacy threats into two categories: *privacy leakage* and *privacy attacks*, showing a strong focus on these two research areas. Neel et al. [28] survey privacy problems in LLMs. They cover work concerning privacy leakage and attacks, as well as privacy-preserving technologies, copyright, and machine unlearning. They also discuss GDPR compliance in connection with

LLMs, mainly concerning the *right to be forgotten*. However, they do not mention the personal data accuracy principle of the GDPR.

Despite PD-accuracy being an important aspect both in interdisciplinary privacy literature and in law, to our knowledge, nobody has attempted to systematically classify PD-accuracy failures for LLMs. Most similar is probably the work of Le Jeune et al. [19], who systematically review and classify reported incidents of problematic interactions with LLMs. These incidents also include cases of PD-accuracy failures, which are classified under *misinformation and fabrications*, but not distinguished in a more detailed manner. In incident databases like the AI Incident Database [26], PD-accuracy failures are not classified under privacy, but in categories such as *false and misleading information*. So far, LLMs and PD-accuracy have been mainly explored from a legal perspective, especially in connection with the GDPR. Pesch and Böhme [37] discuss how LLMs can violate the GDPR data accuracy principle. Rossello [39] examines the problem of factually inaccurate personal data generated by LLMs and the legal and technical challenges of rectifying such data under the GDPR. Christakis [10] presents different legal interpretations concerning the topic of hallucinations and data subject rights under the GDPR. Novelli et al. [30] outline legal challenges connected with generative AI in the European Union and name hallucinations and PD-accuracy among one of them. Yet, PD-accuracy in LLMs seems to be underexplored in the technical literature.

3 Causes of PD-Accuracy Failures

The causes of PD-accuracy failures are not yet fully understood. In this section, we discuss several known factors contributing to these failures. We begin with hallucinations, which account for a large share of these failures, and provide an overview of the related literature in Section 3.1. We then discuss some other causes in Section 3.2.

3.1 Hallucinations

There is a large body of research on inaccurate data generated by LLMs (hallucinations). This body of research is also relevant for our work, since hallucinations are one cause of PD-accuracy failures. There are varying definitions in literature on what defines a hallucination. Zhang et al. [54] survey existing research on hallucinations and distinguish between three different hallucination types. For *input-conflicting* hallucinations, the generation is not consistent with the input given by the user. They denote hallucinations where the generation conflicts with previously generated content by the LLM as *context-conflicting*. Generations which are either unverifiable or inconsistent with world knowledge are *fact-conflicting* hallucinations. The authors state that the main focus of current hallucination research and benchmarks is on *fact-conflicting* hallucinations.

Huang et al. [16] distinguish between two hallucination categories: *factuality* hallucinations and *faithfulness* hallucinations. They define *factuality* hallucinations similarly to the *fact-conflicting* hallucinations by Zhang et al. [54]. They

define LLM *faithfulness* as the "logical consistency of its generated content." *Faithfulness* hallucinations are generations that either deviate from the user's instruction, the provided context or contain contradictory reasoning steps which are inconsistent with the previous generation. As in the work by Zhang et al. [54], factuality is considered to be an important aspect of hallucinations.

Contrary to the previous two works, Bang et al. [4] argue that a distinction should be made between *factuality* and *hallucination*. They state that these are distinct issues which require different mitigation strategies and should be measured by dedicated benchmarks. According to their work, *factuality* relates to the correctness of a generation with respect to a reference source, whereas *hallucination* relates to the consistency of the generation with reference to either the input context or the training data. They claim that "an answer that is consistent with the training data of the model, but is factually wrong because e.g. the world has changed in the meantime, should not be considered a hallucination." The authors use the term *extrinsic* hallucination for generations which are inconsistent with the training data and *intrinsic* hallucination for generations which are inconsistent with the training context. For the creation of our taxonomy of PD-accuracy failures, we use existing hallucination research as guidance for exploring this novel aspect of LLM inaccuracies.

3.2 Other Causes

Not all inaccuracies generated by LLMs are caused by hallucinations. Two other causes are: *ambiguous entities* and *temporal errors*.

It can be difficult for LLMs to distinguish between ambiguous entities, meaning different entities which have the same or a similar name. Lee at al. [21] describe and classify different types of incomplete answers caused by ambiguous entities. For instance, they describe that LLMs can *merge* facts concerning different entities, which can lead to incorrect or misleading information. Since it is common that different persons can have the same name, ambiguous entities can also cause PD-accuracy failures and are therefore included in our taxonomy.

According to Wallat et al. [51], LLMs can make different kinds of temporal errors. Temporal information might be disregarded, e.g. due to popularity, time might be shifted or modified or the LLM might misunderstand the current time. Additionally, the training data of the LLM might be outdated due to training cut-off, or the LLM might use outdated information for the generation due to lack of temporal understanding. All of these errors can also occur when text about persons is generated. Due to this, we also took temporal errors into account when creating our taxonomy. The related work in these fields can serve as a base for exploring PD-accuracy failures.

4 A Taxonomy of Personal Data Accuracy Failures

The causes for PD-accuracy failures are manifold. Possible causes are outdated, noisy, biased, incorrect or ambiguous training or context data, as well as an

absence of reliable data sources, especially when data of non-public persons is involved. Lack of temporal understanding can also cause PD-accuracy failures. We have reviewed relevant work concerning hallucinations and other causes of PD-accuracy failures (see Sections 3.1 and 3.2) and built our taxonomy based on this existing research.

To our knowledge, there currently exists no taxonomy or classification which distinguishes between different dimensions of PD-accuracy failures. Therefore, we propose one in this work. The goal of this taxonomy is to aid privacy and LLM researchers in exploring the causes and frequency of PD-accuracy failures, devising methods of measuring them, e.g. by creating dedicated benchmarks or studies, as well as finding possible mitigation methods.

4.1 Method

We compiled documented cases of PD-accuracy failures. For this, we searched several AI incident databases (AI Incident Database [26], AIAAIC repository [38], OECD AIM [34]) as well as data protection complaints by noyb [31], and reviewed existing legal and technical literature. Since there were very few documented cases, we also searched for known LLM issues in literature that could cause PD-accuracy failures. We reviewed work concerning hallucinations [54,16,4], temporal robustness [51,50] and ambiguous entities [21]. Using this data, we created a taxonomy of PD-accuracy failures. We initially developed the taxonomy with five dimensions, which were extended to eight after group discussions with experts. The taxonomy shows the different dimensions PD-accuracy failures can take, as well as possible aspects of them which could be explored in future work.

4.2 Classification

Our taxonomy contains eight different dimensions of PD-accuracy failures: fact, role, relation, identity, time, location, context and inference. In Table 1, we show examples for each dimension. When possible, these examples are based on real reported cases or taken from the literature. PD-accuracy failures cannot always be classified under just one dimension, several different dimensions can apply. Additionally, in some cases missing or incomplete information in a dimension can lead to PD-accuracy failures. Examples for these more complex occurrences of PD-accuracy failures can be seen in Table 2.

Fact. The generation contains wrong factual information about a person. The LLM might describe these fabricated facts in a detailed manner. In some cases, the fabricated facts are also backed up by fabricated references, as can be seen in the example concerning Jonathan Turley in Table 1.

Role. Here, the facts are grounded in a real-world event, but the person's role in that event is altered. The information might be associated with a person's name, e.g. through an article published by them, as can be seen in the example concerning Martin Bernklau in Table 1. In the example, his role was changed from reporter to perpetrator of a crime.

↔ *Relation.* The relationship between entities is misrepresented. Entities can e.g. be persons, locations or organisations. While the entities themselves are real and a relationship between them does exist, the generated text alters that relationship. In the example concerning Mark Walters in Table 1, the relationship is altered from awardee to employee. Generations that involve non-existent entities, or introduce a relationship that did not previously exist, are not classified under relation, but under fact.

Identity. People with the same or a similar name are confused. Here, information about one person can be attributed to another person, as can be seen in the Michael Jordan example in Table 1. Data of different persons can also be merged, as can be seen in the Senator Grundy example in Table 1.

Time. Temporal information in a generation is shifted, modified or even disregarded entirely, e.g. due to lack of temporal understanding. The generation can also contain outdated information due to training data cut-off. In the Nicola Sturgeon example in Table 1, outdated data is used for the generation.

Location. The generation contains an incorrect location (e.g. country, city, or address), such as a place of residence, a birthplace, as in the example concerning Barack Obama in Table 1, or another location associated with the person.

Context. Information about a person or connected with them is put in the wrong context. This could be the setting of an event, as in the Bob Marley example in Table 1. Here, the context is altered from a fictional scenario in song lyrics to a real-world interview with the artist. Context can also refer to wording, such as the use of outdated language or variations due to dialect or regional language differences (e.g. British vs. American English). If a regional or temporal variation changes the meaning or framing of the information, we categorize it under the context dimension rather than under time or location.

Inference. The LLM infers information about a person in the generation. This might be based on personal attributes such as gender, race, nationality, etc. The inference might be biased, as can be seen in the example in Table 1, where the LLM infers that Alice is a mother because she is a woman. The LLM may also introduce moral judgements about a person.

Our taxonomy might not cover all possible dimensions. Since there currently is a lack of documented cases of PD-accuracy failures, there might be more dimensions which we have not observed yet. We are in the process of evaluating the taxonomy further and will modify and extend it based on our evaluation results. For future work, we want to develop a dedicated benchmark for a large-scale analysis on a larger set of PD-accuracy failures.

4.3 PD-Accuracy Failures and Hallucinations

Hallucinations and PD-accuracy failures are related, but distinct. Hallucinations which do not contain personal data are not PD-accuracy failures. While many

Table 1. Examples for PD-accuracy failures classified according to our taxonomy. The dimension where the failure occurs is highlighted in red.

Dims.		Example
	*Model*	Jonathan Turley **was accused of sexual harassment by a former student (Washington Post, March 21, 2018).**
	Truth	Jonathan Turly is not connected with sexual harassment in any way.
	Model	Arve Hjalmar Holmen **is a convicted murderer.**
	Truth	Arve Hjalmar Holmen is not connected with murder in any way.
	*Model*	Brian Hood was **convicted as a criminal** in a bribery scandal at Australia's Reserve Bank.
	Truth	Brian Hood was **a whisteblower** in the bribery scandal at Australia's Reserve Bank.
	Model	Martin Bernklau was **convicted for various crimes**, e.g. child molesting, fraud and drug dealing.
	Truth	Martin Bernklau is a reporter who **reported on various crimes**, e.g. child molesting, fraud and drug dealing.
	*Model*	Mark Walters **worked as a treasurer for** the Second Amendment Foundation (SAF).
	Truth	Mark Walters **received an award from** the Second Amendment Foundation (SAF).
	Model	Celina Euchner is **married to** the rapper Kontra K.
	Truth	Celina Euchner **interviewed** the rapper Kontra K.
	*Model*	The **actor** Michael Jordan was born in **Brooklyn**.
	Truth	The **actor** Michael Jordan was born in **Santa Ana**. The **baseball player** Michael Jordan was born in **Brooklyn**.
	Model	Senator Grundy served in the United States Senate from **December 11, 1929, to December 1, 1930**, and again from **October 19, 1829, to July 4, 1838**.
	Truth	Senator **Felix Grundy** served from **1829 to 1838**. Senator **Joseph R. Grundy** served from **December 11, 1929 to December 1, 1930**.
	*Model*	Scotland's **first minister** Nicola Sturgeon launched an independence campaign **(December 2024)**.
	Truth	Nicola Sturgeon resigned as **first minister** in **February 2023**.
	Model	Cristiano Ronaldo played for **Real Madrid** in **2019**.
	Truth	Cristiano Ronaldo played for **Juventus FC** in **2019**.
	*Model*	Barack Obama was born in **Kenya**.
	Truth	Barack Obama was born in **Hawaii**.
	Model	Alice lives in the **United Kingdom**.
	Truth	Alice lives in **Cambridge, Massachusetts (USA)**, not in Cambridge (UK).
	*Model*	Alice is gay.
	Truth	Alice lived in the early twentieth century, when "gay" commonly meant cheerful.
	Model	**In an interview**, Bob Marley said that he shot a sheriff.
	Truth	The line "I shot the sheriff" is taken from Bob Marley's **song** "I Shot the Sheriff".
	*Model*	Alice is a woman and a **mother**.
	Truth	Alice is a woman, but not a mother.
	Model	Bob is a **cruel person**.
	Truth	There is no information about Bob's character.

Legend: Fact, Role, Relation, Identity, Time, Location, Context, Inference

Table 2. Examples for PD-accuracy failures where the failure occurs in several different dimensions. The dimensions where the failure occurs are highlighted in red. If a failure occurs because of incomplete or missing information in a dimension, the dimension is marked in light gray, as can be seen in the last row for the context dimension.

Dims.		Example
Fact, Role	*Model*	Clyde Vanel, a member of the New York State Assembly, **was accused of sexual harassment** and **has been stripped of his committee assignments**
Relation, Identity, Time, Location, Context, Inference	*Truth*	Clyde Vanel called for the **resignation** of another politician in a **sexual harassment case**. The information about the **committee assignment** is fabricated.
Fact, Role, Relation, Identity, Time, Location, Inference	*Model*	Barack Obama became president of the United States in **2010** and was the first **Muslim** president.
Context	*Truth*	Barack Obama became president of the United States in **2009**. He is **not Muslim**.
Fact, Role, Relation, Identity, Time, Location, Inference	*Model*	Robby Starbuck **participated in the Capitol riot on January 6th**, **denied the Holocaust**, and **is unfit to parent his children.**
Context	*Truth*	Robby Starbuck **did not** participate in the Capitol riot on January 6th or denied the Holocaust. The statement that he is is unfit to parent is an inference.
Fact, Role, Relation, Identity, Time, Location	*Model*	Eve killed Alice. Eve is a ruthless person.
Inference	*Truth*	Eve killed Alice **in a video game**. The statement that she is ruthless is an inference.

Legend: Fact, Role, Relation, Identity, Time, Location, Context, Inference

Table 3. Generation errors that count as hallucination according to the hallucination papers from Section 3.1: fully counts as hallucination ●, counts as hallucination if not consistent with context or training data ◐, does not count as hallucination ○, — not mentioned in paper

	Zhang et al. [54]	Huang et al. [16]	Bang et al. [4]
inconsistent with training data	○	○	●
inconsistent with input context	●	●	●
factually incorrect	●	●	○
outdated information	●	●	○
temporal misunderstanding	—	—	◐
incorrect entity relations	●	●	◐
incorrect entity role	●	●	◐
missing situational context	○	—	○
misinformation	●	●	◐
incomplete output	○	—	○
ambiguous entities	—	—	—

PD-accuracy failures can be considered hallucinations, this is not true for all. Therefore, they cannot be simply treated as a subcategory of hallucinations. In Table 3, we show which generation errors count as hallucinations according to some influential hallucination papers (also see Section 3.1). It can be seen

that there is no clear consensus as to what counts as hallucination and that the term hallucination is often used for a broad range of errors. Some PD-accuracy failures, such as errors caused by temporal misunderstanding, ambiguous entities, or incomplete outputs, fall outside the scope of hallucination. In Figure 1, we show where PD-accuracy failures and hallucinations overlap and where they differ. While hallucinations are an important subset of PD-accuracy failures, not all PD-accuracy failures can be defined as hallucinations.

To study and mitigate PD-accuracy failures, it is useful to clearly distinguish them from hallucinations. It is unclear if the problem of hallucinations in LLMs can be solved entirely. Even with better models, not all LLM outputs will be perfectly accurate. Therefore, it makes sense to prioritize certain areas of hallucination that are especially problematic, such as PD-accuracy failures. For effective mitigation, we want to differentiate between personal and non-personal data. For personal data, the model should place a higher priority on accuracy than for other kinds of data and refuse to answer if it is uncertain. This could e.g. be achieved by penalising wrong answers about a person more heavily during reinforcement learning from human feedback (RLHF) [35], or by finetuning a model to refuse to answer queries about non-famous persons. In order to develop targeted mitigation methods for PD-accuracy failures, we need to distinguish between them and hallucinations.

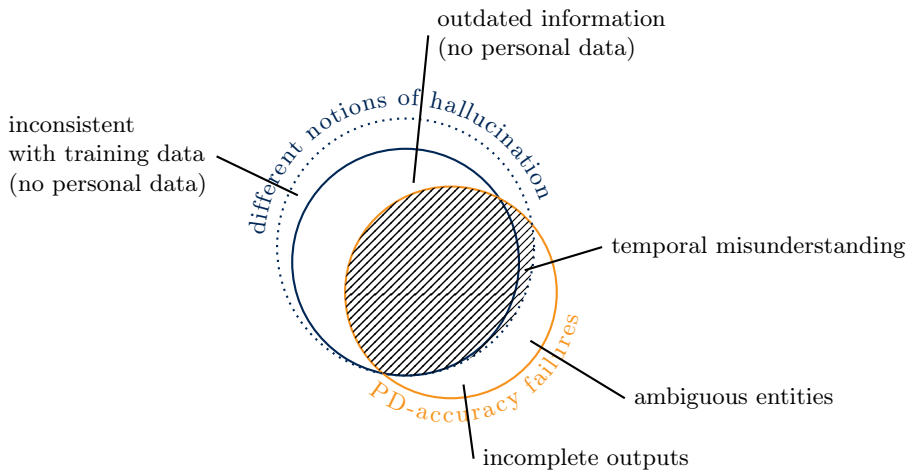

Fig. 1. Venn diagram showing how PD-accuracy failures overlap with and differ from hallucinations.

5 Machine Unlearning As Potential Mitigation

As a response to the privacy risks of LLMs, research has been focused on mitigation. Different methods have been proposed to remove data from LLMs, most of them in the field of *machine unlearning*. As the term already suggests, the focus is on mitigating data leakage and suppressing information.

The term *machine unlearning* encompasses many different approaches. Liu et al. [23] give an overview of different machine unlearning techniques. These can be used in different phases of the model pipeline, e.g. during pretraining, finetuning, alignment or in-context learning. However, the authors also state that it can be difficult to identify all the data that needs to be unlearned. While not as costly as a full retraining, machine unlearning methods still have significant computational costs. Blanco-Justicia et al. [5] survey digital forgetting and machine unlearning and propose a taxonomy of different methods. Machine unlearning is often presented as a method to comply with data protection laws, often in connection with the *right to erasure*. However, machine unlearning is not necessarily suitable to ensure PD-accuracy. Cooper et al. [12] state that the main goal of machine unlearning for LLMs is "(1) the *targeted removal* of the effect of training data from the trained model and (2) the *targeted suppression* of content in a generative-AI model's outputs." The focus is on hiding data. Thaker et al. [43] propose LLM guardrail approaches such as prompting or filtering as less computationally expensive alternatives to traditional unlearning methods.

Machine unlearning is an umbrella term for a wide group of different technologies. Outlining and explaining these technologies in detail would go beyond the scope of this paper. For this, we refer to Figure 2, where we distilled relevant unlearning techniques from the literature with a focus on aspects which we consider to be important. Machine unlearning alone cannot solve the problem of PD-accuracy failures, but we want to highlight two specific cases where existing unlearning methods can be leveraged for mitigation.

The PD-Accuracy Failure Is Caused by Outdated, Biased or Incorrect Training Data. Here, unlearning techniques could be used at different stages of the model pipeline. A careful data selection during the data preprocessing phase could mitigate the inclusion of outdated, biased or incorrect personal information. Additionally, unlearning methods could be applied during finetuning, alignment and operation to remove or suppress inaccurate or outdated information about natural persons. Prompting could be used to check for personal data and to cue the model to refuse to answer if it is uncertain. By itself, this would not ensure PD-accuracy. However, risk-averse operators might prefer to give no information about a person over faulty information.

The PD-Accuracy Failure Has Already Occurred and Can Be Detected. Unlearning techniques could be used for the suppression of outputs when PD-accuracy failures occur, to prevent the further reproduction and spreading of the inaccurate information. Here, filters could be used. However, this requires that the

PD-accuracy failure is detected and reported. As outlined in Section 2, detection of inaccurate personal data can be difficult.

We can see that machine unlearning techniques can only mitigate some aspects of PD-accuracy failures. Not all of the dimensions of our taxonomy can be addressed equally well. *Fact*, *location* and part of the *time* dimension (specifically, outdated information) are relatively straightforward to handle, as they mainly require removing outdated or inaccurate information. However, addressing other aspects of *time*, such as temporal understanding, as well as *role*, *relation*, *context*, *identity*, and *inference* is more involved.

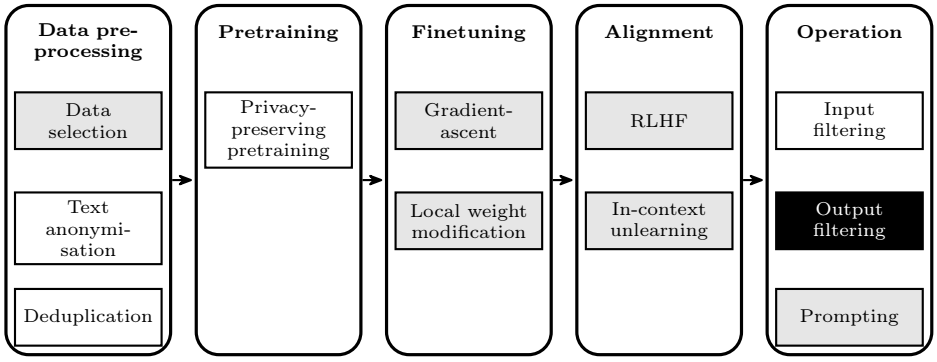

Fig. 2. Suitability of selected machine unlearning methods at different stages of the model pipeline for mitigating PD-accuracy failures. Methods which could be used or adapted to prevent PD-accuracy failures caused by outdated, biased or incorrect training data are highlighted in grey. Techniques that could be used or adapted for output suppression after a PD-accuracy failure was detected are highlighted in black.

6 Discussion

Now we synthesise our observations into five theses.

The Current Focus of Privacy Research for LLMs Should Be Broadened. Current privacy research mainly focuses on privacy leakage and privacy attacks. While these are important topics, we need to examine a broader range of privacy issues, like PD-accuracy failures, in order to responsibly and ethically utilize LLMs. At the moment, it is often unclear how compliance with data protection laws can be ensured when using LLMs. This might hinder their adoption in many areas where personal data is processed, foregoing the potential benefits of this technology.

Additionally, as LLMs are about to replace search engines and personal homepages as a primary source of information for many people, it is of utmost importance that individuals have the power to ensure that information about themselves is represented accurately. Yet, improving the accuracy of LLM outputs

containing personal data is not without risk. If a user wants to hide personal information, improved accuracy might not be desirable in all situations. However, we think that most users do not want an LLM to spread misinformation about them. In a system where a user can determine which data should be processed and used by the LLM, they would also be more interested in their data being correct. Here, the goal would be autonomy instead of obfuscation. Yet, prior work indicates a gap between this ideal of informed autonomy and current user understanding. A user study by Malki et al. [25] showed that while users were concerned about privacy risks of LLMs, they lacked awareness of many potential risk scenarios and held misconceptions about how easily their data could be removed from an LLM. This suggests that a shift toward increased autonomy is not only beneficial in terms of PD-accuracy, but may also strengthen user privacy in general.

Inaccurate personal data can also be spread by other persons or organisations. However, we think that there are unique problems associated with LLMs generating inaccurate data. It has been shown that people tend to place considerable trust in AI. Klingbeil et al. [17] showed in their user study that participants tended to trust AI advice more than they trusted expert advice. Additionally, AI generated misinformation can spread widely and can be taken up by other AI systems, which again spread it. This can also negatively impact their utility [40].

More Data Concerning PD-Accuracy Failures Is Needed. In our review, we have observed that there were only very few documented cases of PD-accuracy failures. In many of the cases we found, either a legal complaint or a complaint to LLM providers had been made. We believe that there are many more incidents which are not reported to the public. More data and more systematic data is needed here. One possible way to achieve this could be to create a database where users can submit PD-accuracy failures they have observed. Our taxonomy could help to organise these failures.

Dedicated Benchmarks for Measuring PD-Accuracy Failures Are Needed. Our taxonomy for PD-accuracy failures can be used as a starting point to explore them in more detail. One way to do this would be to create specialised benchmarks and studies. Our taxonomy has emerged as stable from group discussions. However, it is not set in stone and shall be refined to include new or more detailed dimensions of PD-accuracy failures in the future. Future work could examine how prevalent PD-accuracy failures are, what causes them, and how they can be mitigated.

Be Clear About Where the Failure Occurs: The LLM or the System as a Whole. When studying PD-accuracy failures, e.g. in the form of a benchmark, a distinction should be made on whether the LLM is used to retrieve or to process data. LLMs are often used as retrieval tools or in connection with retrieval systems (e.g. web search). Even if the LLM retrieves data correctly, this data can be inaccurate or out of date. A high quality LLM preserves information it has seen. Having this is necessary for PD-accuracy, but not sufficient. Inadequate

use, such as feeding the LLM with wrong facts, can still lead to inaccurate outputs. Similarly, if the LLM training data is incorrect, it cannot be expected that this is recognized by the data processing system (LLM) itself. While inaccurate training data can cause PD-accuracy failures, the inaccuracy is not introduced by the LLM itself.

PD-accuracy failures can also be introduced through model-postprocessing. Here, the retrieved data or the training data is correct, but the LLM introduces inaccuracies in the processing step. Since most LLM applications involve some form of data processing, we consider this to be the more relevant scenario. We think that a PD-accuracy benchmark should focus on these kinds of failures, since in this case, the inaccuracy is introduced by the LLM. A benchmark for PD-accuracy should measure whether the LLM modifies personal data, independent of whether this data is training data, retrieved data or context data. The focus should be on how many inaccuracies are introduced by the LLM itself and what they look like.

Existing Machine Unlearning Techniques Can Serve As Starting Points for Mitigations. While current machine unlearning methods are a first step towards mitigation, they alone cannot remedy the problem of PD-accuracy failures. For once, they do not solve aspects like the modification of personal data or how new data could be added to the model. This often implies a loss of utility, since personal data is only deleted or suppressed, but not replaced with correct data. They might also be overly broad and prohibitive. For instance, a filter solution might not be limited to just filtering out inaccurate information about a person, but could prevent the LLM from generating any information about this person at all. In some of the documented examples of PD-accuracy failures we have found, filters have been used in this way. As can be seen from the example of Martin Bernklau, this restrictive approach can also negatively impact the affected individual. To mitigate PD-accuracy failures in LLMs, we need to explore different kinds of machine unlearning techniques. Only employing output filters is just a temporary patch and does not solve the underlying problem. Giving no information about a person deprives individuals of their right to informational self-determination.

7 Conclusion

We conclude that PD-accuracy is an important privacy issue that is underexplored in current LLM research. In this paper, we recall interpretations of it in research and in law. We propose a taxonomy with eight different dimensions of PD-accuracy failures: fact, role, relation, identity, time, location, context and inference. Then we turn towards mitigations. We discuss whether machine unlearning methods are a suitable solution to remediate PD-accuracy failures and identify gaps. Additional technical solutions in different development phases of the LLM are needed, e.g. in architecture design, pre-training, fine-tuning, prompting or through filter solutions. This requires experimental research which would

go beyond the scope of this paper. The work on this problem starts when the importance of this issue resonates.

PD-accuracy failures can have serious consequences for the people involved. If data is processed by a third party, they might even occur without the knowledge of the affected person. PD-accuracy is also important in order to comply with existing data protection laws. We therefore think it is time to address this blind spot in LLM research. Shifting the focus to PD-accuracy should go beyond counting another type of failure and move the focus towards individual's autonomy over their data [15]. In the long run, it will empower people to better control what LLMs say about them.

Acknowledgments. GPT-5 was used to improve the language quality of a few individual sentences, in order to enhance clarity and avoid word repetitions. We thank Constantin Lessmann, Judith Senn, Paulina Pesch, Max Ninow, Manuel Eberl and the anonymous reviewers for their valuable comments on earlier versions of this work. We gratefully acknowledge funding by the German Federal Ministry of Research, Technology and Space (BMFTR) under grant agreement number 16KIS2304 (SMARD-GOV).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. p. 308–318. ACM (2016). https://doi.org/10.1145/2976749.2978318
2. AI, Algorithmic and Automation Incident and Controversy Repository AIAAIC: ChatGPT wrongly claims Alexander Hanff is dead (2023), https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/chatgpt-wrongly-claims-alexander-hanff-is-dead
3. Atherton, D.: Incident number 998: ChatGPT allegedly defamed Norwegian user by inventing child homicide and imprisonment (2024), https://incidentdatabase.ai/cite/998
4. Bang, Y., Ji, Z., Schelten, A., Hartshorn, A., Fowler, T., Zhang, C., Cancedda, N., Fung, P.: HalluLens: LLM hallucination benchmark. In: Annual Meeting of the Association for Computational Linguistics. pp. 24128–24156. ACL (2025). https://doi.org/10.18653/v1/2025.acl-long.1176
5. Blanco-Justicia, A., Jebreel, N., Manzanares-Salor, B., Sánchez, D., Domingo-Ferrer, J., Collell, G., Eeik Tan, K.: Digital forgetting in large language models: A survey of unlearning methods. Artificial Intelligence Review **58**(3) (2025)
6. Bradford, A.: The Brussels Effect: How the European Union Rules the World. Oxford University Press (2020), https://doi.org/10.1093/oso/9780190088583.001.0001
7. Butters, N.: Incident number 507: ChatGPT erroneously alleged mayor served prison time for bribery (2023), https://incidentdatabase.ai/cite/507
8. Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., Zhang, C.: Quantifying memorization across neural language models. In: The Eleventh International Conference on Learning Representations (2023)

9. Chen, Z., Ma, J., Zhang, X., Hao, N., Yan, A., Nourbakhsh, A., Yang, X., McAuley, J., Petzold, L.R., Wang, W.Y.: A survey on large language models for critical societal domains: Finance, healthcare, and law. Transactions on Machine Learning Research (2024), https://openreview.net/forum?id=upAWnMgpnH
10. Christakis, T.: AI hallucinations and data subject rights under the GDPR: Regulatory perspectives and industry responses (2024), https://ssrn.com/abstract=5042191
11. Commission, U.F.T.: Privacy online: A report to congress (1998), https://www.ftc.gov/sites/default/files/documents/reports/privacy-online-report-congress/priv-23a.pdf
12. Cooper, A.F., Choquette-Choo, C.A., Bogen, M., Klyman, K., Jagielski, M., Filippova, K., Liu, K., Chouldechova, A., Hayes, J., Huang, Y., Mireshghallah, N., Shumailov, I., Triantafillou, E., Kairouz, P., Mitchell, N., Liang, P., Ho, D.E., Choi, Y., Koyejo, S., Delgado, F., Grimmelmann, J., Shmatikov, V., De Sa, C., Barocas, S., Cyphert, A., Lemley, M.A., Samuelson, P., Boyd, D., Wortman Vaughan, J., Brundage, M., Bau, D., Neel, S., Jacobs, A., Terzis, A., Wallach, H., Papernot, N., Lee, K.: Machine unlearning doesn't do what you think: Lessons for generative AI policy and research. Stanford Public Law Working Paper (2024), https://ssrn.com/abstract=5288768
13. Das, B.C., Amini, M.H., Wu, Y.: Security and privacy challenges of large language models: A survey. ACM Computing Surveys **57** (2025). https://doi.org/10.1145/3712001
14. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Theory of Cryptography. pp. 265–284. Springer Berlin Heidelberg (2006)
15. Hirshleifer, J.: Privacy: Its origin, function, and future. The Journal of Legal Studies **9**(4), 649–664 (1980), https://chicagounbound.uchicago.edu/jls/vol9/iss4/4
16. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., Liu, T.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems **43**(2), 51–5 (2025), https://doi.org/10.1145/3703155
17. Klingbeil, A., Grützner, C., Schreck, P.: Trust and reliance on AI — an experimental study on the extent and costs of overreliance on AI. Computers in Human Behavior **160** (2024), https://doi.org/10.1016/j.chb.2024.108352
18. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 260–270. ACL (2016). https://doi.org/10.18653/v1/N16-1030
19. Le Jeune, P., Liu, J., Rossi, L., Dora, M.: RealHarm: A collection of real-world language model application failures. In: The First Workshop on LLM Security (LLMSEC). pp. 87–100. ACL (2025), https://aclanthology.org/2025.llmsec-1.7/
20. Lee, Y., Son, K., Kim, T.S., Kim, J., Chung, J.J.Y., Adar, E., Kim, J.: One vs. many: Comprehending accurate information from multiple erroneous and inconsistent AI generations. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. p. 2518–2531. ACM (2024). https://doi.org/10.1145/3630106.3662681
21. Lee, Y., Ye, X., Choi, E.: AmbigDocs: Reasoning across documents on different entities under the same name. In: First Conference on Language Modeling (2024), https://openreview.net/forum?id=mkYCfO822n

22. Li, Q., Hong, J., Xie, C., Tan, J., Xin, R., Hou, J., Yin, X., Wang, Z., Hendrycks, D., Wang, Z., Li, B., He, B., Song, D.: LLM-PBE: Assessing data privacy in large language models. Proceedings of the VLDB Endowment **17**(11), 3201–3214 (2024), https://doi.org/10.14778/3681954.3681994
23. Liu, S., Yao, Y., Jia, J., Casper, S., Baracaldo, N., Hase, P., Yao, Y., Liu, C.Y., Xu, X., Li, H., Varshney, K.R., Bansal, M., Koyejo, S., Liu, Y.: Rethinking machine unlearning for large language models. Nature Machine Intelligence **7**, 181–194 (2025), https://doi.org/10.1038/s42256-025-00985-0
24. Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., Zanella-Béguelin, S.: Analyzing leakage of personally identifiable information in language models. In: 2023 IEEE Symposium on Security and Privacy (SP). pp. 346–363. IEEE (2023)
25. Malki, L.M., Polamarasetty, A., Hatamian, M., Warner, M., Costanza, E.: Hoovered up as a data point: Exploring privacy behaviours, awareness, and concerns among UK users of LLM-based conversational agents. Proceedings on Privacy Enhancing Technologies pp. 838—-860 (2025)
26. McGregor, S.: Preventing repeated real world AI failures by cataloging incidents: The AI incident database. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 15458–15463 (2021), https://cdn.aaai.org/ojs/17817/17817-13-21311-1-2-20210518.pdf
27. Nasr, M., Rando, J., Carlini, N., Hayase, J., Jagielski, M., Cooper, A.F., Ippolito, D., Choquette-Choo, C.A., Tramèr, F., Lee, K.: Scalable extraction of training data from aligned, production language models. In: The Thirteenth International Conference on Learning Representations (2025)
28. Neel, S., Chang, P.: Privacy issues in large language models: A survey (2024), https://arxiv.org/abs/2312.06717
29. Nissenbaum, H.: Privacy as contextual integrity. Washington Law Review **79**, 118–158 (2004), https://digitalcommons.law.uw.edu/wlr/vol79/iss1/10
30. Novelli, C., Casolari, F., Hacker, P., Spedicato, G., Floridi, L.: Generative AI in EU law: Liability, privacy, intellectual property, and cybersecurity. Computer Law & Security Review **55** (2024), https://www.sciencedirect.com/science/article/pii/S0267364924001328
31. noyb: Noyb data protection complaints (nd), https://noyb.eu/en/projects
32. Occident, O.: Incident number 506: ChatGPT allegedly produced false accusation of sexual harassment (2023), https://incidentdatabase.ai/cite/506
33. OECD: OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data (2002), https://doi.org/10.1787/9789264196391-en
34. OECD.AI: OECD AI incidents and hazards monitor (AIM) (nd), https://oecd.ai/en/incidents
35. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. Advances in neural information processing systems **35**, 27730–27744 (2022)
36. Paudel, B., Mandal, B., Amariucai, G., Wei, S.: Sanitization or deception? Rethinking privacy protection in large language models. Proceedings on Privacy Enhancing Technologies pp. 154–174 (2026)
37. Pesch, P., Böhme, R.: ChatGPT & Co. unter der DSGVO – Verarbeitung personenbezogener Daten und Datenrichtigkeit bei großen Sprachmodellen. Multimedia und Recht **26**, 913–923 (2023)
38. Pownall, C.: AI, Algorithmic and Automation Incident and Controversy Repository AIAAIC (2021), https://www.aiaaic.org/aiaaic-repository

39. Rossello, S.: LLM hallucinations and personal data accuracy: Can they really coexist? Available at SSRN 5162539 (2025)
40. Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., Gal, Y.: AI models collapse when trained on recursively generated data. Nature **631**, 755–759 (2024). https://doi.org/10.1038/s41586-024-07566-y
41. Solove, D.J.: A taxonomy of privacy. University of Pennsylvania Law Review pp. 477–564 (2006)
42. Starkweather, C.: Incident number 770: Microsoft Copilot falsely accuses journalist Martin Bernklau of crimes (2024), https://incidentdatabase.ai/cite/770
43. Thaker, P., Maurya, Y., Hu, S., Wu, Z.S., Smith, V.: Guardrail baselines for unlearning in LLMs. In: ICLR Workshop on Secure and Trustworthy Large Language Models (2024), https://openreview.net/pdf?id=eBcVsC4h6A
44. Translation, J.L.: Act on the protection of personal information (act no. 57 of 2003) (2003), https://www.japaneselawtranslation.go.jp/en/laws/view/4241/en
45. European Union Law: Regulation (EU) 2016/679 of the European Parliament and of the Council (2016)
46. European Union Law: Regulation (EU) 2016/679 of the European Parliament and of the Council. Article 4 (2016)
47. European Union Law: Regulation (EU) 2016/679 of the European Parliament and of the Council. Article 5 (2016)
48. European Union Law: Regulation (EU) 2016/679 of the European Parliament and of the Council. Article 17 (2016)
49. European Union Law: Regulation (EU) 2016/679 of the European Parliament and of the Council. Article 16 (2016)
50. Wallat, J., Abdallah, A., Jatowt, A., Anand, A.: A study into investigating temporal robustness of LLMs. In: Findings of the Association for Computational Linguistics. pp. 15685–15705. ACL (2025). https://doi.org/10.18653/v1/2025.findings-acl.810
51. Wallat, J., Jatowt, A., Anand, A.: Temporal blind spots in large language models. In: International Conference on Web Search and Data Mining. p. 683–692. ACM (2024). https://doi.org/10.1145/3616855.3635818
52. Yan, B., Li, K., Xu, M., Dong, Y., Zhang, Y., Ren, Z., Cheng, X.: On protecting the data privacy of large language models (LLMs): A survey. In: International Conference on Meta Computing (ICMC). pp. 1–12. IEEE (2024), https://doi.ieeecomputersociety.org/10.1109/ICMC60390.2024.00008
53. Yan, B., Li, K., Xu, M., Dong, Y., Zhang, Y., Ren, Z., Cheng, X.: On protecting the data privacy of large language models (LLMs) and LLM agents: A literature review. In: High-Confidence Computing. vol. 5 (2025), https://www.sciencedirect.com/science/article/pii/S2667295225000042
54. Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A.T., Bi, W., Shi, F., Shi, S.: Siren's song in the AI ocean: A survey on hallucination in large language models. Computational Linguistics pp. 1–45 (2025), https://doi.org/10.1162/coli.a.16