The Hamburg Commissioner for
Data protection and freedom of information

Discussion Paper: Large Language Models and Personal Data

This discussion paper reflects the current state of knowledge and understanding at the Hamburg Commissioner for Data Protection and Freedom of Information (HmbBfDI) regarding the applicability of the General Data Protection Regulation (GDPR) to Large Language Models[1] (LLMs). This paper aims to *stimulate further debate.* It is intended to support companies and public authorities in better navigating complex data protection issues surrounding this subject matter. To this end, this paper explains relevant technical aspects of LLMs, assesses them in light of case law regarding personal data from the Court of Justice of the European Union (CJEU) and highlights the resulting practical implications. From this, three principle theses can be derived:

1. **The mere storage of an LLM does not constitute processing within the meaning of article 4 (2) GDPR. This is because no personal data is stored in LLMs. Insofar as personal data is processed in an LLM-supported AI system, the processing must comply with the requirements of the GDPR. This applies in particular to the output of such an AI system.**

2. **Given that no personal data is stored in LLMs, data subject rights as defined in the GDPR cannot relate to the model itself. However, claims for access, erasure or rectification can certainly relate to the input and output of an AI system of the responsible provider or deployer.**

3. **The training of LLMs using personal data must comply with data protection regulations. Throughout this process, data subject rights must also be upheld. However, potential violations during the LLMs training phase do not affect the lawfulness of using such a model within an AI system.**

[1] This refers exclusively to models as an important, but not sole, component of a comprehensive AI system (e.g. an LLM-based chatbot).

I. Introduction

When an LLM, functioning as a component of an AI system, processes[2] prompts (so-called "inference"), the LLM´s output may contain information relating to natural persons, especially if the prompt specifically asks for it. This raises the question of whether personal data is stored in an LLM.

To answer this question, it is crucial to distinguish between an AI system and any LLM it may incorporate. An AI system consists of multiple components. An LLM is one such component. It cannot be used meaningfully without other components that form an AI system. Chatbots such as ChatGPT exemplify this multi-component structure: Their most important components include the user interface,[3] input and output filters and the LLM. The user input is usually first processed by other components of the AI system before the LLM is inferred. For example, the user input ("prompts") can be enriched with further information from a database, an internet search or by means of Retrieval Augmented Generation (RAG). Only then does the LLM process the modified prompt. The raw output generated by the LLM is then typically further processed by filters before it is - as a rule – being presented to the user via the interface.

The following sections (II. and III.) do not evaluate the processing activities in the entire AI system. They focus exclusively on the question of whether personal data is stored in LLMs.

II. Technical evaluation of LLM

LLMs process language, usually several languages.[4] Initially, they are trained with large amounts of textual input in the relevant languages. In turn, they deliver linguistic results in the output.

1. Tokens as the basic element of LLM's information processing

Understanding how linguistic information is processed and stored in Large Language Models (LLMs) is vital for addressing the question at hand. A key aspect of this process is the “tokenization” of input text. All texts are divided into comparatively small predefined chunks, so-called tokens, before they find their way into an LLM. These pieces are usually smaller than whole words

[2] Cf. also article 3 (1) AI Act.
[3] AI systems can be used via websites or specially developed apps.
[4] This includes various natural languages such as English, French and German as well as computer languages such as Python, JavaScript and Ruby. For the purposes of this paper, only natural language aspects are relevant.

but larger than individual letters. The challenge in developing LLMs is to manage with a finite set of basic elements (usually several tens of thousands) that can be related to one another. Consequently, longer words, phrases or even whole sentences as such are not directly incorporated into an LLM. The German sentence "Ist ein LLM personenbezogen?", which loosely translates to "Does an LLM store personal data?" is divided into 12 tokens by a typical tokenizer,[5] for example, as follows: [I][st][e][in][LL][M] [person][en][be][z][ogen] [?]. These tokens are converted into numerical values,[6] which are used exclusively within the model in the following process.

Hence, within LLMs texts are no longer stored in their original form, or only as fragments in the form of these numerical tokens. They are further processed into "embeddings".[7] These embeddings capture learned correlations by positioning tokens in relation to each other, i.e. assigning them according to probability weights. This describes the core "training" of an LLM. Furthermore, this mathematical representation of the trained input is used for the inference of a prompt. The embeddings represent the learned "knowledge" of the LLM.

Accordingly, during inference the output of an LLM is first produced as a sequence of tokens, which are then converted back into the corresponding letter sequences before being processed further.[8] An output such as the German sentence "Mia Müller hat gelogen" ("Mia Müller has lied") is also constructed by the LLM from tokens. Notably, some of these tokens, such as the first "**M**" and "**ogen**" are the same tokens as those in the example above [I][st][e][in][LL][**M**] [person][en][be][z][**ogen**].[9] In a specific and appropriate context, the token "ogen" might be chosen to follow the token "gel" to produce the word "gelogen" ("lied"), while in another context, for example, the token "b" follows "gel" to produce the word "gelb" ("yellow").

2. Storage of Information in LLMs

The text or token "Mia Müller" is not stored anywhere within the model. Individual tokens such as "M", "ia", "Mü" and "ller" are merely linguistic fragments. The vectorial relationships between the tokens "Mü" and "ller" are such that the token "ller" presumably follows "Mü" more frequently (at least in certain contexts) than, for example, the token "he" to produce the German word

[5] Here OpenAI for GPT-3, see https://platform.openai.com/tokenizer.
[6] In this example, these are the values [40, 301, 304, 259, 27140, 44, 1048, 268, 1350, 89, 6644, 30], i.e. "[I]" the value 40, "[st]" the value 301, etc. - the specific values are LLM-specific and not standardized across the board.
[7] Mathematically, these are vectors in a multidimensional vector space, e.g. [-0.74, 0.42, -0.53, ..., 0.02].
[8] E.g. the output via the user interface.
[9] Specifically, the tokenization here is [M][ia][Mü][ller][hat][gel][ogen][.]

"Mühe" (“effort”). These relationships between the tokens – the embedding – constitutes the core achievement of an LLM and is ultimately what makes them useful. The model's capacity is often measured by the number of parameters, which define token relationships resulting from the training process. Modern LLMs typically contain billions of such parameters.[10] These learned variables of an LLM are complex, not fully interpretable and cannot be specifically adjusted without risking the model’s overall functionality.[11] They represent the "meaning" derived from trained texts without storing the texts themselves.[12] When training data contains personal data, it undergoes a transformation during machine learning process, converting it into abstract mathematical representations. This abstraction process results in the loss of concrete characteristics and references to specific individuals. Instead, the model captures general patterns and correlations derived from the training data as a whole.

In this context, a somewhat contradictory characteristic of such models occurs: On the one hand, they do not store the texts used for training in their original form, but process them in such a way that the training data set can never be fully reconstructed from the model. On the other hand, LLMs process these training texts in a very specific manner based on contextual relationships, which enables the generation of similar and often useful output texts. However, everything that LLMs produce is "created" in the sense that it is not a simple reproduction of something stored (such as an entry in a database or a text document), but rather something newly produced. This probabilistic generation capability fundamentally differs from conventional data storage and data retrieval.

III. Storage of personal data in the LLM

The legal term "personal data" (article 4 (1) GDPR) is a concept that has been substantiated by CJEU case law. It is not synonymous with the general public’s understanding of information on an individual person. Personal data, in the legal sense, refers to information that “relates“ to an

[10] The LLM Llama 3 is available, for example, in a version with 8 billion or 70 billion parameters.
[11] This is also shown by a research project by Anthropic, whose adaptation of the LLM feature "Golden Gate Bridge" led to responses from a chatbot only revolving around this bridge, even though the prompt did not mention it, for examples see https://www.anthropic.com/news/golden-gate-claude; on the research paper Templeton et. al, 2024, https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.
[12] In the current debate, it is occasionally argued that the storage of embeddings is comparable to cryptographic encryption. This is not the case. This is because encryption is a bijective mapping between plaintext and ciphertext. This means that each element of the plaintext set is assigned exactly one element of the ciphertext set and vice versa. With the right key, the plaintext can be uniquely reconstructed from the ciphertext. The information is therefore retained in full, but cannot be read without the key as long as the encryption process has not been broken. This is not the case with the training data of an LLM and its abstract representation in the embeddings. There is no "key" that completely and comprehensively restores the original data of the training phase.

identified or identifiable natural person. For instance, a library card, even if it only displays a series of digits, can be personal data. If the individual associated with the library card number can be identified through other means, such as the library's database, the CJEU would consider the number itself as personal data for someone with access to said database. However, the CJEU stipulates that only lawful means of identification that don't require disproportionate effort in practice should be considered.[13]

The CJEU has not yet ruled on the storage of personal data in LLMs or comparable technologies. Taking into account previous CJEU case law and known methods of attacking LLMs, the HmbBfDI concludes that an LLM does not store personal data within the meaning of article 4 (1), (2) GDPR in conjunction with Recital 26.[14] Although it has been observed that fine-tuned LLMs are occasionally made to reproduce training data through privacy attacks, it is doubtful whether this type of extraction validates the legal conclusion that personal data is stored in the LLM.

1. Relevance of (embedded) tokens in LLMs

The constellations dealt with by the CJEU regarding personal data concerned IP addresses, exam responses, legal memos by public offices, vehicle identification numbers or other coded character strings such as the TC string.[15] These indicate a reference to the identification of a specific person or to objects assigned to persons. They are identifiers or - according to the CJEU - information that “relates“ to the data subject.[16]

This “relation” results from the function of these identifiers and the information they contain.[17] IP addresses are used to assign a device to enable users to send and receive data online. They establish a relational link between an online activity and a physical person.[18] Examination responses and examiner notes are intended to evaluate a specific person to be identified and

[13] CJEU, 19.10.2016, C-582/14, para. 46.

[14] According to the Danish Data Protection Authority, an AI model as such does not contain any personal data. It is merely the result of the processing of personal data. This follows from the fact that a statistical report is also not considered personal data if the report only contains conclusions and aggregated data that are the result of the statistical analysis (Guidelines from the Danish Data Protection Agency on the use of artificial intelligence, p.7, published in October 2023, available at: https://www.datatilsynet.dk/Media/638321084132236143/Offentlige%20myndigheders%20brug%20af%20kunstig%20intelligens%20-%20Inden%20I%20g%C3%A5r%20i%20gang.pdf).

[15] The Transparency and Consent String is used in the online advertising industry to encode users’ consents and preferences regarding data processing for advertising purposes and to transmit them to the parties involved.

[16] CJEU, 20.12.2017 - C-434/16, para. 34.

[17] CJEU, 20.12.2017 - C-434/16, para. 35; see CJEU, 4.5.2023 - C-487/21, para. 24, CJEU, of 8.12.2022 - C180/21, para. 70.

[18] Cf. CJEU, 19.10.2016, C-582/14, para. 36.

their professional competence.[19] A legal memo on the fulfillment of residence requirements provides detailed information about an applicant’s personal circumstances, such as their nationality, marital status or employment situation.[20] A vehicle identification number allows identifying a specific vehicle and indirectly its owner by coding vehicle-specific features such as manufacturer, type, equipment, place of production and year of manufacturing.[21] The TC string is used to communicate specific information about the user's preferences and behavior, such as the processing purposes and providers for which the user has given or refused consent.[22] All of these identifiers therefore serve as placeholders for the identity, characteristics or circumstances of a specific person or object of a person; their function is a targeted association.

Unlike these identifiers addressed in CJEU case law, individual tokens as language fragments ("M", "ia", " Mü" and "ller") lack individual information content and do not function as placeholders for such. Even the embeddings, which represent relationships between these tokens, are merely mathematical representations of the trained input. For instance, the higher frequency of „Mü“ and „ller“ co-occurring compared to „Mü“ and „he“ reflects linguistic patterns rather than information about an individual named Mia Müller. LLMs store highly abstracted and aggregated data points from training data and their relationships to each other, without concrete characteristics or references that “relate“ to individuals. Unlike the identifiers addressed in CJEU case law, which directly link to specific individuals, neither individual tokens nor their embeddings in LLMs contain such information about natural persons from the training dataset. Therefore, according to the standards set by CJEU jurisprudence, the question of whether personal data is stored in LLMs does not depend on the means available to a potential controller. In LLMs, the stored information already lacks the necessary direct, targeted association to individuals that characterizes personal data in CJEU jurisprudence: the information "relating" to a natural person.

2. Effects of privacy attacks and PII extraction

It has been observed that LLMs optimized for specific tasks through fine-tuning can, under certain circumstances, reproduce training data – including information relating to natural persons.[23]

[19] CJEU, 20.12.2017 - C-434/16, para. 38.
[20] Cf. CJEU, 17.7.2014 - C-141/12, C-372/12, para. 48.
[21] CJEU, 9.11.2023 - C-319/22, para. 49.
[22] CJEU 7.3.2024 - C-604/22, para. 21.
[23] On this subject matter, cf. Das et al., 2024, https://arxiv.org/html/2402.00888v1; such extraction attacks are to be distinguished from model inversion or membership inference attacks, which attempt to draw conclusions about statistical properties or the affiliation to a training data set through targeted queries to the LLM, without extracting training data sets as such. They do not make any statement about the possible storage of data in the LLM itself, but about the probability that data with certain properties or a certain data set has been used for training, cf. Maini et. al. 2024,

This reproduction of training content has led to the conclusion that such information relating to natural persons must be stored ("memorized") in the LLM, despite being represented in the form of tokens or embeddings. However, it is doubtful whether this type of extraction also allows the legal conclusion that personal data is stored in the LLM.

First of all, the mere presence of plausible personal information in LLM output is not conclusive evidence that personal data has been memorized, as LLMs are capable of generating texts that coincidentally matches training data.

Moreover, reproduction typically occurs only through targeted attacks on LLMs,[24] such as "privacy attacks" or "PII extraction".[25] According to CJEU case law, however, data can only be classified as personal if identification is possible through means of the controller or third parties that are not prohibited by law and do not require disproportionate effort in terms of time, cost and manpower.[26] Conducting such attacks could be considered a practically disproportionate effort and potentially a legally prohibited means of determining whether personal data is stored in the LLM.

Generally, designing and executing effective privacy attacks on LLMs require substantial technical expertise and time resources that the average user lacks.[27] In order to verify whether the texts generated by an LLM have actually been extracted from the training data or hallucinated, current privacy attacks require access to the original training data.[28] Only through direct comparison of LLM output with original training data it can be argued that personal data from the training dataset may have been stored in an LLM.[29] However, training datasets from LLMs are generally not fully accessible to the public.[30] At the same time, validating the correlation between

https://arxiv.org/abs/2406.06443; for the difference between model inversion attacks and membership inference attacks, cf. Yang et. al. 2020, https://arxiv.org/pdf/2005.03915.

[24] For other attack variants, see Das et al., 2024, https://arxiv.org/html/2402.00888v1.

[25] Whereby the acronym "PII" means Personally Identifiable Information, a term used in various US laws that cannot be equated with the term personal data under the GDPR.

[26] CJEU, 19.10.2016, C-582/14, para. 46.

[27] This is all the more true as known attacks from the past are regularly taken up by developers in order to take further protective measures and therefore require considerable expertise in order to develop new attacks, see also below.

[28] Cf. Das et al., 2024, https://arxiv.org/html/2402.00888v1.

[29] Such attacks, which do compare outputs to training data, are also laborious – not least because model developers implement safeguards - and may constitute a prohibited means under German law.

[30] Shi et al, 2024, "Although large language models (LLMs) are widely deployed, the data used to train them is rarely disclosed", https://arxiv.org/html/2310.16789v3, not all training data is disclosed even for LLMs labeled as "open source": "for Llama, the corporate preprint notes that fine-tuning was done based on "a large dataset of over 1 million binary comparisons based on humans applying our specified guidelines, which we refer to as Meta reward modeling data", and which remains undisclosed", Liesenfeld/Dingemanse, Rethinking open source generative AI, open-washing and the EU AI Act, https://pure.mpg.de/rest/items/item_3588217_2/component/file_3588218/content.

the vast array of possible LLM outputs and the training dataset requires immense amounts of training data. These prerequisites for an attack are not only practically demanding but also cost-intensive,[31] constituting a disproportionate effort currently undertaken only for scientific research into LLM technology.[32] Furthermore, such attacks often lead to LLM-developers improving protective measures to detect and prevent such extractions.[33] If such technical protective measures are deliberately overcome, it cannot be ruled out that this could constitute prohibited means under German law.[34]

IV. Practical Implications

The thesis that LLMs do not store personal data carries significant implications for practical applications, as illustrated by the following examples:

1. Consequences of unlawful training of LLMs

A company or public authority deploys an LLM developed by a third party. It later emerges that the third party used personal data in training the model without a legal basis.

Potentially unlawful[35] processing of personal data during the training of a model does not affect the legality of using said model. Data protection violations during the training of an LLM are not attributable to the controller who deploys the LLM, but exclusively to model's developer. The latter, like companies and authorities wishing to deploy and fine-tune LLMs, must comply with data protection regulations.

2. Significance for data subjects

A person enters their name into a company's or authority's LLM-based chatbot. The LLM-based chatbot provides incorrect information about them. Which data subject rights can they assert in relation to which subject matter?

[31] Cf. Das et al, 2024, https://arxiv.org/html/2402.00888v1; the statement of another research project that a PII attack only triggered costs of USD 200 ultimately only concerns the costs of the 25 million queries, but not the technical expertise, the time required, the remuneration required for this and the comparison of the extracted data with nine terabytes of partial training data sets, see Nasr et al, 2023, https://arxiv.org/pdf/2311.17035.

[32] Research also states: "Memorization Can Be Hard to Discover.", Carlini et al., 2021, https://www.usenix.org/system/files/sec21-carlini-extracting.pdf.

[33] Cf. on suitable red teaming practices as a method for the (further) development of protective measures, Feffer et al https://arxiv.org/pdf/2401.15897; on current limitations of such privacy attacks through protective measures, Das et. al, 2024, https://arxiv.org/html/2402.00888v1.

[34] Cf. § 202a StGB; on the legal risks and ethical reservations of privacy attacks, see Carlini et. al, 2021, Extracting Training Data from Large Language Models, https://www.usenix.org/system/files/sec21-carlini-extracting.pdf; Nasr et al, 2023, https://arxiv.org/abs/2311.17035.

[35] E.g. due to the lack of a legal basis for processing of personal data in the training data.

Organizations must ensure GDPR compliance when processing personal data. As LLMs don't store personal data, they can't be the direct subject of data subject rights under articles 12 et seq. GDPR.[36] However, when an AI system processes personal data, particularly in its output or database queries, the controller must fulfill data subject rights.

In the case outlined above, this means that the data subject can request the organization to provide, at least regarding the input and the output of the LLM chatbot,

- that information is provided in accordance with article 15 GDPR,
- that their personal data is rectified in accordance with article 16 GDPR,
- if applicable, that their personal data is erased in accordance with article. 17 GDPR.

3. Requirements for fine-tuning LLMs

A company or public authority wants to fine-tune an LLM developed by a third party with their own training data for a specific use case.

For the company or the public authority, this means in particular:

- It is recommended that the training data contains as little personal data as possible. Synthetic data should be preferred, as long as it is suitable for training.[37]
- If personal data is used for fine-tuning, organizations must have a legal basis[38] and ensure that data subject rights can be fulfilled.

4. Requirements for local LLM operation

A company or public authority would like to use a locally operated LLM to deploy an internal knowledge management tool via a web interface.

For the company or the authority, this means in particular:

1. The storage of an LLM on the server of a company or public authority is not relevant under data protection law.

[36] Likewise, the processing principles of article 5 GDPR, in particular the principle of data accuracy pursuant to article 5 (1)(d) GDPR, are not applicable to an LLM itself.

[37] The European Data Protection Supervisor derives this from the principle of "Privacy by Design", https://www.edps.europa.eu/press-publications/publications/techsonar/synthetic-data_en, cf. article 25 (1) GDPR.

[38] For possible legal bases, see the discussion paper "Rechtsgrundlagen im Datenschutz beim Einsatz von Künstlicher Intelligenz" by the State Commissioner for Data Protection and Freedom of Information of Baden-Württemberg, available at: https://www.baden-wuerttemberg.datenschutz.de/rechtsgrundlagen-datenschutz-ki/.

2. However, the AI system must enable the fulfillment of data subject rights, in any case, regarding its input and output.
3. Controllers should ensure prevention of extractions, such as through privacy attacks and PII extraction:[39]
    a. The safeguards provided by the LLM developer should be implemented.
    b. In addition to the developer's protective measures, the responsible party should take its own measures, such as filters, to prevent privacy attacks and PII extraction.

5. Requirements for the operation of third-party LLMs

A company or public authority enters into a contract with a third-party provider for the provision of an LLM, for example via an application programming interface (API), to enable employees to create text summaries through a web interface.

For the company or the authority, this means in particular:

1. The AI system must enable the fulfillment of data subject rights, in any case, regarding its input and output.
2. When selecting a provider, it is important to ensure that protective measures are in place to prevent privacy attacks and PII extraction.
3. Before the LLM is put into operation, the responsibilities must be clarified (data processing on behalf, joint controllership, or independent controllership).[40]

[39] But here, too, it remains doubtful whether any extractable data records constitute personal data, see III. 2. above.

[40] Cf. DSK Guidance of 06.05.2024, "AI and data protection", para. 32 ff., https://www.datenschutzkonferenz-online.de/media/oh/20240506_DSK_Orientierungshilfe_KI_und_Datenschutz.pdf.