

# Navigating the Privacy Landscape of Large Language Models: Challenges, Technologies, and Policy Directions

Moges Kelklie

May 6, 2024

## Contents

<b>1</b>	<b>Foundations of large language models</b>	<b>2</b>
1.1	Overview of LLM	2
1.2	Limitation of LLMs	3
<b>2</b>	<b>Personal data and Large language models</b>	<b>3</b>
2.1	LLM training Dataset	4
2.2	LLM operation dataset	5
<b>3</b>	<b>Data privacy regulations and LLMs</b>	<b>6</b>
3.1	European General Data Protection Regulation	6
3.2	US state laws and sectorial approach	7
3.3	EU AI Act and the US AI bill of rights	8
<b>4</b>	<b>Solving the data privacy challenge</b>	<b>8</b>
4.1	privacy enhancing technologies	8
4.2	LLM data privacy policy recommendation	9
4.2.1	Centralized opt out from LLM models	10
4.2.2	Central registry of LLM model providers	10
4.2.3	Anonymized and synthetic personal data services	12
4.3	Evaluation of the proposed policy	12
4.4	Limitation of the proposed solution	15
<b>5</b>	<b>Future work and the U.S. draft privacy legislation</b>	<b>16</b>
5.1	American privacy rights act	16

## Abstract

Data privacy has become a key concern of large language models (LLMs), both in the large trove of information they can infer and in the inherent inflexibility of models in forgetting the learned data. LLMs do not have an easy way to delete information; sensitive data could be inferred through prompt engineering, thus raising concerns about data privacy. This article attempts to address the challenge of LLM data privacy and how policy can help mitigate some of the privacy concerns.

# 1 Foundations of large language models

Large language models (LLMs) have created a great interest in the potential of artificial intelligence in natural language processing, multimodal content creation, identifying patterns in massive data sets, and many other exciting use cases. LLM tools can summarize written text and generate text, audio, and video from a prompt [Al-Amin et al., 2024]. LLMs can also be manipulated and thus can do what they are not intended for, which calls for appropriate governance and technical evaluations. [Yao et al., 2024]

## 1.1 Overview of LLM

Large language models (LLMs) are models that are part of foundational models and have become the new way of training neural networks. Foundational models are part of a deep learning model that has been studied for decades with many applications. Deep learning is a specialized type of learning with artificial neurons, a sub specialty of machine learning. Generative AI sits at the intersection of Deep Learning and the Natural Language Processing specialty of Artificial Intelligence.

LLM models can be auto regressive, which is used to predict the next most likely word in a sentence, and auto-encoding, which is used to reconstruct the original word from a missing input. The transformer architecture [Vaswani et al., 2023] that is used to build LLMs can two components, namely, a decoder and an encoder. The encoder is good for understanding text, while the decoder is good for generating text. GPT models, which are the most popular, are decoder models. BERT [Devlin et al., 2019] is an example of the encoder model, and T5 [Raffel et al., 2023] has both the encoder and decoder packages.

LLMs are pre-trained on a massive corpus of data to understand the mechanism of language and then fine-tuned on specific tasks of the use case. LLMs are good for text generation, translation, software code generation, etc. LLM output should always be inspected for accuracy and should not be used for an automated decision without a human in the loop.

As the training data size for LLMs exponentially increased, LLMs started to exhibit an enhanced phenomenon called emerging abilities. These abilities were discovered by accident and not by design. Many of the commercial and open-source LLMs on the market exhibit capabilities that were not possible just a decade ago. The most common of these LLM powered tools are ChatGPT from OpenAI, Gemini from Google, and Claude Anthropic and OpenSource LAlMA from Meta.

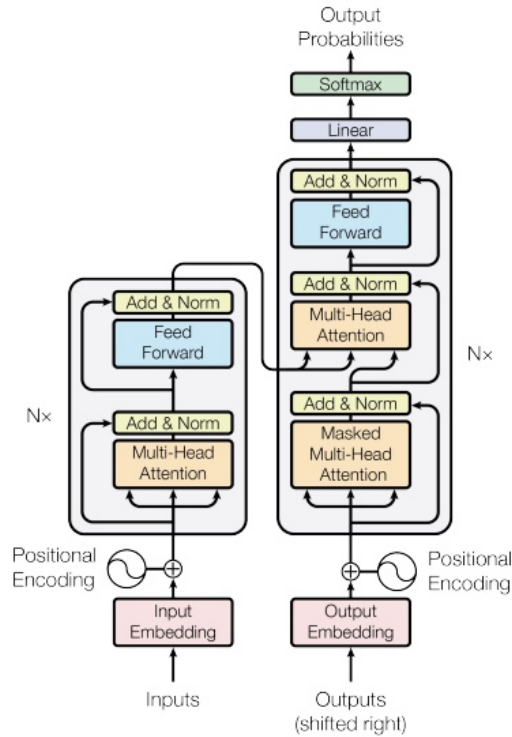


Figure 1: Attention is all you need. [Vaswani et al., 2023]

## 1.2 Limitation of LLMs

Although LLMs have become powerful and have found many applications in the real world, many limitations should be addressed as the technology matures. LLMs are susceptible to hallucinations, data security attacks, and deep fakes; data privacy, and could also come with some bias and ethical concerns[Jia et al., 2023].

LLMs can memorize training data [Staab et al., 2023], which can be used for a purpose not intended in the mode. Some of the other limitations, such as hallucinations, can be minimized by grounding the models, but not all limitations can be solved by technology alone.

## 2 Personal data and Large language models

Building LLMs requires a vast amount of specialized computing and a large corpus of data. Many papers [Neel and Chang, 2024] have been written to address privacy from the technological and mathematical point of view, but this paper focuses on a higher level of abstraction for policy formulation.

In any LLM, there will be three areas in which the LLM interacts with the data. The first and most significant one occurs during the training of the LLM models. Just as humans learn by observation, so do LLM models. In this process, LLMs are exposed to data from different types of data to gain a good understanding of human language. Training an LLM is a continuous process that includes fine-tuning after the initial training.

The second exposure for LLM to the dataset is during prompt by inputting data into the context window. The interaction with the content window can happen with humans interactively or with an application programming interface (API) programmatically. Data input into the context is entirely the responsibility of the end-user or the application designer, as the LLM provider cannot prevent the user from supplying data. LLM providers rely on the consent of the end user when interacting with the model, and users can provide sensitive information during the process.

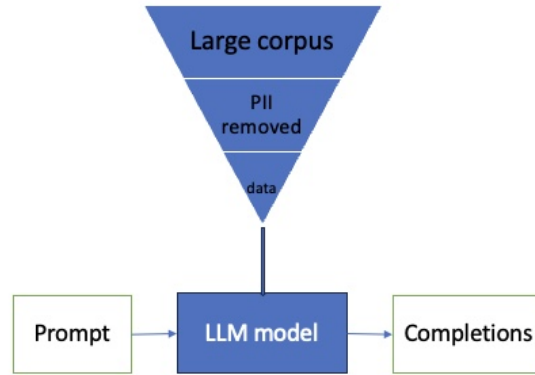


Figure 2: Large Language Model data interface.

The third type of interaction of LLM with the data is when generating an output. The primary purpose of the LLM is to generate a meaningful and contextual output for the prompt in the context window. Depending on the prompt, the LLM will produce some data or refuse, or it may be tricked into providing sensitive information.

It should be noted that LLM companies could be exposed to more data than those discussed above during user registration, payment, access to websites, and other personal identifiers. These data sets fall outside the scope of the LLM unless used for training and will not be discussed further in this paper.

## 2.1 LLM training Dataset

Most publicly available LLMs are trained on a massive corpus of data from the Internet collected by third parties, licensed and freely available contents. There is no central registry of LLM training data, but the table compiled by the Dolma paper contains the most widely used[Soldaini et al., 2024]. Another compilation is published by a team of researchers called Pile[Gao et al., 2020], which is a data set of 22 data sources with a total size of 800 GB.

Common Crawl(CC) is an online web archive run by a non-profit called Common Crawl. They have been collecting Internet data since 2007 using a CCbot crawler. It is one of the most extensive data sets in size and complexity used for machine learning. It is unknown how much PII the CCbot crawler collects, and analyzing the data set will require years of work. BookCorpus is a collection of [Bandy and Vincent, 2021]books with only the author’s email address.








Source	Doc Type	UTF-8 bytes (GB)	Documents (millions)	Unicode words (billions)	Llama tokens (billions)
Common Crawl	 web pages	9,022	3,370	1,775	2,281
The Stack	 code	1,043	210	260	411
C4	 web pages	790	364	153	198
Reddit	 social media	339	377	72	89
PeS2o	 STEM papers	268	38.8	50	70
Project Gutenberg	 books	20.4	0.056	4.0	6.0
Wikipedia, Wikibooks	 encyclopedic	16.2	6.2	3.7	4.3
<b>Total</b>		<b>11,519</b>	<b>4,367</b>	<b>2,318</b>	<b>3,059</b>

Figure 3: Dolma paper corpus data for LLM training [Soldaini et al., 2024]

The Reddit data set is the most widely used in data science research and is usually given a higher weight than the Common Crawl data set. It consists mainly of the Reddit community chat [Baumgartner et al., 2020], but the collection has stopped since the company announced a subscription system for bulk access. The Enron corpus [Noever, 2020] is a collection of a million and a half email exchanges between former Enron employees over three and a half years. Google assembled a data set called the Clean Colossal Crawled Corpus (C4) [Dodge et al., 2021], a cleaned version of the Common Crawl data set. The data set used to train LLM models contains personally identifiable information (PII) [Elazar et al., 2023] such as email address, phone numbers, IP address, etc.

## 2.2 LLM operation dataset

Large language model services are provided for both the consumer and enterprise markets, whereas latter provides better data privacy protections. Open AI, Microsoft, through an open source delivery of OpenAI, and Google provide LLM Enterprise services with the option for the data to live within the organization’s premises.

Data privacy protection provided by the LLM companies varies by the type of the service even if it is the same product. If we take the example of OpenAI services between the consumer and enterprise editions, there is a varying level of data privacy protection. The table below shows that the open version of OpenAI has a different level data privacy protection. Data inputted during prompt for the free edition while the Enterprise and the Teams editions can exclude the option. The Enterprise provides the most flexibility as it allows consumers to limit how long the input data remain, and access to this edition is subject to SOC2 audit.

# Enterprise privacy at OpenAI

Updated  
January 10, 2024

Trust and privacy are at the core of our mission at OpenAI. We're committed to privacy and security for ChatGPT Team, ChatGPT Enterprise, and our API Platform.

Privacy feature	Free	Plus	Team	Enterprise
Input data excluded from training	No	No	Yes	Yes
Can control how long input data is stored	No	No	No	Yes
Audited for SOC 2 compliance	No	No	No	Yes

Figure 4: OpenAI data privacy feature by edition  
<https://openai.com/chatgpt/enterprise>

## 3 Data privacy regulations and LLMs

In this digital world, data is a critical asset. Companies collect a huge amount of data from people's interactions with various digital products, as well as from data brokers. Data privacy is an essential aspect of digital services and requires principle-based governance. The United States currently has state-by-state data privacy regulations, while Europe follows an European Union wide rule called the General Data Protection Regulation (GDPR). Many countries around the world, such as China, Brazil, and India, have adopted some variations of the GRPR regulation.

Privacy regulations are designed to provide a set of rights to individuals and limit how their data should be processed, mainly by private companies. For example, GDPR is based on the following set of principles: purpose limitation, lawfulness, fairness and transparency, security and data minimization, individual rights, and accountability. [Kotsios et al., 2019] The same applies to other privacy regulations, as the fundamental tenet is the privacy of the individual with a common sense approach.

### 3.1 European General Data Protection Regulation

The General Data Protection Regulation (GDPR) is the first and most comprehensive data regulation in the world. It came into force in May 2018 and has been touted as one of the most successful privacy regulations.

The European GDPR outlines specific terms that are crucial for all parties involved in the design and deployment of a large-language model to understand. For example, a data controller is the person or entity that defines the purpose and means of processing the personal data of a European Union citizen. This entity carries the ultimate responsibility for data processing. On the other hand, a data processor is delegated to processes personal data on behalf of the data controller. In some cases, there may be a sub processor that receive delegated responsibility from the data processor. It is important to note that all entities in this supply chain share the responsibility to ensure the proper handling of personal data.[Degeling et al., 2019]

# OpenAI Personal Data Removal Request

Under certain privacy or data protection laws, such as the GDPR, you may have the right to object to the processing of your personal data by OpenAI's models. You can submit that request using our [Privacy Portal](#).



Figure 5: OpenAI data deletion request form <https://privacy.openai.com>

In the case of a new LLM development, the LLM company establishes the purpose of the data processing and thus acts as the data controller during the initial training of the model. As discussed in the previous section, companies sourced data from various sources, including a raw archive of internet data from a nonprofit Common Crawl(CC) foundation. The CC data set contains many sensitive and nonsensitive data sets, which the controller has to set as the basis for processing the data to build a large language model.

According to Article 6 of the GDPR, three legal bases are used to process personal data: contracts, legal interest, and consent. [Cabral, 2021] Consent and contracts are not easy to implement, as they require the direct participation of the person whose data is being processed or a company that can form a contractual agreement with the LLM vendor. Legitimate interest is the only possible source of the established purpose of data processing for data scraped from the Internet. LLM companies will need to prove that their interest in processing personal data outweighs the interest of the impacted person. Currently, most LLM companies use a legal basis as the primary consent mechanism, but the case still needs to be settled. [Novelli et al., 2024]

GDPR grants many rights to EU citizens, which are activated through data subject access requests(DSAR). One of the rights in GDPR under Article 17 is the right to be forgotten (RTBF). The right to be forgotten is a challenging concept to apply in large-language models, but it applies to LLMs from the GDPR point of view. [Zhang et al., 2023] OpenAI allows European citizens and other countries to be forgotten and to remove their data from their systems.

In addition to those listed above, GDPR grants EU citizens the right to access personal information (article 15), rectification, and correction (article 16), and to object to automated decision making. Please note that GDPR rules apply to LLM providers and thus LLM model providers should comply to operate in Europe. [Hacker et al., 2023]. It is also possible to leverage automated policy implementation for GDPR which can help in the age of Generative AI. [Amaral et al., 2021]

## 3.2 US state laws and sectorial approach

The United States took a sectorial approach ( e.g., health care, financial, education, etc.) to data privacy and took the lead in filling the void left by the lack of federal data privacy policy by introducing state-by-state comprehensive data privacy laws. As of April 20, 2024, 14 states have enacted data privacy laws, and the list is growing by month. The figure below from the International Association of Privacy Professionals(IAPP) shows the most recent update of data privacy map in the United States.

The 14 states that adopted the data privacy regulation at the time of this document are California, Virginia,Colorado, Connecticut, Delaware, Florida, Indiana, Iowa, Montana, New Jersey, Oregon, Tennessee, Utah, and Texas. Each state has some variation in the type and scope of data privacy protection granted to citizens, which will inevitably make it hard for companies to comply. However, many states have standard provisions regarding the fundamental protection of data privacy, but they fall short of matching the European GDPR rules.



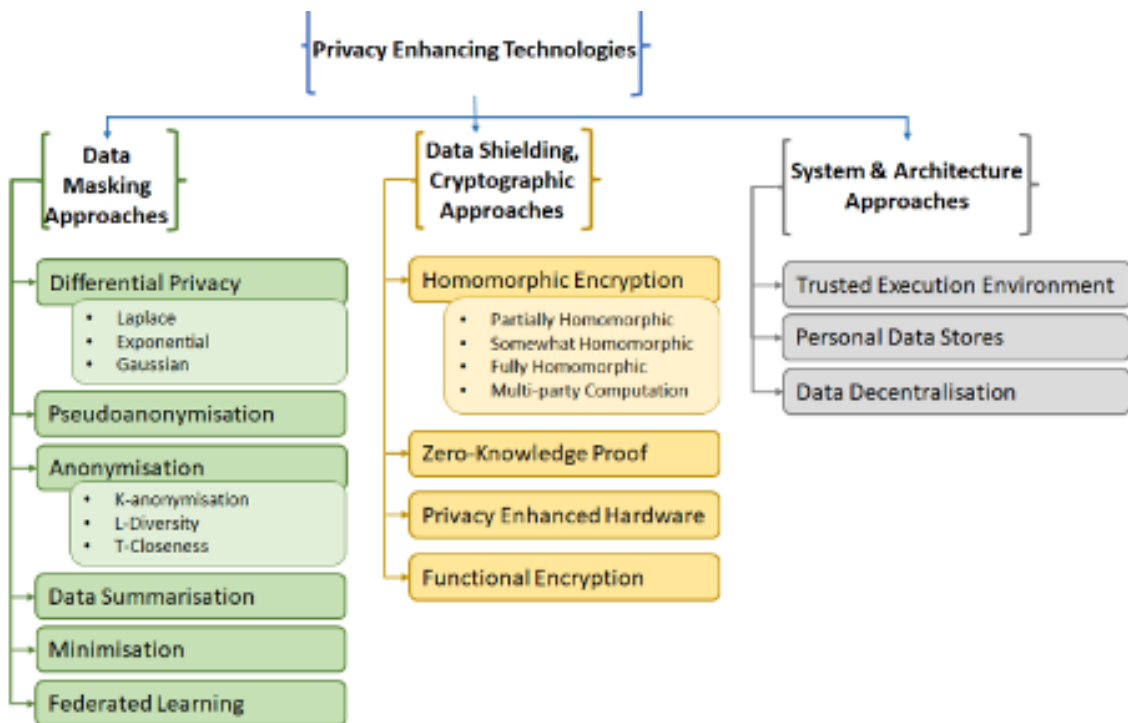


Figure 7: Categories of privacy-enhancing technologies [Ahmadi-Assalemi et al., 2022]

categories: data masking, data shielding, and cryptographic and system approaches. The figure above describes the subcomponents of each; however, the full analysis is not within the scope of this paper.

One type of PET applicable for large language models is differential privacy, which is a method that adds noise to the original data so that the source information cannot be inferred [Behnia et al., 2022] or lined back. LLMs are prone to inference attacks, and if the data set contains sensitive data that can be attributed to an individual, differential privacy can be used to minimize the risk of re-identification. In many use cases, LLMs should not be trained on personal data, but in cases where LLMs should operate in such use cases, differential privacy helps as a privacy enhancing technology.

Many other methods can minimize data leakage and maximize data privacy, such as scrubbing personally identifiable information [Lukas et al., 2023] and federated learning [Chen et al., 2023]. One method that gets the least attention is data minimization, which uses proper technology to automatically discard data once it is used for the intended purpose. If data minimization is done correctly, it will have as much impact as consent at a lower operating cost [Ullah et al., 2023].

## 4.2 LLM data privacy policy recommendation

Data privacy is a complex concept that requires a technical and non-technical solution to address adequately. The recommendation does not exclude the need for privacy-enhancing technologies, but coupling data privacy policy with privacy-enhancing technologies can achieve maximum effect.

Many open-source and private models are currently being used in production. Most of these models are benchmarked against standardized tests at some level of accuracy, but protecting personally identifiable information that they infer is not benchmarked.

Currently, there is no easy way to opt out of the various large language models especially for Citizens that do not mandate private rights of action. Some of the most advanced and well-funded companies, like Open AI and Meta, allow users to opt out of their models, but they do not cover the

many other models that are already out and will not apply retroactively. The policy recommendation I am proposing is to establish a legal framework that allows users to opt out of model training or processing their personally identifiable information.

As discussed earlier, companies use one of the three mechanisms to process personal data: legal basis, consent, and contracts. The legal basis and the contract do not require a direct interaction with the person whose data is being processed. Using a legal basis as the primary vehicle for processing is an open debate, which will leave consent as the most reliable approach to processing personal data.

The first question regarding consent is whether the opt-in versus opt-out model of policy choices is more applicable. Generally, opt-in provides the most privacy protection, as the data subject should provide affirmative consent before processing begins. This opt-in model requires companies that design and implement the LLM model to interact directly to collect opt-in, which is practically impossible. The opt-out model, on the other hand, allows for indirect communication between model developers and end-users.

The opt-out model has already been used for many privacy-related use cases. The most notable ones are the Federal Trade Commission's Do Not Call registry and the one used by credit reporting agencies to prescreen credit offers. In both cases, a centralized agency acts as a mediator between end users and service providers by providing a simple access form.

In this policy recommendation, the three credit reporting agencies (CRAs), private companies, are proposed to be the clearinghouse of opt-out, giving the consumer an option. The policy proposal consists of three sections: a centralized opt-out for the LLM training model managed by CRAs, a centralized LLM registry for those who process more than ten thousand users, and a synthetic data or anonymized data service in lieu of personal data. The proposed solution can be deployed within one year, including building the opt-out platform, establishing a data exchange schema, and educating the public about the choices they have. The figure below describes the three components labeled one, two, and three as well as the three parties at play.

#### **4.2.1 Centralized opt out from LLM models**

The first step in the opt-out consent model is to build a user profile that can serve as a registration platform. Users access the opt-out system and set preferences based on preset choices. The interaction between end users and the three CRAs delegated to handle consent and preference management is based on a better user experience, accessible to any citizen, and handles all forms of modality.

The registration portal acts as a single interface point for end users to interact with current and future language model developers, and users do not need to make frequent changes. Their choices can be time-bound ( three or five years) or permanent. The advantage of leveraging credit card processing companies as opposed to an entity set up for this is that they already hold most of the personally identifiable information of the US population. In addition, CRAs are private companies that can operate with much better efficiency and continue to update the platform with changes in technology. The other important aspect is that since there is more than one Credit Reporting Agencies, they can allow redundancy for end users. Users will only need to signal in one of the three major CRA's where the responsibility of checking across three falls under the LLM companies.

#### **4.2.2 Central registry of LLM model providers**

The second component of the policy recommendation is the registration of LLM providers, which includes those who process the personal records of more than a hundred thousand users. Registration includes signing a data processing agreement (DPA) with designated credit card reporting agencies. Few players develop large-scale models, usually limited by the resources needed to develop a model and the data size that goes along with it. This centralized registration of LLM model providers helps address challenges beyond data privacy, including approaches such as the EU in risk classification.

# Three part policy options

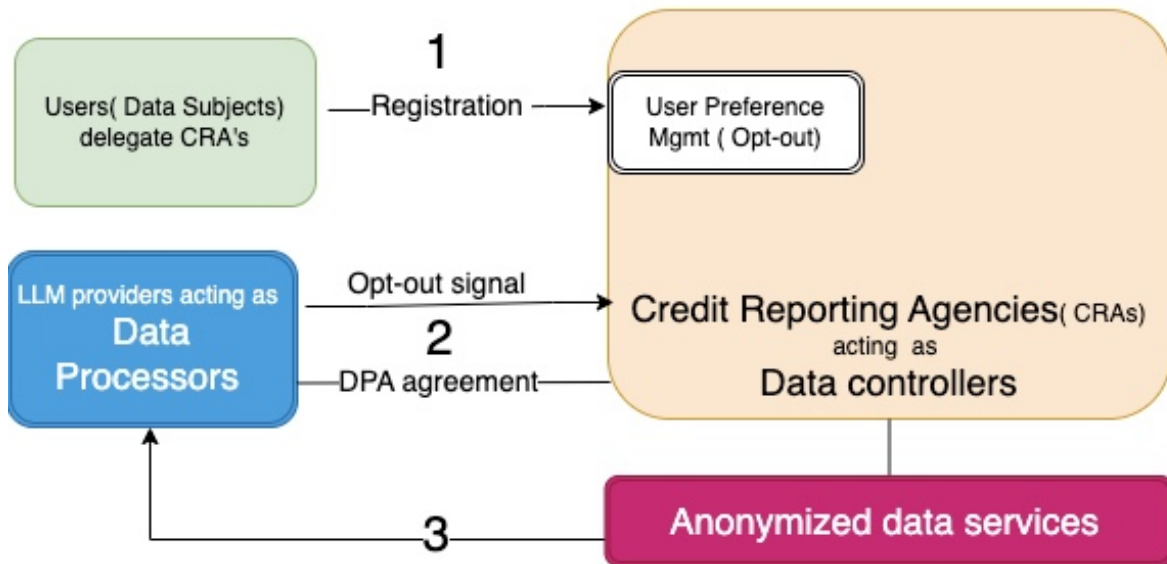


Figure 8: Three part policy options

Only registered and approved companies have the ability to collect an opt-out signal from credit card processing agencies.

Unlike the interaction between credit card processing agencies (CPAs) and end users in the previous recommendation, the interactions between LLM providers and CPAs should be fully automated using the application programming interface (API). The automated interaction needs a well-defined schema of personal and sensitive personal data.

Here is an example schema representation of personally identifiable information (PII) and sensitive PII.

```
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <!-- Personal Identifiable Information (PII) -->
  <xs:element name="PersonalIdentifiableInformation">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="Name" type="xs:string"/>
        <xs:element name="Address" type="xs:string"/>
        <xs:element name="TelephoneNumber" type="xs:string"/>
        <xs:element name="EmailAddress" type="xs:string"/>
        <xs:element name="DateOfBirth" type="xs:date"/>
        <xs:element name="SSN" type="xs:string"/>
        <xs:element name="DriverLicenseNumber" type="xs:string"/>
        <xs:element name="PassportNumber" type="xs:string"/>
        <xs:element name="EmployeeID" type="xs:string"/>
        <xs:element name="IPAddress" type="xs:string"/>
        <xs:element name="Username" type="xs:string"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
</xs:schema>
```

```

</xs:element>

<!-- Sensitive PII -->
<xs:element name="SensitivePII">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="FinancialInformation" type="FinancialInformationType"/>
      <xs:element name="MedicalRecords" type="MedicalRecordsType"/>
      <xs:element name="BiometricData" type="BiometricDataType"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>

<!--Sensitive PII categories -->
<xs:complexType name="FinancialInformationType">
  <xs:sequence>
    <xs:element name="CreditCardNumber" type="xs:string"/>
    <xs:element name="BankAccountNumber" type="xs:string"/>
  </xs:sequence>
</xs:complexType>

<xs:complexType name="MedicalRecordsType">
  <xs:sequence>
    <xs:element name="HealthHistory" type="xs:string"/>
    <xs:element name="MedicalTestResults" type="xs:string"/>
  </xs:sequence>
</xs:complexType>

<xs:complexType name="BiometricDataType">
  <xs:sequence>
    <xs:element name="Fingerprint" type="xs:string"/>
    <xs:element name="FacialRecognitionData" type="xs:string"/>
    <xs:element name="RetinaScan" type="xs:string"/>
  </xs:sequence>
</xs:complexType>

```

After the schema is defined, there is also a need for a standardized exchange of information between LLM providers and the three credit reporting agencies. The exchange can take the form of publisher and subscriber architecture using existing technologies and products with an appropriate level of security. This will be future work for those who want to implement the proposed opt-out solution.

#### 4.2.3 Anonymized and synthetic personal data services

Credit reporting agencies, banks, and the government have huge amounts of personal data that are collected for a variety of purposes. A recent study by the Government Accountability Office(GAO) shows a significant PII in the hands of the financial sector. The figure below describes the study by the GAO and its recommendations for better handling of PII. However, these same data can be used to create a synthetic information service that can be equally useful to model creators.

### 4.3 Evaluation of the proposed policy

The evaluation of the proposed policy proposal is based on the efficacy of the prior implementation of opt-out in adjacent areas. In this specific case, opt-out has been used in two key areas of data

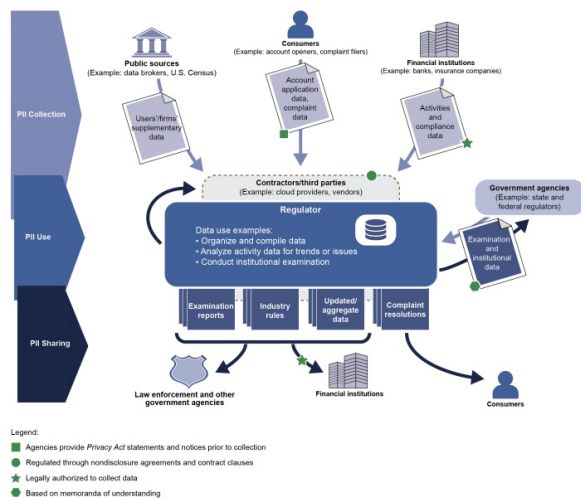


Figure 9: Personally identifiable data by financial regulators and institutions  
<https://www.gao.gov/products/gao-22-104551>

privacy. The first is the Do Not Call(DNC) Registry located at Donotcall.GOV and mandated by the FTC, which allows consumers to opt out of marketing calls. Consumers are expected to register their phones using a portal provided by the FTC, verify prior registration, and report unwanted calls should there be a violation of the DNC rule. There is a separate website for telemarketers, which provides the opt-out signal based on subscription basis.

The FTC has conducted an analysis of the DoNotCall registry and reported that it has 249 million consumers. The figure below describes the portal and the link to access the DoNotCall registry.

**FEDERAL TRADE COMMISSION**  
PROTECTING AMERICA'S CONSUMERS

Back to [ftc.gov](#) | [Español](#)

Resources | Privacy & Security | Home

## National Do Not Call Registry

[En Español](#)

Report Unwanted Calls    Verify Your Registration    Register Your Phone

**The National Do Not Call Registry gives you a choice about whether to receive telemarketing calls**

- You can [register](#) your home or mobile phone for **free**.
- After you register, **other types of organizations may still call you**, such as charities, political groups, debt collectors and surveys. To learn more, read our [FAQs](#).
- If you received an unwanted call after your number was on the National Registry for 31 days, [report it to the FTC](#).

**Sellers and telemarketers:**  
Go to <https://telemarketing.donotcall.gov> to access the National Do Not Call Registry.

Figure 10: Do not call registry from Donotcall.Gov  
<https://www.donotcall.gov/>

The second example of opt-out is credit card prescreening, in which consumers can opt out of using one of the major credit reporting agencies (CRAs). The credit card prescreening process is much more complex, as it needs to capture more data and has many players, including credit card companies, banks, and CRAs. FTC published a paper titled [The Effectiveness of Opt-Out Disclosures in Prescreened Credit Card Offers](#) showing that the opt-out method works.

The other factor in evaluating the efficacy of the proposed policy is if there are early indications of solving the problem in a specific segment or by leading vendors. In this regard, Meta is already experimenting with an opt-out [Generative AI data subject access request](#) that allows an opt-out from the Meta LLM model. Meta is a significant player in the LLM space with its widely adopted LaLama model; thus, it has the opportunity to see the problem ahead of others.

←
Generative AI Data Subject Rights

---

**Generative AI Data Subject Rights**

This form is where you can submit requests related to your personal information from third parties being used to train Meta's generative AI models.

In general, personal information is information about you. Examples include your name, home address, phone number or email address.

Information from third parties includes what is publicly available on the internet and licensed information that someone else owns and gives Meta permission to use. This information does not include information from Meta or our products and services.

Generative AI models are computer programs. They use predictions and patterns to create new content. To be able to spot these patterns, models are trained on billions of pieces of data from a variety of sources. These sources may include third parties. When we do get personal information as part of this data that we use to train our models, we don't specifically link this data to any Meta account.

You should only use this form for personal information from third parties being used to train Meta's generative AI models. To learn more about your rights related to information from Meta's products or services, visit our [Privacy Policy](#).

We don't automatically fulfill requests sent using this form. We review them consistent with your local laws. Enter your details below to submit your request.

What best describes your request?

I want to access, download or correct any personal information from third parties used for training Meta's generative AI models

I want to delete any personal information from third parties used for training Meta's generative AI models

I have a concern about my personal information from third parties that is related to a response I received from Meta's generative AI models

I have a different issue

Figure 11: Meta Generative AI data subject access rights  
<https://www.facebook.com/help/contact/1266025207620918>

OpenAI provides the right-to-be-forgotten capability to countries under the GDPR umbrella and signs data processing agreements ( DPA) with enterprise customers. The proposed solution extends the data protection provided to Enterprise customers to consumers, with the three CRAs acting on behalf of the consumers. A comprehensive policy of OpenAI for Enterprise is attached to the appendix of this paper.

In summary, the opt-out method has already been tested on a large scale, but it needs legislative support to provide it as a central service to consumers.

#### 4.4 Limitation of the proposed solution

The proposed solution assumes that consumers whose data are being processed by the large language model developer acting as a data controller and the consumer whose data is being processed participate in the opt-out platform as a primary means of communication. This approach may cause consumers who need help to interact with web-based services. An approach suggested is to leverage community services that already serve the same population in similar cases like accessing health extension sites.

The second limitation is that opt-out is good only when accompanied by other solutions like data minimization, which mandates the life span of the data, and an additional opt-in when sensitive data are transferred. Policy solutions that look to implement opt-out should put additional guard rails in place to prevent abuse of the permission given.

## 5 Future work and the U.S. draft privacy legislation

Large language models are here to stay and have already revolutionized the way humans communicate with intelligent bots. A good comprehension of English is needed to converse with chatbots instead of a highly scripted dialog that can answer only what was programmed. Large language models are trained in a data set that is collected for a purpose other than training a model and may not discriminate personally identifiable data from other data sets. This has caused an undue burden on data preparation teams to clean up PII before it makes it to the LLM. The future of LLMs will start by setting the data collection purpose, and thus, vendors may collect the source data by setting up their own crawler bots or licenses from the data owners. Closed models will be scrutinized closely using the research API, which can be mandated or done voluntarily.

### 5.1 American privacy rights act

On 7 April 2024, the United States Congress introduced [a new draft bill](#) to address data privacy after a failed attempt in 2022. The new data privacy act, the American Privacy Rights Act (APRA), is a comprehensive draft federal regulation that preempts most state-level privacy regulations. The APRA introduces new terminology and approaches to address data privacy, including consent and data minimization. APRA exempts the government, government contractors, and small businesses. The FTC is recommended as the lead agency in implementing the APRA.

According to the APRA draft regulation, a covered entity is an organization that determines the purpose of collecting and processing personal data. A covered entity is analogous to a data controller under GDPR. A small business is a company that has less than 40 million dollars in revenue and is exempt from the APRA compliance requirement. A company that makes more than 40 million is a data holder, and this will be a covered entity. A company with a revenue of more than 250 million is considered a large data holder.

Covered data and sensitive covered data are two data labeling methods for identifying and enumerating personal data applicable to APRA coverage. The ARPA analogy is similar to that used for the health care data privacy, which has undergone extensive adoption.

APRA recommends a centralized opt-out to transfer non-sensitive covered data, algorithmic decisions, and targeted advertisements. It calls for an opt-in for the transfer of covered sensitive data. This is, in fact, in line with my recommendation, which I developed independently of the ARPA draft. The draft also calls for FTC to produce a mechanism for centralized, for which I recommend an XML-based schema and a high-level architecture.

As part of future work, I highly recommend developing a centralized opt-out technical solution and, if possible, collaborating with the FTC.

## References

- [Ahmadi-Assalemi et al., 2022] Ahmadi-Assalemi, G., Al-Khateeb, H., and Aggoun, A. (2022). Privacy-enhancing technologies in the design of digital twins for smart cities. *Network Security*, 2022:Not available.
- [Al-Amin et al., 2024] Al-Amin, M., Ali, M. S., Salam, A., Khan, A., Ali, A., Ullah, A., Alam, M. N., and Chowdhury, S. K. (2024). History of generative artificial intelligence (ai) chatbots: past, present, and future development.
- [Amaral et al., 2021] Amaral, O., Abualhaija, S., Torre, D., Sabetzadeh, M., and Briand, L. C. (2021). Ai-enabled automation for completeness checking of privacy policies.
- [Bandy and Vincent, 2021] Bandy, J. and Vincent, N. (2021). Addressing "documentation debt" in machine learning research: A retrospective datasheet for bookcorpus.
- [Baumgartner et al., 2020] Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. (2020). The pushshift reddit dataset.
- [Behnia et al., 2022] Behnia, R., Ebrahimi, M. R., Pacheco, J., and Padmanabhan, B. (2022). Ew-tune: A framework for privately fine-tuning large language models with differential privacy. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE.
- [Cabral, 2021] Cabral, T. S. (2021). Ai and the right to explanation: Three legal bases under the gdpr. *Data Protection and Privacy*, Not available:29–56.
- [Chen et al., 2023] Chen, C., Feng, X., Zhou, J., Yin, J., and Zheng, X. (2023). Federated large language model: A position paper.
- [Degeling et al., 2019] Degeling, M., Utz, C., Lentzsch, C., Hosseini, H., Schaub, F., and Holz, T. (2019). We value your privacy ... now take some cookies: Measuring the gdpr's impact on web privacy. In *Proceedings 2019 Network and Distributed System Security Symposium*, NDSS 2019. Internet Society.
- [Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.
- [Dodge et al., 2021] Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., and Gardner, M. (2021). Documenting large webtext corpora: A case study on the colossal clean crawled corpus.
- [Elazar et al., 2023] Elazar, Y., Bhagia, A., Magnusson, I., Ravichander, A., Schwenk, D., Suhr, A., Walsh, P., Groeneveld, D., Soldaini, L., Singh, S., Hajishirzi, H., Smith, N. A., and Dodge, J. (2023). What's in my big data?
- [Gao et al., 2020] Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. (2020). The pile: An 800gb dataset of diverse text for language modeling.
- [Hacker et al., 2023] Hacker, P., Engel, A., and Mauer, M. (2023). Regulating chatgpt and other large generative ai models.
- [Jia et al., 2023] Jia, J., Liu, H., and Gong, N. Z. (2023). 10 security and privacy problems in large foundation models.
- [Kotsios et al., 2019] Kotsios, A., Magnani, M., Rossi, L., Shklovski, I., and Vega, D. (2019). An analysis of the consequences of the general data protection regulation (gdpr) on social network research.
- [Lukas et al., 2023] Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., and Zanella-Béguelin, S. (2023). Analyzing leakage of personally identifiable information in language models.

- [Neel and Chang, 2024] Neel, S. and Chang, P. (2024). Privacy issues in large language models: A survey.
- [Noever, 2020] Noever, D. (2020). The enron corpus: Where the email bodies are buried?
- [Novelli et al., 2024] Novelli, C., Casolari, F., Hacker, P., Spedicato, G., and Floridi, L. (2024). Generative ai in eu law: Liability, privacy, intellectual property, and cybersecurity.
- [Raffel et al., 2023] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2023). Exploring the limits of transfer learning with a unified text-to-text transformer.
- [Soldaini et al., 2024] Soldaini, L., Kinney, R., Bhagia, A., Schwenk, D., Atkinson, D., Authur, R., Bogin, B., Chandu, K., Dumas, J., Elazar, Y., Hofmann, V., Jha, A. H., Kumar, S., Lucy, L., Lyu, X., Lambert, N., Magnusson, I., Morrison, J., Muennighoff, N., Naik, A., Nam, C., Peters, M. E., Ravichander, A., Richardson, K., Shen, Z., Strubell, E., Subramani, N., Tafjord, O., Walsh, P., Zettlemoyer, L., Smith, N. A., Hajishirzi, H., Beltagy, I., Groeneveld, D., Dodge, J., and Lo, K. (2024). Dolma: an open corpus of three trillion tokens for language model pretraining research.
- [Sovrano et al., 2021] Sovrano, F., Sapienza, S., Palmirani, M., and Vitali, F. (2021). *A Survey on Methods and Metrics for the Assessment of Explainability Under the Proposed AI Act*. IOS Press.
- [Staab et al., 2023] Staab, R., Vero, M., Balunović, M., and Vechev, M. (2023). Beyond memorization: Violating privacy via inference with large language models.
- [Ullah et al., 2023] Ullah, I., Hassan, N., Gill, S. S., Suleiman, B., Ahanger, T. A., Shah, Z., Qadir, J., and Kanhere, S. S. (2023). Privacy preserving large language models: Chatgpt case study based vision and framework.
- [Vaswani et al., 2023] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need.
- [Veale and Zuiderveen Borgesius, 2021] Veale, M. and Zuiderveen Borgesius, F. (2021). Demystifying the draft eu artificial intelligence act — analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, 22(4):97–112.
- [Walters et al., 2023] Walters, J., Dey, D., Bhaumik, D., and Horsman, S. (2023). Complying with the eu ai act.
- [Yao et al., 2024] Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., and Zhang, Y. (2024). A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 4(2):100211.
- [Zhang et al., 2023] Zhang, D., Finckenberg-Broman, P., Hoang, T., Pan, S., Xing, Z., Staples, M., and Xu, X. (2023). Right to be forgotten in the era of large language models: Implications, challenges, and solutions.