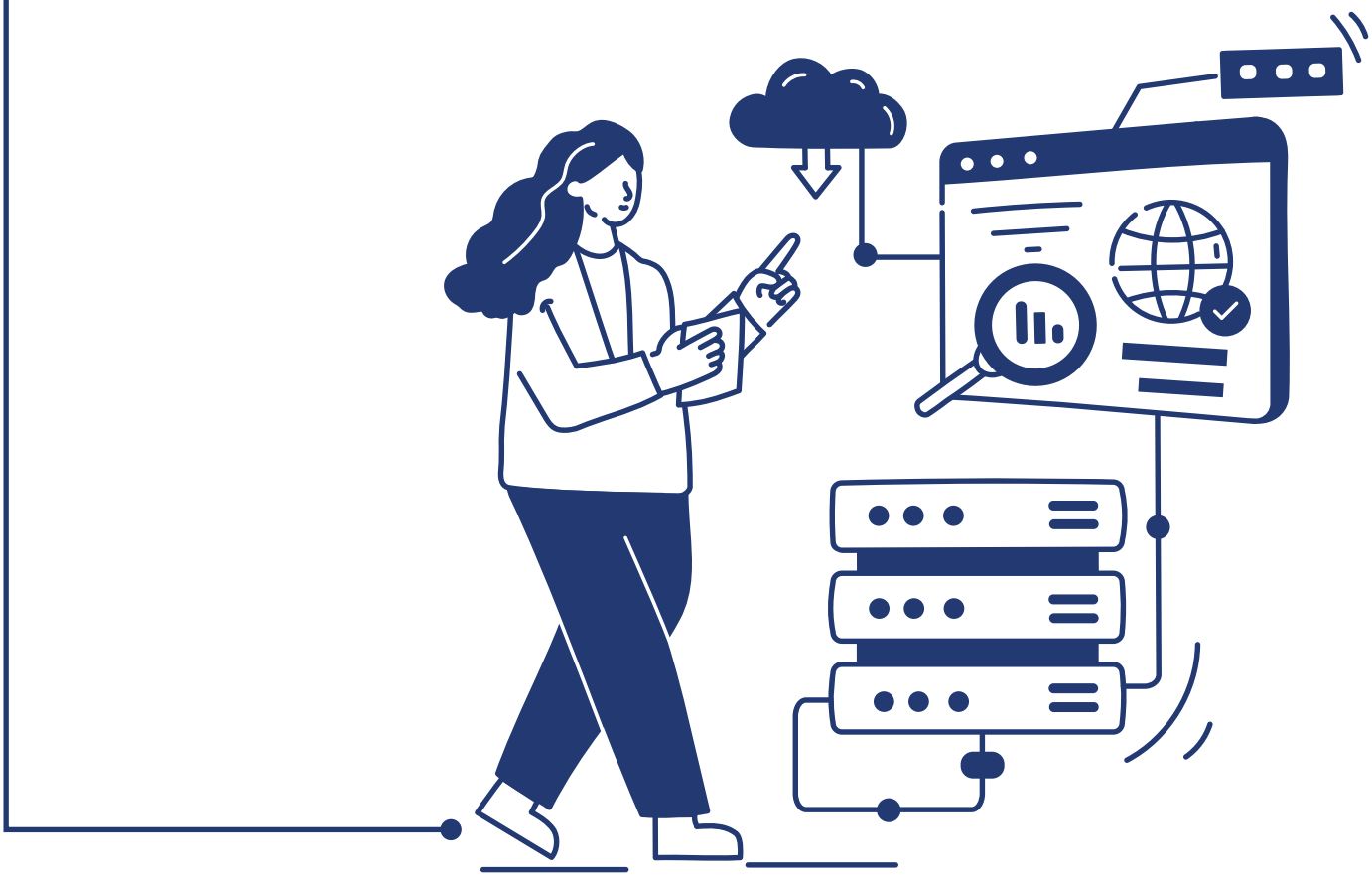
DSCI
PROMOTING DATA PROTECTION
A nasscom Initiative

EXPLORATORY NOTE ON

PRIVACY, DATA PROTECTION, AND LARGE LANGUAGE MODELS



Objective

This exploratory note is intended to map out the functioning of Large Language Models (LLMs), which form the basis of some of the popular generative AI services, against key data protection requirements globally. The note takes a law-agnostic approach and maps out the functioning of LLMs against conceptual elements which are common to most data protection regulations. Given that this is a rapidly evolving space, this document is inherently dynamic and future iterations may outline other risks and include recommendations for addressing the same.

Contributors

- Shivangi Malhotra
- Devanshi Singh

Table of CONTENTS

I

Data Protection Concerns Arising from Publicly Available Large Language Models

Large Language Models (LLMs) refer to algorithms that can perform several tasks, such as generating texts and other content, based on the learnings from massive datasets. These datasets usually consist of large troves of information taken from the internet, which runs the risk of including personal data within the algorithm's training data set. With the massive increase in accessibility and adoption of these large language models, there are several data protection and privacy-related concerns that arise.

This exploratory note aims to shed light on some of these concerns by mapping the operation and functionality of Large Language Models against some of the common principle-level requirements of data protection laws globally. Our endeavour is to share more nuanced perspectives on the privacy, data protection, and security considerations that may arise in the future with the rapid advancement of artificial intelligence in the industry.

1. Training of LLMs on 'publicly available' personal data

Contours of 'Publicly Available' Data: To what extent do data protection regulations apply?

Developers of Large Language Models state that these models are developed based on 'publicly available' data, which may include personal data. It's also important to note that personal data used in for training may not necessarily have been made publicly available by data subject themselves. For instance, personal data made available on the internet because of a data breach may also be included in the training data set. These models are typically trained on massive datasets- for instance in case of GPT-3 it was 60 million domains.It is therefore important to understand how, and to what extent, data protection regulations apply to 'publicly available' personal data.

- **GDPR (European Union):** Under the European Union's GDPR, Article 14 requires that data subjects be notified of certain information where their personal data has not been obtained from them. Such notifications should include information such as what personal data is processed, its source, how long it will be retained, etc. This would also include scenarios where personal data has been sourced from publicly available databases or other resources. This is exemplified in multiple case decisions in the EU. For instance, in 2019, the Polish data protection authority fined a company for failing to provide a privacy notice to individuals whose personal data it had scraped off public databases

  Additionally, under Article 9(2)(e), there is a general prohibition to process 'special categories of personal data', however when such personal data is manifestly made public by the data subject, this prohibition does not apply.

- **Draft Digital Personal Data Protection Bill (India):** In India, the draft DPDPB permits processing of personal data for a lawful purpose, based on either consent given by the data principal or when consent is deemed to have been given. It can therefore be concluded that the general rule for processing of publicly available personal data is that consent of the data principal is required for the same. However, where publicly available personal data is processed 'in public interest', consent of the data principal will be deemed to have been given.

- **Personal Information Protection and Electronic Documents Act (Canada):** In Canada, Section 7(1)(d), PIPEDA stipulates that publicly available personal information may be collected, used or disclosed by an organization without the consent of individuals, only if it falls within the regulated categories of telephone directories, professional or business directories, court and tribunal records, and books, magazines and newspapers. It was further clarified in 2009 by the OPC that the collection, use or disclosure of such information must also relate directly to the purpose for which it appeared publicly. Otherwise, the specific knowledge and consent of individuals is a prerequisite for processing of personal data found publicly from any other source.

What kind of personal data do LLMs use?

- **In training data sets:** Millions of websites are used to train LLMs. For instance, the GPT-3 model was reportedly trained on 570GB of data scraped from the internet, through sources like books, texts, articles, etc. GPT-4's training set, admittedly, 'may' include publicly available personal information, and by combining capabilities of the model, "GPT-4 has the potential to be used to attempt to identify individuals when augmented with outside data."

  Independent investigations into the sources of public LLM's data sets reveal information has been taken from over 15 million websites including general informative sites, piracy sites, subscription-only sites, private voter database sites, political and religious sites. This indicates the possibility of sensitive

personal data such as political opinions or religious beliefs being included in the training data sets of public-facing LLMs. The training data sets for such natural-language processing models can also include data scraped from public dialog sites or forums such as Reddit, Facebook, Quora, etc.

- **Personal data collected and processed through end-users:** While a part of the concern stems from use of personal data in the original training data set, the personal data collected from end users also raises some concerns. Particularly in scenarios where prompts by users may contain personal information, which is then used for refining and training the model. For instance, OpenAI's privacy policy states that personal information may be collected through inputs and file uploads which may then be used to improve existing services, to develop new services, and to train the models that power ChatGPT. More recently, a feature has been introduced to allow end users from disabling their conversations being used for training of the algorithm. However, this is an opt-out feature, and by default, users' conversations are used to train the model.

  Tools like 'ProfileGPT' have emerged to analyse and summarise the information that is collected about individual users by large language models. For instance, the creators of ProfileGPT claim that users' interactions with ChatGPT may enable extraction of personal data related to the user's life summary, their hobbies and interests, political/religious views, mental health, etc. More generally, in big data analytics, it is well-settled that it is not just the data provided by individuals which can be used for analysis, but also observed data, derived data, and inferred data.

  Concerns have also been raised about the possibility of hackers 'poisoning' the dataset to create security vulnerabilities, which can then be exploited to extract sensitive personal data.

Mechanisms used by developers to protect Personal Data vis-à-vis LLMs

- **Prohibiting web scraping exercises:** The terms of use of Facebook, LinkedIn & Twitter prohibit web scraping of the personal data available on their platforms. Some research argues that language models should be trained on data that has been explicitly produced for public use, instead of scraping publicly available data.

- To remove PII, and train AI to remove other unwanted features, developers use **filters** and carry out Reinforced Learning from Human Feedback (RLHF).

- The use of masking/anonymisation or manually removing PII has been suggested as a fix, but it does not address the issue of collection, purpose limitation and use itself.

- Data redaction and use of synthetic PII: Two possible solutions to address some of the privacy and data protection concerns arising out of use of LLMs are data redaction and the use of contextually accurate synthetic data.However, data redaction of unstructured data at scale is difficult at scale, while manually redacting data is slow and expensive, as well as often inaccurate.

- Developers of LLMs state some of the measures currently being undertaken by them to address privacy and data protection issues in this context. For instance, OpenAI states that it fine-tunes models to reject certain types of requests, removes personal information from the training dataset where feasible, creates automated model evaluations, monitoring and responding to user attempts to generate personal information, and restricts this type of use in our terms and policies. Another measure undertaken for this purpose is the use of human raters to rate personally identifiable information as unsafe.

- Children's safety: In Italy, one of the reasons cited by the Italian Data Protection Authority, Garante, for its temporary ban on ChatGPT in the country was the absence of any age verification requirements for accessing the services of ChatGPT. In response to the same, OpenAI introduced new requirements for Italian registered users, requiring self-declaration of age and confirmation of parental consent for users between the ages of 13 and 18 years.

2. Application of Data Protection Principles

The processing of personal data by LLMs imputes the responsibility upon the developers of the LLM to comply with the personal data protection principles of legality, purpose and storage limitation, transparency, data minimisation, accuracy, and integrity, accepted as the cross-jurisdictional standard.

Purpose Limitation

- To adhere with the principle of purpose limitation, all personal data processed by LLMs must be within the sphere of specific delineated objectives, clearly stated, and adhered to for the scope and duration of data collection and processing. For Large Language Models, which are often used as the foundational models for several use cases across industry, predetermining and defining all purposes for which such models may be used is not always feasible. The broad scope nature of tasks for which foundational models can be adapted, leaves little scope for reliably limiting its purposes.

For Large Language Models, which are often used as the foundational models for several use cases across industry, predetermining and defining all purposes for which such models may be used is not always feasible.

- Further, in the case of non-API use of LLMs, data is collected and used not just in compiling the training data sets for the language models, but also through user prompts. Ostensibly, developers use it for research, marketing and advertising, product development and improvement of services. However, repurposing of prompts for analysis and improvement training may be violative of purpose limitation, as conceptualised under the GDPR, since the EDPB has previously held that service improvement is not a legitimate basis necessary for 'core' processing under Article 6, GDPR.

- It is also important to consider that LLM developers, such as OpenAI, may be able to limit the purposes for which their services are used to some extent. For instance, OpenAI's privacy policy states that it uses personal data only for the purposes of providing and improving services, communication with users, maintaining security, compliance with legal obligations, research and similar purposes. However, in the terms of use for its API version, developers may only be able to enumerate purposes for which their language models may not be used, i.e. prohibiting deployers from using the language models in a manner which violates any person's rights, or to develop competing products and services, etc. It may not be practically feasible for developers to ensure adherence of purpose limitation principle by the downstream users of their language models.

Storage Limitation

The principle of storage limitation requires that organisations only retain personal data for as long as it is required to fulfil the purposes for which it was collected. However, as established in the above section, in case of foundational models such as large language models, it is not always possible to limit and define the purposes for which the service will be used. Therefore, as such, there is no temporal limit on how long personal data can be retained, as the business purposes may continue to exist in perpetuity.

For instance, OpenAI's privacy policy states that personal data will be retained for as it is needed to provide services to the users or for other legitimate business purposes. As such there is no defined time for how long a user's personal data may be retained. Additionally, there is also the question around application of the storage limitation principle vis-à-vis personal data contained in training data sets, as this personal data can arguably be stored perpetually, unless an individual exercises their right to deletion of personal data.

Transparency and Fairness

The principle of transparency and fairness in data protection requires that personal data is not processed in a manner which is detrimental to the interests of the data principal and there is transparency about why the personal data is being collected

and the purposes for which it will be used. Generally, there are some concerns about the 'black-box' algorithms that power these LLMs as there is not a complete understanding of their functioning. Additionally, there is a lack of transparency about the sources from which data is taken for forming the training data sets for these large language models. Some recommendations that may be considered for improving fairness and transparency in the functioning of LLMs include the documentation of data sources, and ensuring that the characteristics of data taken to train the AI are identified, documented, and justified.

Data Minimisation

The principle of data minimisation requires that data controllers that collection of personal data is limited to what is necessary and relevant to fulfil stated purposes. However, with massive training data sets being used to train Large Language Models, this principle is difficult to achieve. At the pre-training stage of these language models, some technological interventions may make it possible to identify, detect, and mask/remove personally identifiable information, thereby facilitating data minimisation to some extent.

3. Entity's Relationship with Personal Data

Traditional conceptual notions of data controller/fiduciary and data processor may be challenged in the context of large language models. Broader concerns surrounding the determination of personal data ownership and control are key focus areas in the context of large language models.

- **Collection of personal data for training data sets:** Developers of LLMs may argue they are not data controllers because the personal data used in the original training dataset, its collection and the purposes for which it was collected and the output were not determined by the developer of the AI model, e.g. OpenAI.

  However, it may also not be feasible to categorise them as data processors. Data processors are generally understood as entities which process personal data on behalf of the data controller/fiduciary, and such processing takes place based on a contractual relationship. Both these elements are generally absent in this context, making it difficult to qualify developers of large language models as 'data processors'.

  In the context of API use of large language models, LLM developers may be considered as the processor, processing data on behalf of the controller, i.e. the enterprise.

- **Processing of personal data of end users:** In the context of personal data that LLM developers collect directly from end users, such as through account information, or user content, these entities may qualify as data fiduciaries as they determine the means and purposes for processing of such personal data.

4. Exercising Data Subject Rights

Generally, the exercise of data subjects' rights is contingent upon the jurisdiction from which they are accessing the services of a public-facing language model. In thinking about the data protection rights of data subjects, it is important to consider the rights of both sets of individuals; end-users of the language models as well as individuals (i.e. non-users) whose personal data may be a part of the original training data set.

- **Right to access/rectification:**Practically, it is not possible for individuals to know if their data was part of the training data set, unless a data subject rights request is made to the developer of the LLM. For instance, OpenAI allows individuals to file a request to access or rectify their personal data. However, there is some discretion to refuse such requests.
- **Right to erasure:** For end users, one of the ways to ensure deletion or erasure of their personal data is to delete their accounts on the service provider's platform. OpenAI also permits individuals in certain jurisdictions to fill out a form to exercise their right to request deletion of their personal data. More recently, the organisation also announced some new features to allow users to exercise more control over their personal data, this includes: 1. enabling users to turn off chat history to ensure that user inputs are not used to train the language models, and 2. ensuring that data shared by enterprises using the API version of the service users' data is not used to train models by default.
- **Data portability:** For end users, OpenAI permits exporting of personal data, including exporting of complete prompt/response history and data.

5. Subject Matters of Privacy

Lawful Basis for Processing

LLM developers may rely on a number of bases as the justification for data processing activities. This can include the following; contractual obligation, legitimate interest (in conducting research, developing new services etc.), consent (in the case of data shared by users in utilising the service), etc. For claiming legitimate interest as the lawful basis for processing, it is important to weigh the interests of data subjects against the interests of the data controller and take into account any adverse impact on the data subjects.

Notice and Consent

As such, notice and consent requirements are not replicable in the context of scraping of personal data from publicly available sources for training language models. Given the magnitude of data that is scraped and processed for developing and training LLMs, there may be a need to explore mechanisms other than a notice and consent framework to achieve the goals of greater transparency and accountability. This is because notice-and-consent mechanisms may not be technically feasible in the context of LLMs.

II

Enterprise use of Large Language Models

The previous section of this note highlighted general data protection concerns arising from public use of large language models and most of these concerns would also reflect in the enterprise use of LLMs. However, private or enterprise use of LLMs adds another layer of complexity to the questions surrounding application of data protection principles in this context. Some of these complexities are highlighted below.

- **API use of Large Language Models and affixing responsibility on entities**: Enterprises using APIs, for instance the ChatGPT API, and integrating the same into their services, are arguably acting as data controllers, as they define the purposes for which personal data is processed. When using OpenAI services via the OpenAI API platform, an order processing relationship is established between the data controller as the client and OpenAI as the processor. For these purposes, OpenAI provides a Data Processing Agreement. However, it is also conceivable that the developer of the language model and the enterprise user be treated as joint controllers. OpenAI currently does not yet provide a template for a contract pursuant to Art. 26 GDPR for establishing a joint controller relationship. An enterprise integrating LLM APIs into its product and services offerings is arguably acting as a data controller. However, it is also conceivable that the LLM developer and enterprise user act as joint controllers.

An enterprise integrating LLM APIs into its product and services offerings is arguably acting as a data controller. However, it is also conceivable that the LLM developer and enterprise user act as joint controllers.

- Affixing accountability and responsibility along the AI value chain can be difficult to navigate as the relationships between developers and deployers are complex. For instance, when a large language model is deployed by an enterprise, which then modifies it for its own internal use, it may be difficult to pin down which performance issues are to be attributed to the enterprise (i.e. the deployer) and which are to be attributed to the developer of the language model.

- **Risk management:** Enterprise use of large language models, at scale, increases the risks of leakage of proprietary information, inadvertent de-anonymisation and breach of PII, training data poisoning, unauthorized access to the data inside the organization, bias and inaccuracy, training data extraction attacks, etc.
- **Employee misuse:** Enterprise use of large language models carries with it the risk of employees entering personal identifying information into the language model as prompts. While, the language model may not retain this information, it does 'learn' from it. It is also difficult for security applications to monitor data that employees input as prompts, as most security products are designed to safeguard files, and not necessarily safeguard the information contained in the files from being copied into a browser window. These scenarios raise concerns about not just sharing of personal information, but also confidential information of enterprises. To address some of these challenges, organisations like Morgan Stanley have put in place mechanisms to ensure granular fine tuning of language models, curated document bases, imposed restrictions on types and number of prompts to be entered into system, and started conducting weekly audits.

Ultimately, in the context of generative artificial intelligence and large language models, a number of data protection principles, and foundational concepts may require a thorough re-examination. Firstly, there are considerations surrounding the complexity involved in determining the roles and corresponding accountabilities of each stakeholder involved in the processes from training and development of algorithms to deployment for various purposes across different use cases in the industry. Secondly, depending on the use case, the risk attributable to the large language model may also vary. Finally, the frameworks developed around exercise of data principals' rights, notice and consent requirements, etc. may not be exactly replicable in this context, because of a lack of technical feasibility. It is therefore pertinent to accommodate for these considerations as India thinks about enacting regulations for the digital realm.

References

1. What are large language models used for? https://blogs.nvidia.com/blog/2023/01/26/what-are-large-language-models-used-for/
2. Training Data used to train LLM models https://indiaai.gov.in/article/training-data-used-to-train-llm-models
3. https://openai.com/blog/our-approach-to-ai-safety
4. https://techcrunch.com/2020/08/07/here-are-a-few-ways-gpt-3-can-go-wrong/
5. https://iapp.org/news/a/publicly-available-data-under-gdpr-main-considerations/
6. https://www.technologylawdispatch.com/2019/04/privacy-data-protection/processing-publically-available-personal-data-without-telling-data-subjects-the-polish-data-protection-authority-has-bad-news-for-you/
7. Clause 5, draft Digital Personal Data Protection Bill,https://www.meity.gov.in/writereaddata/files/The%20Digital%20Personal%20Data%20Potection%20Bill%2C%202022_0.pdf
8. Clause 8(8)(f), draft Digital Personal Data Protection Bill, https://www.meity.gov.in/writereaddata/files/The%20Digital%20Personal%20Data%20Potection%20Bill%2C%202022_0.pdf
9. Section 1, Regulations Specifying Publicly Available Information, SOR/2001-7 [CA] http://www.canlii.org/en/ca/laws/regu/sor-2001-7/latest/sor-2001-7.html .
10. PIPEDA Case Summary #2009-020 https://www.priv.gc.ca/en/opc-actions-and-decisions/investigations/investigations-into-businesses/2009/pipeda-2009-020/ ; PIPEDA Case Summary #2010-004 https://www.priv.gc.ca/en/opc-actions-and-decisions/investigations/investigations-into-businesses/2010/pipeda-2010-004/ .
11. Office of Privacy Commr of Canada, 'Interpretation Bulletin: Publicly Available information https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/pipeda-compliance-help/pipeda-interpretation-bulletins/interpretations_06_pai/
12. https://www.sciencefocus.com/future-technology/gpt-3/
13. https://cdn.openai.com/papers/gpt-4.pdf#page=53
14. https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/
15. https://www.searchenginejournal.com/google-bard-training-data/478941/#close
16. https://openai.com/policies/privacy-policy
17. https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance
18. https://github.com/sahbic/profile-gpt
19. https://sahbichaieb.com/profilegpt/
20. Information Commissioner's Office (Big Data, Artificial Intelligence, Machine Learning and data protection) https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf
21. Melissa Heikkila, "What does GPT-3 know about me?" https://www.technologyreview.com/2022/08/31/1058800/what-does-gpt-3-know-about-me/
22. https://www.facebook.com/legal/terms
23. https://www.linkedin.com/legal/user-agreement
24. https://twitter.com/en/tos
25. What does it meanfor a language model to preserve privacy? https://arxiv.org/pdf/2202.05520.pdf
26. https://cdn.openai.com/papers/gpt-4.pdf

27. https://wandb.ai/wandb_gen/llm-data-processing/reports/Processing-Data-for-Large-Language-Models--VmlldzozMDg4MTM2#personal-identifiable-information-control:~:text=Depending%20on%20the%20application%20it%27s%20important%20to%20either%20mask%20or%20removes%20such%20information%20before%20pre%2Dtraining%20language%20models; https://fastdatascience.com/sensitive-data-machine-learning-model/
28. Article 29WP, Opinion 05/2014 on Anonymisation Techniques (2014) https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf p 10, 11.
29. https://www.private-ai.com/2023/01/18/addressing-privacy-and-the-gdpr-in-chatgpt-and-large-language-models/
30. GPT-4 Technical Report, https://cdn.openai.com/papers/gpt-4.pdf#page=53
31. https://www.technologyreview.com/2022/08/31/1058800/what-does-gpt-3-know-about-me/
32. https://www.gpdp.it/web/guest/home/docweb/-/docweb-display/docweb/9870847#english
33. https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9881490#english
34. https://www.ohchr.org/en/documents/thematic-reports/a77196-principles-underpinning-privacy-and-protection-personal-data
35. https://blogs.nvidia.com/blog/2023/03/13/what-are-foundation-models/ ; https://ai.facebook.com/blog/large-language-model-llama-meta-ai/
36. Binding Decision 5/2022 on the dispute submitted by the Irish SA regarding WhatsApp Ireland Limited (Art. 65 GDPR)[para 51, 52] https://edpb.europa.eu/system/files/2023-01/edpb_bindingdecision_202205_ie_sa_whatsapp_en.pdf
37. https://openai.com/policies/terms-of-use
38. https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/storage-limitation/ ; https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Advisory-Guidelines/the-retention-limitation-obligation---ch-18-(270717).pdf
39. https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/lawfulness-fairness-and-transparency/
40. https://wandb.ai/wandb_gen/llm-data-processing/reports/Processing-Data-for-Large-Language-Models--VmlldzozMDg4MTM2#documentation
41. https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/data-minimisation/
42. https://wandb.ai/wandb_gen/llm-data-processing/reports/Processing-Data-for-Large-Language-Models--VmlldzozMDg4MTM2#personal-identifiable-information-control
43. Clause 2(7), draft Digital Personal Data Protection Bill 2022
44. Clause 9(9), draft Digital Personal Data Protection Bill 2022
45. https://openai.com/policies/privacy-policy
46. https://openai.com/policies/privacy-policy
47. https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed
48. https://share.hsforms.com/1UPy6xqxZSEqTrGDh4ywo_g4sk30
49. https://openai.com/blog/new-ways-to-manage-your-data-in-chatgpt
50. https://help.openai.com/en/articles/7260999-how-do-i-export-my-chatgpt-history-and-data
51. https://www.linkedin.com/posts/alexanderhanff_openai-dsar-response-activity-7054052957578833922-rYvJ
52. https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/legitimate-interests/how-do-we-apply-legitimate-interests-in-practice/
53. https://www.brookings.edu/research/protecting-privacy-in-an-ai-driven-world/
54. https://openai.com/policies/data-processing-addendum
55. https://newsroom.ibm.com/Whitepaper-A-Policymakers-Guide-to-Foundation-Models

56. https://techcrunch.com/sponsor/zero-systems-company/bridging-the-cosmic-gap-between-generative-ai-and-the-enterprise/
57. https://news.bloomberglaw.com/us-law-week/employers-should-consider-these-risks-when-employees-use-chatgpt
58. https://www.cyberhaven.com/blog/4-2-of-workers-have-pasted-company-data-into-chatgpt/#:~:text=Identifying%20what%20data%20goes%20to%20ChatGPT%20isn%E2%80%99t%20easy
59. https://www.forbes.com/sites/tomdavenport/2023/03/20/how-morgan-stanley-is-training-gpt-to-help-financial-advisors/?sh=5c0fd68c3fc3