Data Pseudonymization in the Generative Artificial Intelligence Environment

Jaroslava Kniežová
Bratislava, Slovakia
Jaroslava.Kniezova@gmail.com

Patrik Hrkút, Emil Kršák
University of Žilina
Žilina, Slovakia
Patrik.Hrkut@fri.uniza.sk, Emil.Krsak@fri.uniza.sk

Abstract — **This paper deals with data pseudonymization when using generative artificial intelligence, specifically ChatGPT. The aim of the research was to experimentally verify whether ChatGPT can maintain the protection of pseudonymized data by not maintaining the link between the anonymized part and personal data. Four requests were tested in the experiment, in which ChatGPT was gradually provided with various combinations of pseudonymized and complete data about customers and their orders. The results showed that ChatGPT does not store previous inputs and is not able to reconstruct personal data retroactively if they were not explicitly provided to it in each request. Based on the results obtained, it can be assumed that the use of ChatGPT when processing pseudonymized data does not pose a risk of violating the protection of personal data. This research seeks to contribute to the discussion on the security of generative artificial intelligence in processing sensitive data and confirms the effectiveness of pseudonymization as a mechanism for protecting personal data.**

I. INTRODUCTION

The present era is characterized by the creation of a large amount of data in all areas of human activity. On the one hand, this data represents an almost unlimited set of opportunities for its use, whether in business or private sphere for individuals or groups of people, or. organizations. On the other hand, data that is suitable for processing for this purpose is constantly generated in large quantities with relatively accessible options for their further processing. The amount of data is continuously processed within the framework of business, state regulatory processes or research.

Since the processed data can carry information not only about the processed and recalculated indicators in business, for example, or. research, but also about individuals who can be clearly identified with certain types of such data and possibly damaged by inappropriate publication of such data. Therefore, the processing and archiving of data is subject to legal regulations, contractual restrictions and regulations. However, to maintain the effectiveness of the costs incurred, it is necessary to plan (and implement) the process of processing the necessary data in advance and thoughtfully.

When a company processes data containing personal data, it must comply with the regulation called GDPR – General Data Protection Regulation [1]. GDPR represents the rules for processing and using personal data that can contain information about a precisely identified individual. The main purpose of GDPR is to ensure that data of a specific individual is not processed, published or otherwise used without the person's consent.

Since consent in the case of a larger number of identifiable people can be complicated to obtain and, moreover, in some activities, requiring consent can be counterproductive (for example, in research, only data that has been agreed upon could be used, the rest would not be taken into account, which could significantly distort the results), there is a need to process such data without the need to obtain consent. In many cases, this means modifying this data so that it is not possible to identify specific persons to whom the data relates.

Data processing in this sense is called anonymization. As part of it, identifiers of specific persons are usually removed from the data. Individual identifiers are not necessary for calculating various indicators, trends, or other interpretations of data. However, there are many cases where completely removing the link to an individual is not a suitable solution, since, for example, business entities often need to have a link to the individual to whom the data relates as their customer. On the other hand, they need to publish certain interpretations of the data publicly, or include them in various outputs, which are then sent and processed in other places. In these cases, so-called pseudonymization comes into play.

Data treated in this way does not contain personal identifiers but is still somehow linked to the data that identifies these individuals. However, these are stored in another place. In such a case, the entity can publish a group of data without personal identifiers in public places as needed, but the other group, which would allow these persons to be identified, is kept separately by the entity. If necessary, this data can be linked again to specific people and the entity can then, for example, personalize its activities towards its customers.

In connection with the GDPR, it should be noted that such protection does not make sense for every data processing. Especially in the case of statistical processing, or even the processing of so-called big data, data that would identify a specific individual is lost in these processes, therefore in these cases their protection loses its meaning (more in [2]).

In other areas, such as research, it cannot be argued that individual identifiers lose their significance. Although when data is manipulated to obtain an evidentiary interpretation, specific individuals are usually eliminated when applying the generalization method, this may not always be the case and, for

example, in case studies in the field of health, where specific cases are described, this connection may be of a highly personal nature.

Several sources are dedicated to ensuring data protection, as not only organizations deal with the processing of personal data during their business, but also various research institutions that need to process personal data that could lose value for research through anonymization, e.g. [3].

This article focuses on the need to protect personal data when it is necessary to maintain ties to specific individuals, which means that the content of the research will be exclusively pseudonymization. The solution to pseudonymization is assessed in the article in connection with the use of generative artificial intelligence, which has become a phenomenon of our time. Its free availability in its basic form is an attraction for working with data in all areas, from hobby activities to professional activities within the processes of a given entity.

The content of the research will be an experiment, whether at all, or to what extent artificial intelligence can breach data protection in the case of applying pseudonymization, if the processing of this data is used within a publicly available service, specifically ChatGPT [4]. ChatGPT is:

I am ChatGPT, an advanced language model developed by OpenAI. My purpose is to assist users by answering questions, providing advice, analyzing texts, generating content (such as articles, emails, scripts, and other documents), helping with programming, retrieving up-to-date information, and much more.

I do not have a personal identity or consciousness; I process text based on a vast amount of training data. My knowledge is current up to June 2024, but I can search the web for more recent information if needed.

Let me know how I can help, and I'll tailor my response to your needs. 😊

As part of the experiment, ChatGPT will be used to process a selected sample of pseudonymized data for a specific business case of the company. The sample of data is intentionally small to create the least complicated conditions for ChatGPT to capture the link that binds the completely anonymized part of the data and the separate (but linked) part of the data containing personal data. This data will be provided for processing and the outputs because of the experiment will be used to confirm or refute the hypothesis.

II. Related Work

As mentioned, pseudonymization is the process by which identifying data is transformed into a special form so that it is not possible to directly identify who it concerns without some decoding procedure [5], [6].

The issue of using pseudonymization in scientific articles has been around for a long time. Its need arose mainly for medical research, because it was necessary to hide data that came from patients during research. On the other hand, it was necessary to preserve the original information about who the data came from.

One of the first works was the work [7] from 2007, where the authors considered how to anonymize images from the field of radiology for the purpose of clinical research, but at the same time not to lose the connection to the patient, so that they would be able to connect the results back to him. In another work Riedl at el. [8] presents a detailed description of the new system pseudonymization of Information for Privacy in e-Health which can securely integrate primary and secondary usage of health data especially for use in electronic health records. Research in this area is ongoing, as evidenced by publications from recent years, e.g. [9], [10], [11], [12].

Although the use of generative AI is currently only at its beginning, there are already some scientific articles in this area. For example, [4] addresses the protection of personal data when using large language models. Paper [13] conduct an analysis of tagsets that have previously been utilized in anonymization and pseudonymization in the field of Natural Language Processing (NLP) applications. Zhou at al. in [14] in his survey provides an overview of the current state of research on security of using ChatGPT, with aspects of bias, disinformation, ethics, misuse, attacks and privacy.

The results of similar work [15] to ours demonstrate GPT-4's potential as a powerful tool for safeguarding patient privacy while increasing the availability of clinical data for research. This work sets a benchmark for balancing data utility and privacy in healthcare data management.

III. Methodology

At the beginning of the research, we define the research problem and formulate a hypothesis that we will verify using an experiment. For the sake of the experiment, we will create a data set that will contain personal data, as well as an anonymized data sample. We will insert this data into a conversation with ChatGPT and, using controlled communication, we will try to get back the information that we previously entered into the conversation. Based on the responses from ChatGPT, we will then evaluate the experiment and confirm or refute the established hypothesis.

A. Hypothesis formulation

This part of the research is crucial, and it is necessary to choose appropriately what will be investigated. The hypothesis formulation is an expression of our opinion on how ChatGPT treats the data that the user provides to it. It is an estimate of how we think the object under investigation works. It should be defined by a statement that states what we want to find out through the research. In our case, we will try to prove or disprove whether ChatGPT also stores data that is of a personal nature and will use it in later communication.

B. Experiment design

The research will consist of a series of question/answer inputs, in which we will gradually enter various tasks related to the processing of the data provided. We have designed 4 questions that we will enter into a conversation with ChatGPT. The goal will be to verify whether ChatGPT can reconstruct the connection between the anonymized and personal part of the data from various inputs. We will gradually evaluate these

requirements and, using various manipulative techniques, we will try to force ChatGPT to use previously entered inputs in its answers. For the experiments, we have chosen the ChatGPT language model with the latest version currently available. This is the ChatGPT-4-turbo model, which is optimized to run faster and cheaper than the GPT-4 version, while maintaining high quality of answers.

Of course, we realize that it would be possible and probably appropriate to test multiple generative AIs, but this is the beginning of our research and for its purpose we will consider the choice of ChatGTP AI sufficient.

C. Evaluation of the experiment

After entering the inputs into ChatGPT, we will therefore expect answers that either confirm or reject our hypothesis. We will analyze the results of the experiment and investigate whether ChatGPT does not use non-anonymized data from previous conversations to generate answers. Based on the analyzed results of communication with ChatGPT, we will then be able to determine whether the generative artificial intelligence (ChatGPT) has stored personal data and used it in subsequent answers. Finally, we will deduce how ChatGPT handled the data in this specific case and to what extent there is a risk of misuse of personal data in such cases.

IV. RESULTS

A. Hypothesis formulation

The prerequisite for data processing using artificial intelligence is that personal data will only be used in the processing if they are provided at the time of entry. This means that if ChatGPT has already processed data that includes personal data protected under the GDPR, it does not store this data after processing. The learning process therefore does not include collecting and archiving data from queries for later use. ChatGPT itself answers the question of whether it stores data from processing for later use in the negative. It only allows their short-term storage within a single conversation. The answer in Fig. 1

> I don't have the ability to learn in real-time or remember past queries long-term, but I do retain context within a single conversation. This means I can remember what we've been discussing during our current chat.
>
> If you want me to remember something for future conversations (e.g., the topic of a project you're working on), I can save that information. Just let me know! 😊

To verify the above fact, we decided to formulate a hypothesis in the following form:

When processing data in pseudonymized form using ChatGPT, the data protection applied to them will be maintained and ChatGPT will not retain the link between the anonymized part of the data and the part with full personal data. This is even though the task that will be required of ChatGPT will assume the use of this link.

This hypothesis assumes that ChatGPT, despite being able to work with the context in which the entire conversation takes place, does not store sensitive information from previous inputs, which should maintain the protection of personal data in accordance with the rules on personal data protection.

B. Sample data and solution for their pseudonymization

To implement the experiment, data that is processed within the framework of common economic cases will be used, namely a small database of customers and their orders. For the processing of customer data, an indicator of the total amount of orders for the evaluation of customers within the business intelligence processes of this entity is important for the given company in our business case. However, for publishing these outputs in reports, it is necessary that it is not possible to identify specific customers, due to the protection of their personal data, because the given entity wants to publish the outputs on its website as well as with each promotion of its products. On the other hand, the company wants to further develop a very effective customer program and reward customers with larger subscriptions. The basic model of the operational database is shown in Fig. 2.

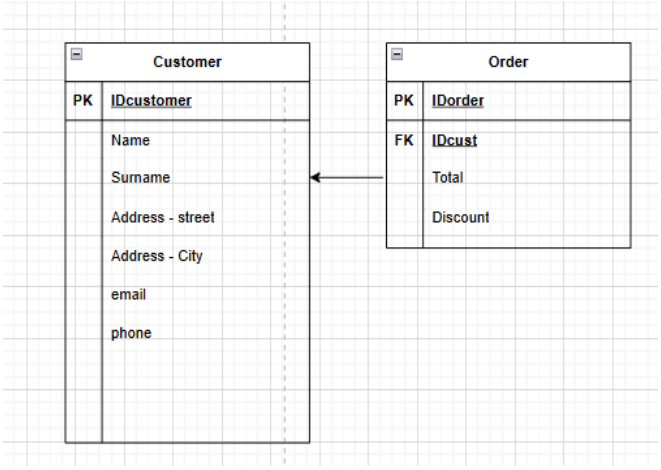

Fig. 2. Operational database model for recording customers and their orders

For our business entity, these data were modified to be pseudonymized, and the model was divided into two parts. One part contains full data for individual customers. This part will not be published publicly. The second part is modified so that it does not contain identifiers of specific individuals but remains linked to the first part, see Fig.3.

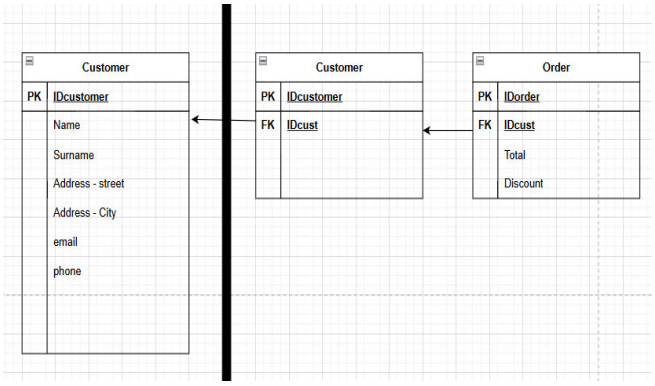

Fig.3. Pseudonymous data – split model into two parts while maintaining mutual connectivity

We chose MS Excel as the file format for saving, because it is one of the most common data formats. The correctness of choosing this format is also confirmed by the fact that the mutual connection of the two parts is not implemented by the database engine software (which would be used in the case of saving data to a relational DB) and therefore this connection can be more easily "understood" by ChatGPT. Fig. 4 and 5 show the specific data that were used in the experiment.

	A	B	C	D	E	F	G
1	IDcustomer	Name	Surname	Address-street	Address - City	email	phone
2	1	Andrew	Stape	Green 89	Brenn	And.S@goho.com	
3	2	Peter	Second	Hund 99	Bann	Psec@gm.com	
4	3	Alice	Crann	Paw 5	Bonny		55866982
5							

Fig. 4. Customer data including personal identifiers

	A	B	C	D
1	Idorder	Idcust	Total	Discount
2	44-55	2	100	10
3	44-56	1	500	5
4	44-57	3	200	0
5	44-58	2	10	0
6	44-59	1	100	10
7				

Fig. 5. Customer order data without personal identifiers, including a link to a section with complete customer data

The data is organized in two separate files so that pseudonymization is maintained. In addition to the sheet mentioned above, the file with pseudonymized order data also contains one more sheet so that it is not just a single-sheet file. However, the order amounts to themselves and the link are contained in the sheet shown in the image.

C. Experiment

As part of the experiment, a total of 4 requests were entered into ChatGPT within the conversation. At the beginning of the conversation, in the first request, complete data was provided to ChatGPT, which means that both separate files were inserted for analysis and obtaining the result. In subsequent requests and communications, the file with complete personal data of customers was no longer provided to ChatGPT.

a) Task No.1 for ChatGPT

Task: The first request for ChatGPT contained 2 tasks: The first required compilation of a customer overview with their names and the total number of orders, which ChatGPT had to calculate from all orders of a given customer. The second task required the creation of a formula in one of the input files that would create such a customer overview.

Expected output: ChatGPT had to use the connection between both parts of the data for both tasks. Two separate files were provided as input. Fig. 6 shows the output – the response to this request, and the output can be considered correct.

Result: ChatGPT used and loaded the connection between the two parts.

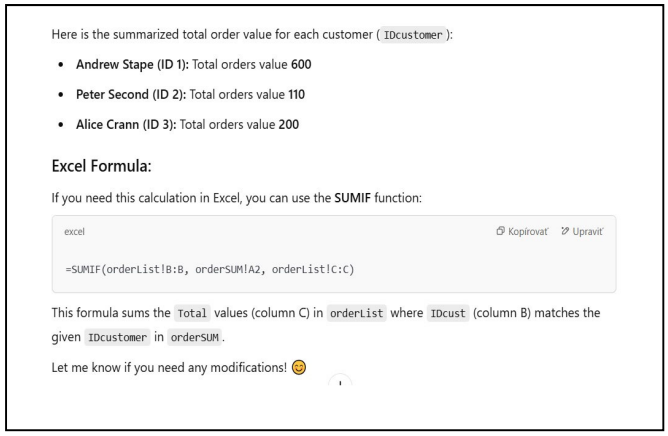

Fig. 6. Answer to the first request - recalculating amounts depending on customers and building a formula for this calculation

b) Task No.2 for chatGPT

Task: Within the second request, ChatGPT was given a task to create an overview of customers with the total amount calculated from their orders (the same task as the first task from the previous request). However, only a file with data without personal data was provided as input. Therefore, the names and surnames of these customers were missing.

Expected output: The goal was to test whether ChatGPT would not return these names based on their storage from processing from the previous task as part of its learning processes for improving outputs.

Result: ChatGPT returned an output that it called an analysis of the input file, but did not create the required output, citing the fact that it lacked data with the names and surnames of individual customers. ChatGPT therefore did not use the data link that it loaded in the previous task, and it can be assumed that it did not store this link and thus did not violate the protection of personal data. The output from the second task is shown in Fig. 7.

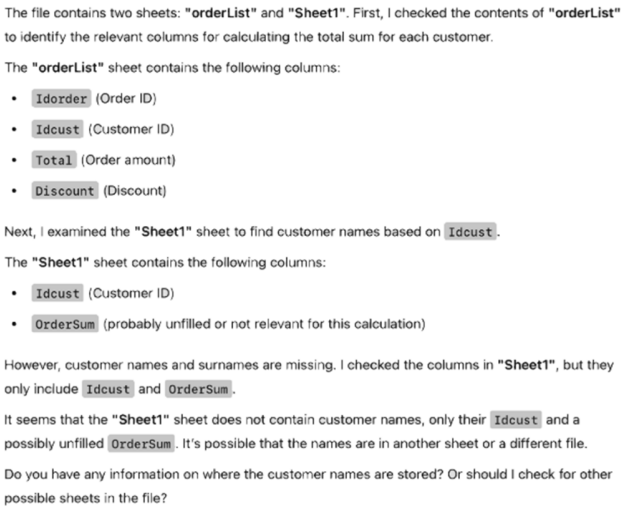

Fig.7 Answer to the second request: recalculating amounts depending on customers without providing a file with full customer data at the input

c) Task No.3 for ChatGPT

Task: In the third request, ChatGPT was given the task of compiling a list of customers with the total amounts of their orders, while there was no file as input, but a list of customers whose orders were recalculated in the first task was provided. The list was provided in text form, the names and surnames exactly matched the names and surnames from the file containing customer data that ChatGPT received as part of the first task.

Expected output: The goal was to test whether ChatGPT would not use the loaded data from the first task and would not assign individual previously calculated order amounts to individual customers in the provided list.

Result: The output from this request is shown in Fig. 8. ChatGPT stated in its response that it needed data that it did not receive as input to compile the list according to the assignment.

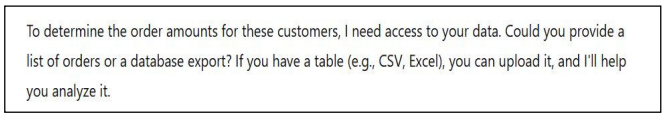

Fig. 8. Response to the third request to recalculate amounts depending on customers without providing files, only listing customer names at the input

It is clear from the response that chatGPT did not use the loaded data and it can be assumed that it did not save the previously loaded and calculated data.

d) Task No.4 for ChatGPT

Task: In the last request, chatGPT was given the same task as in the second request, but with a greater time interval, namely 1 day.

Expected output: ChatGPT should have responded whether it had complete data for the given task from the input.

Result: Again, however, it can be stated that chatGPT did not save and reuse the data from the first task and did not violate the protection of personal data in the sense of their pseudonymization. Interestingly, unlike the output of the same task from the previous day, this time it reports an error in the analysis in the response. The response from the fourth task can be seen in Fig. 9.

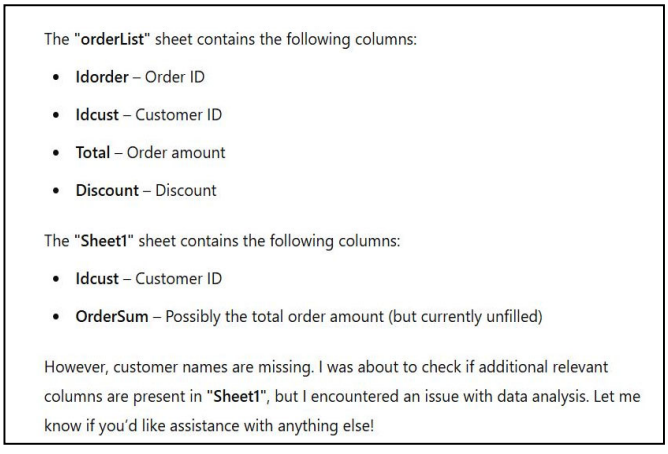

Fig. 9. Response to the fourth request for recalculation of amounts depending on customers without providing a file with complete customer data at the input, made on another day

We have summarized the entire communication during the experiment in Table I, where all questions and answers are listed, and it is possible to get an overall overview of how the communication took place and ended.

TABLE I. POPULAR DRAWBACKS FOR SUBMISSIONS OF FRUCT13

Task No.	Task	ChatGPT response
1	Create a list of customers with their names and the total number of orders. Create a formula that will create such a list.	The response contained the requested data.
2	Create a customer overview with the total amount deducted from their orders (same task as the first task from the previous request), but only one file was provided (without personal data).	The output was not created with information that the data required for such an assignment is missing.
3	Compiling a customer overview with the total amounts of their orders, where there was no file as input, but a list of customers in text form was provided.	Output was not created, ChatGPT said it did not have the necessary data.
4	Same assignment as for the second task but with a certain time interval (the next day).	The output was not generated with information that an error occurred in the analysis.

V. CONCLUSION AND DISCUSSION

Many entities work with personal data that fall under the protection regulated by the GDPR. At the same time, however, this data is the subject of analyses, calculations and various strategic interpretations that individual entities need to carry out. In these processes, it is still necessary to comply with the protection of this data. Moreover, artificial intelligence is now relatively commonly used in the processing of this data, and it is a legitimate assumption that individual entities will use artificial intelligence more and more even in the case of processing personal, sensitive data. Therefore, the question arises whether the use of artificial intelligence will not violate the protection of this data.

Through an experiment, when data falling under the protection of the GDPR was provided to ChatGPT in pseudonymized form, the hypothesis

When processing data in pseudonymized form using ChatGPT, the data protection applied to them will remain intact and ChatGPT will not retain the link between the anonymized part of the data and the part with full personal data. This is even though the task that will be required from ChatGPT will assume the use of this connection.

was confirmed. ChatGPT worked with personal data and included this data in the requested output only if it was also available at the input. It can be assumed that it only includes data processed from the provided inputs in the output. It does not store and does not use this data in requests to process the same data, if this was not provided directly at the input.

Although at first glance it may seem that the data sample is relatively small, in our opinion it is sufficient to verify our hypothesis. In the case of a larger amount of data, whether in the form of a larger amount or a more complex structure, we do not assume that ChatGPT would proceed differently. In our opinion, the volume of data does not play a role in this case. However, if an organization were to create its own solutions based on LLM models, there is nothing preventing it from integrating data provided by users into its solution. Moreover, in such a case, there is a high probability that there will be no violation of data anonymization, since we assume that the company's employees will not enter their personal data, but data that is somehow related to the functioning of the company and will be so cautious that they will not even enter sensitive data that would reveal the most important company secret.

We believe that this issue is currently very relevant, because generative AI is increasingly penetrating our lives and draws its information from various sources, not only available on the Internet, but also learns from conversations that are entered by communicating users. This information can often be intimate and sensitive, so there is a real threat that it can be misused. We will continue our research, try other generative AI and compare and evaluate the results.

Based on our results, it is not possible to directly generalize the conclusions to all generative AIs, but we have selected the most widely used and widespread AI. We assume that other AIs will behave similarly, but this requires further research.

On the other hand, we would like to warn against premature optimism and the claim that based on this small research it can be said with certainty that generative AI does not impose any input on the user is not sufficiently proven. Just as publishing information on websites brings with it certain risks (e.g., once published information is difficult to remove from websites), similar caution should be observed in the case of ChatGTP and other AIs. ChatGPT itself claims that it obtains data from publicly available and licensed sources, but also from conversations with users.

REFERENCES

[1] REGULATIONS REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance).

[2] L. Bolognini and C. Bistolfi, “Pseudonymization and impacts of Big (personal/anonymous) Data processing in the transition from the Directive 95/46/EC to the new EU General Data Protection Regulation,” Computer Law and Security Review, vol. 33, no. 2, pp. 171–181, Apr. 2017, doi: 10.1016/j.clsr.2016.11.002.

[3] F. Kohlmayer, R. Lautenschläger, and F. Prasser, “Pseudonymization for research data collection: Is the juice worth the squeeze?,” BMC Med Inform Decis Mak, vol. 19, no. 1, Sep. 2019, doi: 10.1186/s12911-019-0905-x.

[4] G. Sebastian, “Privacy and Data Protection in ChatGPT and Other AI Chatbots,” International Journal of Security and Privacy in Pervasive Computing, vol. 15, no. 1, pp. 1–14, Jul. 2023, doi: 10.4018/ijsppc.325475.

[5] J. Montagnat et al., “Medical Images Simulation, Storage, and Processing on the European DataGrid Testbed,” 2004.

[6] K. A. Taipale, “TECHNOLOGY, SECURITY AND PRIVACY: THE FEAR OF FRANKENSTEIN, THE MYTHOLOGY OF PRIVACY AND THE LESSONS OF KING LUDD,” 2004.

[7] R. Noumeir, A. Lemay, and J. M. Lina, “Pseudonymization of radiology data for research purposes,” J Digit Imaging, vol. 20, no. 3, pp. 284–295, Sep. 2007, doi: 10.1007/s10278-006-1051-4.

[8] B. Riedl, V. Grascher, S. Fenz, and T. Neubauer, “Pseudonymization for improving the Privacy in e-Health Applications,” 2008.

[9] T. J. M. Chin, G. X. M. Chin, J. Sutherland, A. Coon, C. Morton, and C. Fleming, “BT24 Pseudonymization for artificial intelligence skin lesion datasets: a real-world feasibility study,” British Journal of Dermatology, vol. 191, no. Supplement_1, pp. i199–i200, Jun. 2024, doi: 10.1093/bjd/ljae090.421.

[10] M. Al-Zubaidie, Z. Zhang, and J. Zhang, “PAX: Using Pseudonymization and Anonymization to Protect Patients’ Identities and Data in the Healthcare System,” 2019, doi: 10.3390/ijerph16091490.

[11] H. Abu Attieh, M. Halilovic, T. Meurers, A. Müller, D. T. Neves, and F. Prasser, “Rapid Deployment of a Pseudonymization Service in a Distributed Research Data Infrastructure - Lessons Learned,” Stud Health Technol Inform, vol. 316, pp. 1248–1249, Aug. 2024, doi: 10.3233/SHTI240637.

[12] O. Yermilov, V. Raheja, and A. Chernodub, “Privacy-and Utility-Preserving NLP with Anonymized Data: A case study of Pseudonymization,” 2023. [Online]. Available: https://huggingface.co/bert-base-cased

[13] M. I. Szawerna, S. Dobnik, T. L. Tiedemann, R. Muñoz Sánchez, X.-S. Vu, and E. Volodina, “Pseudonymization Categories across Domain Boundaries,” 2024.

[14] W. Zhou et al., “The security of using large language models: A survey with emphasis on ChatGPT,” IEEE/CAA Journal of Automatica Sinica, 2024, doi: 10.1109/JAS.2024.124983.

[15] B. Altalla et al., “Evaluating GPT models for clinical note de-identification,” 2025, doi: 10.1038/s41598-025-86890-3.