



OPEN Man vs the machine in the struggle for effective text anonymisation in the age of large language models

Constantinos Patsakis^{1,2,4}✉ & Nikolaos Lykousas^{2,3,4}

The collection and use of personal data are becoming more common in today's data-driven culture. While there are many advantages to this, including better decision-making and service delivery, it also poses significant ethical issues around confidentiality and privacy. Text anonymisation tries to prune and/or mask identifiable information from a text while keeping the remaining content intact to alleviate privacy concerns. Text anonymisation is especially important in industries like healthcare, law, as well as research, where sensitive and personal information is collected, processed, and exchanged under high legal and ethical standards. Although text anonymisation is widely adopted in practice, it continues to face considerable challenges. The most significant challenge is striking a balance between removing information to protect individuals' privacy while maintaining the text's usability for future purposes. The question is whether these anonymisation methods sufficiently reduce the risk of re-identification, in which an individual can be identified based on the remaining information in the text. In this work, we challenge the effectiveness of these methods and how we perceive identifiers. We assess the efficacy of these methods against the elephant in the room, the use of AI over big data. While most of the research is focused on identifying and removing personal information, there is limited discussion on whether the remaining information is sufficient to deanonymise individuals and, more precisely, who can do it. To this end, we conduct an experiment using GPT over anonymised texts of famous people to determine whether such trained networks can deanonymise them. The latter allows us to revise these methods and introduce a novel methodology that employs Large Language Models to improve the anonymity of texts.

In today's data-driven society, the collection and use of personal information are becoming increasingly prevalent. While this has numerous benefits, such as improved decision-making and better service provision, it also raises important ethical concerns related to privacy and confidentiality. Indeed, harvesting user data is a common practice of far too many online platforms and services with a significant impact on citizens. This has been one of the pillars that led to the introduction of the General Data Protection Regulation (GDPR)¹ and other relevant legislation around the world as a means to address the privacy issues that emerged. The GDPR mandates using privacy-preserving methods and processes throughout the data management lifecycle, from collection and processing to sharing and publishing. One of these fundamental methods is anonymisation. Given that modern organisations continuously deal with documents, the above has served as a catalyst in the emergence of text anonymisation as a research topic with many practical applications. The general concept is that given a text, one has to remove or mask identifiable information while preserving the remaining content. Text anonymisation is particularly relevant in healthcare, law, and research, where personal and sensitive information is overwhelming and must be protected to comply with privacy regulations and ethical guidelines.

Although text anonymisation has been widely adopted in practice, it still faces significant challenges. These methods must strike a balance between the need to protect the privacy of individuals and the need to preserve the data utility. Let us consider this with an example where the anonymisation task is to anonymise the sentence “Volodymyr Zelenskyy is the president of Ukraine”. Clearly, simply removing the name is not enough. If one is given the sentence “[NAME] is the president of Ukraine”, it is trivial to recover the missing information. Therefore, the anonymised sentence would be “[NAME] is the president of [COUNTRY]”. To this end, named entity

¹Department of Informatics, University of Piraeus, 80 Karaoli & Dimitriou str, 18534 Piraeus, Greece. ²Management Systems Institute of Athena Research Centre, Marousi, Greece. ³Data Centric Services, Bucharest, Romania. ⁴These authors contributed equally: Constantinos Patsakis and Nikolaos Lykousas. ✉email: kpatsak@unipi.gr

recognition methods are used to identify possible identifying information, such as names and locations, to mask them. Let us go back to the example.

While the sentence is anonymised, it is easy to understand that the anonymity of the masked person is bounded. Given that the sentence refers to a country's president, there are 193 possible choices, which in terms of k -anonymity² would be perfect. Nevertheless, even fragments of information from other sentences in the context of a text could significantly limit the possible choices. Therefore, the challenge is not only whether an algorithm finds clearly identifiable information but whether the remaining information is enough to deanonymise an individual and reveal the rest of the pruned information. For instance, using non-identifying information, e.g., referring to his acting career or studies, can limit the candidates to just a handful of people. Therefore, the challenge comes down to how much information is available and whether this can be properly extracted to infer the identity of individuals.

In the past two years, endless discussions have sparked with the introduction of GPT from OpenAI, which boomed with the introduction of ChatGPT. The GPT-3 model was trained on a text dataset of more than 8 million documents and over 10 billion words and is 175 billion parameters in size. This allows it to perform many text-generative tasks efficiently, astonishing people worldwide. Based on the above, a natural research question is to ponder whether such systems, trained over texts with sensitive content that are available from one organisation, would be able to deanonymise published anonymised texts from another organisation. Evidently, the latter could have a catastrophic impact on individuals, despite applying best practices from the publishing organisation.

We performed a scoped experiment using GPT targeting famous individuals to assess this threat scenario. Since efficiently and effectively training such systems is by itself a challenge, we opted to use the available closed but off-the-shelf GPT to deanonymise texts that had been anonymised by a state of the art algorithm, namely Textwash³. The choice of testing on deanonymising famous people is that this maximises the chances of GPT to have been trained on relevant documents and that, this way, we depend solely on public information. Then, we use GPT to improve text anonymisation, increasing the actual anonymity of the texts without significantly impacting their quality.

The rest of this work is structured as follows. The next section provides an overview of the current state of the art. Then, in Section “[Threat scenario](#)”, we detail our threat scenario. Section “[Dataset and methodology](#)” introduces the dataset and our methodology. Next, Section “[Experimental results](#)” presents the results of our experiments. Finally, the article concludes by discussing our findings and ideas for future work.

Related work

In the next paragraphs, we discuss personal data and identifiers. Then, we provide an overview of text anonymisation algorithms. Finally, we describe two essential parts of our work, Textwash, an open-source text anonymisation tool, and the GPT model.

Personal data

Personally Identifiable Information (PII) refers to any piece of information that can directly or indirectly identify a specific individual. In this regard, direct identifiers can provide an explicit link to an individual and identify them. In most cases, direct identifiers are unique values. Typical examples of direct identifiers are identity, passport, driver's license, and Social Security numbers. The name is also a direct identifier, however, they are not unique. Quasi-identifiers are attributes that do not uniquely identify individuals on their own; nevertheless, once someone combines them with other quasi-identifiers or other data, they can narrow down the possible individuals to the point of uniquely identifying the individual. Typical quasi-identifiers are birthday, gender, ethnicity, postal code, and occupation. Some legal frameworks, such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States tried to define some quasi-identifiers in medical documents that have to be removed to protect the privacy of individuals. Nevertheless, as outlined by Narayanan and Shmatikov “*any attribute can be identifying in combination with others*”⁴ so subsequent legislations do not directly refer to specific identifiers and regulators acknowledge the challenges that data anonymisation faces, as well as its limitations and possible ephemeral nature⁵.

Due to the wide harvesting and exploitation of PII by organisations worldwide, data protection and privacy regulations, such as the General Data Protection Regulation (GDPR) in the European Union and the California Consumer Privacy Act (CCPA) in the United States

According to the GDPR (Article 4), personal data are defined as follows¹:

“Personal data” means any information relating to an identified or identifiable natural person (“data subject”); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;

Therefore, the GDPR considers that personal data are pieces of information that individually or with other information can be used to identify, contact, or locate an individual. As a result, personal data are widely described as Personally Identifiable Information (PII). Due to the sensitivity of the underlying data, data protection and privacy regulations, such as the General Data Protection Regulation (GDPR) in the European Union or the California Consumer Privacy Act (CCPA) in the United States make specific provisions on the collection, handling, and management of such data. As a result, organisations must adopt appropriate data protection techniques, such as anonymisation, pseudonymisation, or encryption, subject to constraints on the type and volume of data, user roles, and jurisdiction, to name a few. They should also follow the principles of data minimisation and privacy by design, which means collecting and processing only the necessary data and incorporating privacy

considerations from the outset of any project. Of specific interest is consent management, as organisations are obliged to have the direct and informed consent of the data subjects to collect and process their data. Nevertheless, the aggregated and anonymised version of user data can be published without their consent, as theoretically, the identifying information has been removed and individuals cannot be identified. Indeed, according to the GDPR: “*This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes*”.

Data anonymisation methods

To anonymise data there are various methods that depend on the underlying information and how it would be published. For instance, there are specific methods for tabular data which may greatly vary depending on whether an updated version of the dataset would have to be published again. Such methods may generalise or suppress records, introduce noise, slice the data etc.⁶. Nevertheless, such data are well-defined and have very specific and structured information. Similarly, transactional and trajectory data may have to employ similar methods, e.g. noise addition and generalisation, yet in different contexts to make the data reusable.

However, for textual data, the approaches are radically different as information is not well structured, at least for a machine. Therefore, the first step to applying a text anonymisation algorithm is to identify the structure of the document that would allow one to trace potentially sensitive information and then prioritise these pieces of information, e.g., to purge the ones that seem more appropriate. Apparently, the above describes a natural language processing task, and named entity recognition plays a central role^{7–9}.

While mainly dictionary and pattern-based, Scrub¹⁰ was the first such method and targeted the anonymisation of medical records. Neamatullah et al.¹¹ followed the same path as their approach is also based on dictionaries, however, they use regular expressions and simple heuristics to locate possible sensitive information, including doctor names and years of dates. Similarly, Ruch et al.¹² utilised semantic lexicon to anonymise medical documents. Dernoncourt et al.¹³ used a long short-term memory (LSTM) to anonymise patient notes that were tokenised text using the Stanford CoreNLP tokeniser¹⁴. LSTMs alongside conditional random field (CRF) and regular expressions were also utilised in¹⁵ to anonymise clinical notes. Sánchez et al.^{16–18} leverage Information Content, disclosure risk, and knowledge bases to detect possible sensitive content in tokenised texts. Uzuner et al.¹⁹ introduced Stat De-id which uses support vector machines to anonymise medical discharge summaries by assessing whether each word in the text is a sensitive feature.

In a parallel line of thought, Carrell et al.²⁰ introduced the Hiding in plain sight (HIPS) method. Practically, this method tries to conceal residual identifiers by introducing additional noise in the form of falsified information in the text to prevent the identification of individuals. To this end, the method replaces detected PII with fictitious, yet semantically similar surrogates. The principle behind this approach is that even if an adversary successfully detects potential PII within the text, the task of differentiating genuine identifiers from falsified ones is obfuscated due to their random substitutions. This could mislead the adversary towards incorrect or non-existent individuals, or result in contradictory information, effectively neutralizing a deanonymisation attack.

For more details, the interested reader may refer to^{21–23}.

Textwash

One of the latest tools for text anonymisation is Textwash³, an open-source tool introduced by Kleinberg et al., specifically addressing the problem of privacy-preserving data sharing. The general concept of Textwash is to identify sensitive and potentially sensitive information and redact it while maintaining semantic coherence to ensure that the anonymised out remains usable for various downstream text analysis tasks. Textwash is based on supervised machine learning, leveraging pre-trained contextualised word representations provided by a fine-tuned BERT language model. Much like the NER-based text anonymisation tools, the categories of redactable text data considered by Textwash comprise a pre-defined set of 11 possible tags, the annotations provided by domain experts. More precisely, the possible tags are PERSON_FIRSTNAME, PERSON_LASTNAME, OCCUPATION, LOCATION, TIME, ORGANIZATION, DATE, ADDRESS, PHONE_NUMBER, EMAIL_ADDRESS, and OTHER_IDENTIFYING_ATTRIBUTE. The last tag comprises a meta-category which encapsulates the **potentially sensitive information** (PSI) concept. Concretely, the PSI notion captures the full spectrum of textual information that could reveal an identity which cannot be attributed to a well-defined category of PII. This is made possible by leveraging the contextual awareness of its Transformer-based architecture^{24–26}.

To preserve the semantic properties of anonymised texts while removing any identifiable information, Textwash implements a two-stage anonymisation process. Specifically, after a token is classified as one of the 11 possible categories (e.g. John), it will be replaced with the relevant tag (e.g. `PERSON_FIRSTNAME`), with a numeric suffix incremented for each different instance of the specific category. For instance, if in the document there is another name, e.g. George, then each occurrence of John would be replaced by `PERSON_FIRSTNAME_1`, and each occurrence of George would be replaced by `PERSON_FIRSTNAME_2`.

GPT

GPT-3 is a large language model (LLM)²⁷ that relies on techniques such as tokenisation, part-of-speech tagging, named entity recognition, and syntactic parsing to understand the structure and meaning of natural language text. The underlying architecture of GPT-3 is based on the transformer model²⁸, which is highly effective for natural language processing tasks. According to OpenAI, the training data for GPT-3 includes various sources such as books, articles, and websites, with a primary source being the Common Crawl (<https://commoncrawl.org/>), a repository of web pages and documents that is regularly updated and maintained and its training process

is described in²⁹. Currently, GPT-3 is also used in various practical applications, including chatbots, with Chat-GPT being in the spotlight, language translation, and text completion.

However, GPT-3 and other AI models also raise ethical concerns. There has been a growing body of research on topics such as fairness, accountability, and transparency in AI to ensure that models are developed and deployed in an ethical and transparent manner. For instance, given the increasing role of AI, negligence and liability^{30,31} must be reconsidered. The above concerns led Brundage et al.³² to propose ways to ensure that AI models are developed and deployed ethically and transparently and the EU Commission to push towards the development of the AI Act³³, the first law on AI by a major regulator anywhere in the world, to mitigate possible risks from the wrong use of AI and regulate its development and deployment in a well-defined legal and regulatory framework.

Threat scenario

Currently, GitHub’s Copilot³⁴ uses OpenAI’s Codex³⁵ a descendant of GPT-3, as it is based both on a corpus from both physical language and code documents. Interestingly, Copilot was reportedly leaking secrets and only recently did GitHub issue measures to stop it from doing so³⁶. The above means that LLMs can now efficiently understand sensitive information and its context. Therefore, we face the following threat. If an LLM is trained on a large corpus of sensitive information, then it may have the capacity to deanonymise relevant anonymised texts. By merely removing obviously sensitive parts, we cannot guarantee that the result would provide the privacy guarantees that we would expect. Due to the continuous increase in the quality and size of training datasets, as well as the sophistication to calibrate them, LLMs are managing to perform many tasks with high efficiency. Therefore, we must also consider them part of the attack tooling. Given the sheer amount of data that modern organisations hold, it is only a matter of time until some of them try to feed this data to such systems. Going back to the Copilot use case, where secrets were leaked when one had to fill in a password, we can assume that these LLMs could fill in the gaps of missing information, including the case of anonymised texts.

Based on the above, we consider the following threat scenario. Let there be two organisations \mathcal{A} and \mathcal{B} having two text datasets $T_{\mathcal{A}}$ and $T_{\mathcal{B}}$, respectively. These sets refer to their clients, denoted as $C_{\mathcal{A}}$ and $C_{\mathcal{B}}$, respectively and $C_{\mathcal{A}} \cap C_{\mathcal{B}} \neq \emptyset$. The two organisations exchange information as a basis of their collaboration, but due to legal and regulatory constraints; they anonymise them. Therefore, having in hand two anonymisation functions $Anon_1$ and $Anon_2$ (not necessarily the same), they send to each other $Anon_1(T_{\mathcal{A}})$ and $Anon_2(T_{\mathcal{B}})$. Setting aside the legal and ethical constraints, \mathcal{A} trains/fine-tunes (depending on its capacity) $LLM_{\mathcal{A}}$ with $T_{\mathcal{A}}$. Our new threat scenario considers the exposure of $C_{\mathcal{B}}$ from the use of $LLM_{\mathcal{A}}$ on $Anon_2(T_{\mathcal{B}})$ that \mathcal{A} receives from \mathcal{B} . In what follows, we consider that $Anon_2$ is a black box; however, further tuning could be performed knowing how it works.

Dataset and methodology

In our experiments, we use the dataset from³. More precisely, we use the data from the second study of Kleinberg et al., which contains 1080 descriptions of 20 famous individuals in the UK and their anonymised versions. Kleinberg et al. assigned 200 participants to write a description of some random subset of these celebrities in English. As a result, for each individual, there are 46 to 61 descriptions, with an average of 54. Then, these descriptions were anonymised by Textwash. The text in Table 1 is a sample anonymised text from this dataset. Afterwards, Kleinberg et al. recruited 222 participants to deanonymise ten random texts from this dataset and recorded their success rate. By using this dataset, we refrain from processing any private information, as the whole dataset consists of information about famous individuals such as Adele, David Beckham, and Luis Hamilton, and the information is provided by individuals who have no relation to them. Therefore, the information is not sensitive, it is public, and the data subjects have chosen to make it public. Moreover, due to the fact that they are celebrities, it is highly possible that GPT has already been trained on documents containing related information. Clearly, the relevant documents that might have been used for the training of GPT are different from those used in the dataset.

Using this dataset, we aim to assess whether the anonymised version can lead to the deanonymisation of a document using an LLM. According to OpenAI, GPT-3 does not use online information; therefore, all its responses are based on what the model has learned through its training. Indeed, OpenAI in its “API data usage policies”³⁷ states:

OpenAI will not use data submitted by customers via our API to train or improve our models, unless you explicitly decide to share your data with us for this purpose.

PERSON_FIRSTNAME_2	PERSON_LASTNAME_1	is an	LOCATION_1	singer, songwriter, actor and film producer who was born on	DATE_1	DATE_1			
DATE_1	and is now	NUMERIC_1	years old.	PERSON_FIRSTNAME_2	PERSON_LASTNAME_1	is the lead singer of rock band,	ORGANIZATION_2		
PERSON_FIRSTNAME_2	PERSON_LASTNAME_1	is known as a rock legend and for	PRONOUN	charismatic stage presence and dancing. So much so,					
		that	ORGANIZATION_1	released a song after	PRONOUN	dancing, called	OTHER_IDENTIFYING_ATTRIBUTE_1	like	PERSON_LASTNAME_1
PERSON_FIRSTNAME_2	PERSON_LASTNAME_3	has	NUMERIC_4	children, and has had multiple partners, and	NUMERIC_1	spouse.	PERSON_FIRSTNAME_2		
PERSON_LASTNAME_1	has been with	PRONOUN	current partner	PERSON_FIRSTNAME_1	PERSON_LASTNAME_2	since	DATE_2		

Table 1. An anonymised version of the text for Mick Jagger in the dataset.

Therefore, the information that we used in our experiments was not used to retrain the model.

Having the above dataset at hand, we created a set of tasks for GPT-3, asking it to guess the name of the person that would fit most to each of the provided anonymised descriptions. This is illustrated in Fig. 1.

The results, discussed in the following section, led us to consider the revision of text anonymisation algorithms. We argue that the existence of LLMs and their ability to extract knowledge from large quantities of text, in conjunction with their zero-shot reasoning capacity, offer advanced features for text anonymisation. Therefore, we consider a revised text anonymisation methodology, as illustrated in Fig. 2. Practically, an LLM is queried to report which pieces of identifying information should be pruned. Then, we remove these pieces of information from each text, leading to an improved anonymised version of the dataset. While some records could still be identified, since the adversary would not have access to the same dataset to train the LLM, the risk exposure would be significantly less than in our experiments.

Experimental results

In what follows, we detail our two experiments and their results, illustrated in Fig. 3.

Deanonimisation

The first experiment is a ‘motivated intruder’ test, focused on a zero-shot deanonymisation task. More concretely, we perform two deanonymisation attacks. In the first one, we attempt to replicate the deanonymisation task presented in³, but we modify the experiment by substituting Prolific Academic participants with an LLM. In the second attack, we try to attack the HIPS method by feeding the LLM with falsified documents but instructing it that some of this information is false. In this context, we used the 1080 anonymised descriptions of the 20 celebrities and assigned GPT to guess the person this information refers to.

Deanonimisation of suppressed PII

To reduce the noise in the input of the first attack, before submitting each text, we used a regular expression to remove any tags that were inserted by Textwash, e.g. PERSON_FIRSTNAME_1, PRONOUN, DATE, etc. Specifically, we evaluated the most recent iteration of GPT-3 models provided by OpenAI, namely `text-davinci-003` and `gpt-3.5-turbo`, both being part of “GPT-3.5” series. At the time of writing, contrary to previous GPT versions, these recent models based on reinforcement learning from human feedback (RLHF)³⁸, do not support fine-tuning using the OpenAI API. To establish which one of these pre-trained LLMs performed best for this task, we randomly sampled 100 anonymised person descriptions from the dataset and tried different prompts replicating the motivated intruder task of identifying the described person. In our trials, the latest `gpt-3.5-turbo` model consistently performed better than `text-davinci-003`, by a margin of 25–30%. Moreover, we observed that for the first model, syntactic differences in prompts describing the task had negligible impact on the model’s accuracy. As such, henceforth we run every experiment using the `gpt-3.5-turbo` model. The prompt template we used for this experiment is shown in Fig. 2a.

In total, GPT deanonymised 784 texts correctly, which is 72.6% of the total. Notably, the humans in the original experiments of Kleinberg et al.³ deanonymised 285 texts, which is 26.39% of the total. Practically, using an LLM such as GPT the deanonymisation almost tripled since the correctly anonymised texts were 2.75 times the ones that humans anonymised. As it can be observed in Fig. 4, the results are consistent. For every celebrity, GPT

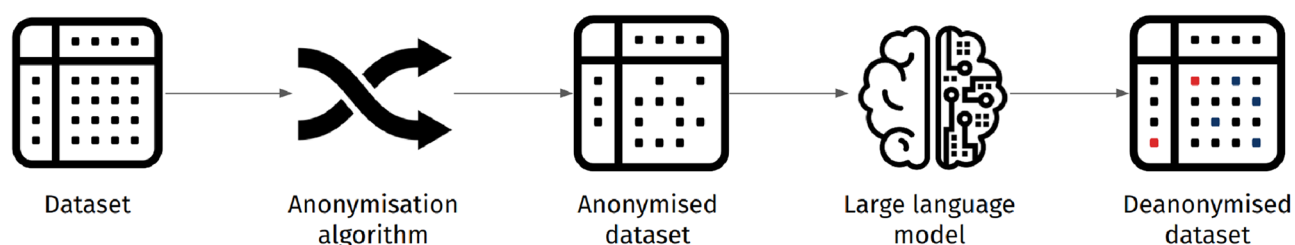


Figure 1. Deanonymisation methodology.

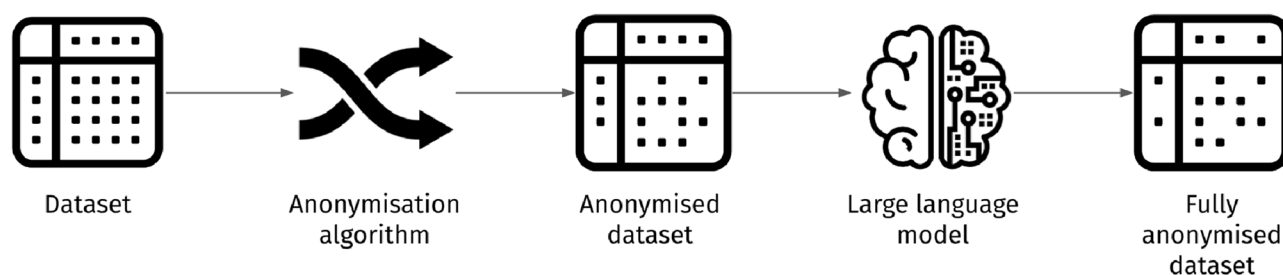


Figure 2. Proposed anonymisation methodology.

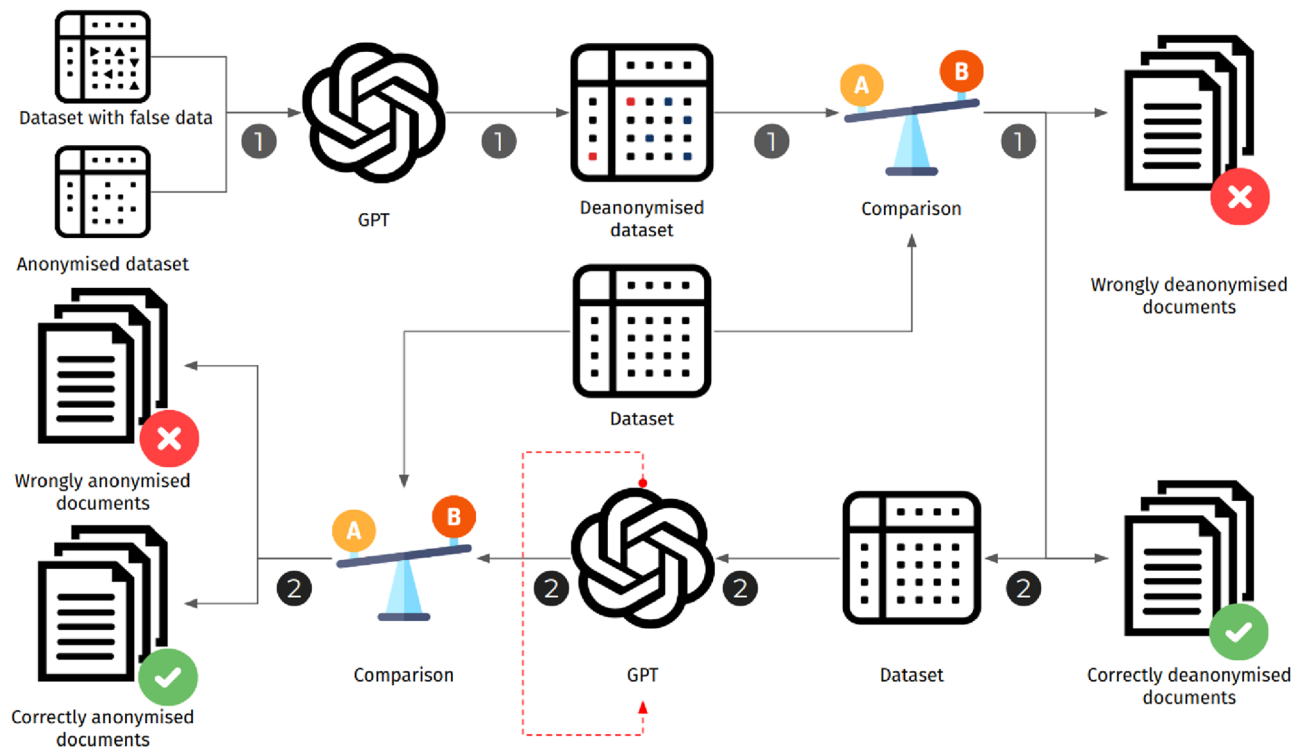


Figure 3. The data flow of the two experiments.

clearly outperforms humans in the motivated intruder test. Notably, in terms of percentages, the closest result was for D. Beckham (166% better), and the biggest difference was for D. Radcliffe (almost eight times better).

Beyond the outstanding results in deanonymisation, of specific interest are the misclassifications, illustrated in Fig. 5a. On the left hand side with blue colour, we have the 20 celebrities from the original dataset, and on the right hand side and in red, we have all the recorded misclassifications. The most often misclassifications are of J. Dench with M. Streep and D. Radcliffe with D. Craig. Nevertheless, for most of the misclassification, there are well-justified arguments why these two celebrities could be mixed with each other. Of specific interest was that in our initial experiments, we noticed several misclassifications between JK Rowling and Sam Smith. Given that they both belong to the dataset, the excellent performance of GPT in the task, and that the link between them is not direct, we decided to inspect the relevant texts. Interestingly, we noticed that the dataset providers wrongly classified eight texts in the dataset; they were attributed to JK Rowling while the text was actually describing Sam Smith and vice versa³⁹. The latter perfectly exhibits the prevalence of the deanonymisation attack and the risk from this attack vector.

Finally, we asked GPT again for a second guess for the misclassified celebrities, and 77 more were correctly identified, reaching a correct deanonymisation of 79.72%. Since there is no similar result for humans, we cannot have a fair comparison. Yet, it illustrates that there is potential for further improvements as, e.g. confidence levels for each response can be provided for each response.

Hide in plain sight

In this experiment, we aim to assess the effectiveness of the HIPS approach against LLMs, hypothesising that the introduction of fabricated data would lead to a misdirection of the GPT and significantly reduce the risk of potential privacy breaches. To this end, we utilised the anonymised data from Textwash, which includes labels indicating the intended insertion points for surrogate PII, and employed Faker⁴⁰, to generate the relevant fabricated data corresponding to the tags identified by Textwash. For each tag in the text, we used the corresponding Faker provider to generate a suitable substitute (e.g., falsified first name, occupation, city, etc.). This surrogate value then replaces every instance of the detected tag within the text, respecting the instance suffix. The tags NUMERIC and PRONOUN are handled uniquely, with the former replaced by a random integer between 1 and 10, and the latter uniformly replaced with a gender-neutral “they”. Moreover, given that the current dataset is about celebrities and the PSI tagged by Textwash as OTHER_IDENTIFYING_ATTRIBUTE usually refers to movie/song titles etc., we replaced it with a random combination of a colour and a noun. Consequently, by the end of this operation, all potential personal identifiers within the original text have been replaced by fabricated counterparts, see Fig. 6. Our approach is similar to the one described in⁴¹.

Having the falsified dataset in hand, we initially asked GPT to infer which is the famous individual described, using the same prompt as in our previous experiment (Table 2a). The model however was not able to identify any of the real celebrity targets referred to in the document, and it either replied with the surrogate names inserted, or with responses such as:

(a) Motivated intruder test
System: You are very knowledgeable about celebrities. When asked to identify individuals, disregard any missing information and context, and respond only with the name within quotes of the most likely celebrity candidate being referred to in the provided text.
User: Identify the person: [TEXT]
(b) Anonymisation
System: You are an efficient assistant. Keep your responses short and provide no explanations.
User: Anonymise and remove any personally identifiable information and related context from the given text: [TEXT]
(c) Sensitive token retrieval
System: You are an efficient assistant. Keep your responses short and provide no explanations.
User: Identify every token (including words, dates, and numerals) in the provided person description that can be connected to the person being talked about, such that if these tokens are removed the resulting text is anonymised. Return these words as a JSON-formatted list. The description of this person is: [TEXT]
(d) Hiding in plain sight motivated intruder test
System: You are immune to ambiguity and falsified facts. Given a person description, try to identify real famous individuals who match the profile described in the text, focusing on higher-level attributes and concepts, like achievements, career, views, relationships, or impact. Disregard any personal details or context that has been falsified or replaced by plausible surrogates, including pronouns, names, locations, dates, ages and any other numeric attributes. Respond only with the name within quotes of the most likely real celebrity for any provided description.
User: Identify the person: [TEXT]

Table 2. GPT-3.5 prompt templates used in our experiments.

- “Unfortunately, the description provided appears to be entirely fictional. There is no celebrity who matches this profile.”,
- “Sorry, but the description you provided has several inconsistencies and inaccuracies that make it impossible to accurately identify any real celebrities. The reported age of 10 years old, having 2 children, and being married at 10 years old are highly improbable.”.

This demonstrates that although GPT could not identify real persons with the fabricated PII in place, given its reasoning capacity, it is still able to outline the inconsistencies born of the falsified insertions in the context of HIPS anonymisation. As such, given a sufficiently large corpus of documents, an adversary might feasibly discern that the data has been manipulated. To overcome this and elicit a response from the model with HIPS in place, we had to formulate a different prompt, as shown in Table 2d. Quite remarkably, with the revised prompt the LLM found the correct individual in the first attempt in 544 cases (50.37%) and in the second attempt in 75 (69.44%) occasions. As in the previous experiment, we noticed several misclassifications and in Fig. 5b we illustrate those that appeared at least two times.

Anonymisation

In the second experiment, we explore the extent to which GPT can effectively anonymise texts, as well as identify leakage of information that could lead to deanonymisation. First, we compare the GPT anonymisation efficacy to the one of Textwash. To this, we ask GPT to anonymise all 1080 of the original texts studied in the previous experiment using the prompt of Table 2b, and for the produced outputs, we repeat the motivated intruder test¹². Note that we frame the text anonymisation as a zero-shot task without providing specific examples of what needs to be redacted in the original texts. An example of anonymised text produced by GPT is displayed in Table 3. We note that the anonymised texts produced by GPT are more readable and semantically coherent than the ones typically produced by anonymisation tools, as the redacted information is not just replaced by a tag, but the surrounding text is rewritten to maintain continuity of meaning.

In total, GPT successfully deanonymised 738 texts (729 in the first attempt and 9 in the 2nd), resulting in a deanonymisation rate of 68.3%, only 11.4% lower than what we observed for the texts anonymised with Textwash. As such, we can assume that zero-shot anonymisation using LLMs, while slightly more successful than using a state-of-the-art purpose-built anonymiser, cannot be considered an effective counter-measure mitigating the threat scenario described in Section 3. Moreover, since the texts produced are largely altered compared to their originals, it is difficult to directly compare with the outputs of Textwash and evaluate the capacity of GPT to identify PII and sensitive words.

Original: Adele is a well known British singer and songwriter. She has won an amazing amount of recognition for her work including 15 grammy awards and 9 brit awards as well as academy awards and even a golden globe for Skyfall which was released in 2012 for the James Bond film by the same name. Adele was born in London and went to school with the likes of Leona Lewis and Jessie J. Adele is very down to earth and connects to a lot of her fans on a personal level. She appears very kind and considerate of others and does a lot of charity work.
Anonymised: A well-known singer and songwriter has achieved a great deal of recognition for their work, including multiple awards. They were born in a certain city and attended school with other notable people. This individual is known for being down to earth and connecting to their fans on a personal level, as well as being kind and considerate of others and doing charity work.

Table 3. An example of anonymised text produced by GPT.

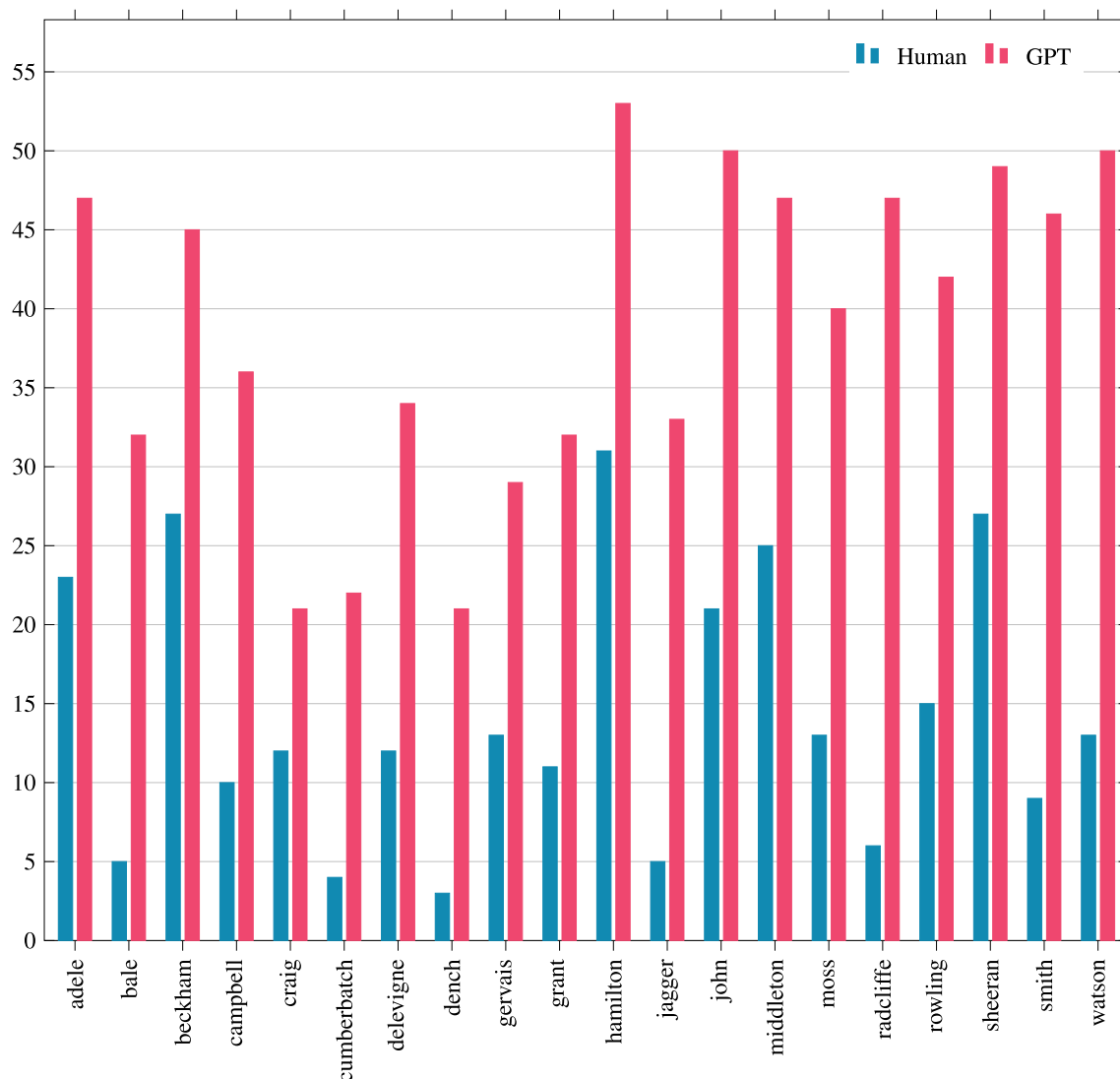


Figure 4. Successfully deanonymised text per celebrity by GPT and humans.

To this end, we focus on the 784 texts that were successfully deanonymised in the first trial of the motivated intruder test, see previous section. We prompt GPT to report the most relevant excerpts in the original texts that provide identifying information regarding the described person; see Table 2c. Note that we explicitly requested JSON-formatted output from GPT to facilitate the process of matching the exact excerpts in the original texts, as well as comparing them with the entities redacted by Textwash and their tags. Finally, we remove the returned excerpts from the original texts and repeat the motivated intruder test.

Evaluation

In total, 459 out of the 784 person descriptions were deanonymised (58.54%), which indicates that while text anonymisation can be substantially improved compared to the Textwash baseline (all of these texts were successfully deanonymised by GPT), there is still work that needs to be done to consider text anonymisation as an adequate mechanism for ensuring the privacy of text data in the era of LLMs. Next, to better understand the capacity of GPT to find identifying information, we perform a series of comparisons with the output of Textwash for the corresponding texts. To this end, we parse token by token the pairs of the original and anonymised by the Textwash texts and extract the tags of the redacted tokens, which we then compare with the ones returned by GPT to capture the subset of tokens returned by both methods and their categories. We plot the results in Fig. 7.

Interestingly, the most prevalent tag returned by Textwash, PRONOUN, is largely absent from the GPT output (2365 vs 59 instances, respectively). This means that -against human intuition- redacting the pronouns has a negligible impact on the text anonymisation for an LLM. An exception to the latter is that GPT returned the “they/them” pronouns, commonly used for non-binary individuals (see Table 4), which Textwash does not capture. For the other categories, we observe that the tokens returned by GPT are aligned with the tokens redacted by Textwash to a large extent (> 70%), further highlighting the capacity of LLMs to capture sensitive information from text, even under a zero-shot setting without task-specific training.

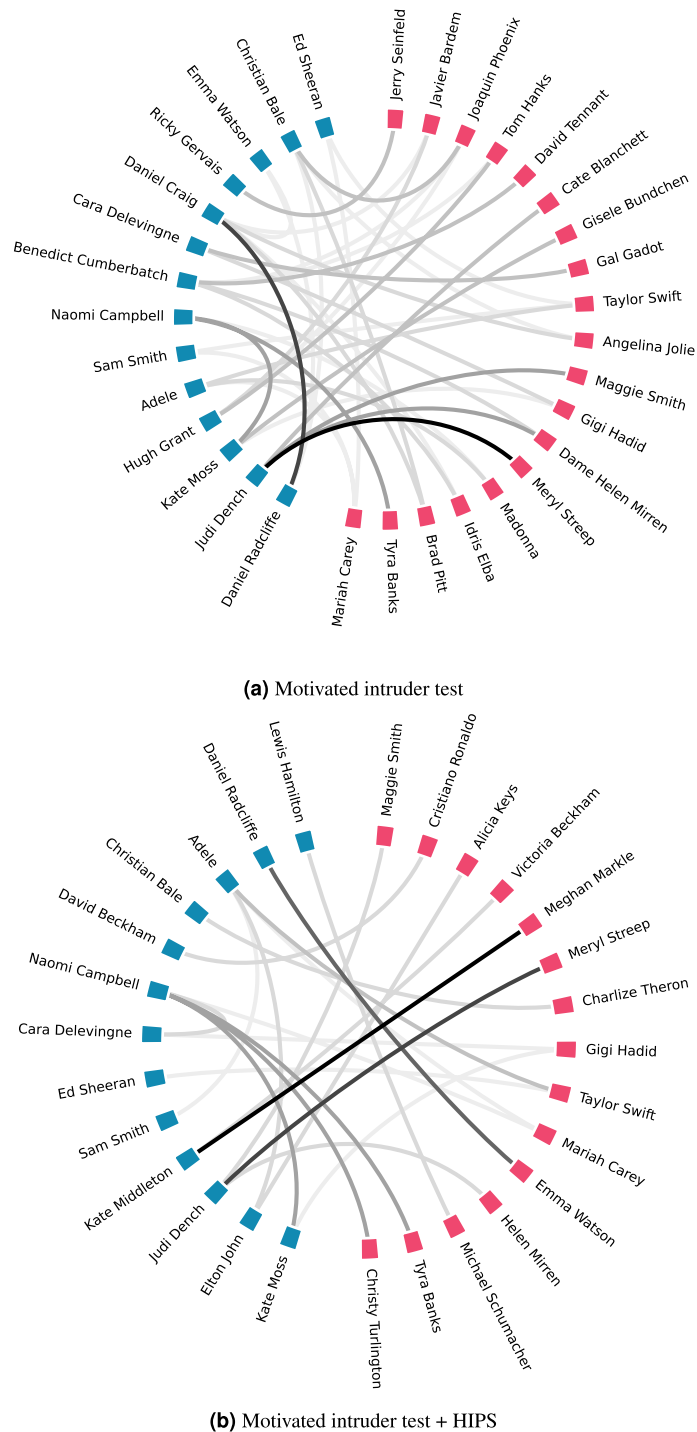


Figure 5. Misclassifications of the LLM in the two deanonymisation experiments.

Next, we measure the differences in text anonymisation based on the tokens that each method removed. We particularly focus on the proportion of anonymised text identified by Kleinberg et al. as the only statistically significant variable correlated with the deanonymisation success in their motivated intruder test. Concretely, this metric is calculated as

$$P_{anon} = 1 - \frac{ntok_{original} - ntok_{anonymised}}{ntok_{original}}$$

, where $ntok$ denotes the number of tokens in each text (original or anonymised). To this end, in Fig. 8, we plot the cumulative distribution functions (CDFs) of $P_{anon}(TW)$ (Textwash) and $P_{anon}(GPT)$ w.r.t the success our

She is an English model, actress, and singer. won Model of the Year at the British Fashion Awards in 2012 and 2014. She was at Victoria's secret for a long time after that she started acting and played in a couple of good movies. She is generous and funny. Her best acting was I believe in the movie called Paper Towns. As she models, she was in a lot of commercials which made people buy the products as she looked so beautiful. She sings plays the drums and the guitar. She has her own albums. She had her first modelling job at age ten which is quite early for someone. She has been in a busy schedule

Andrew is an Lake Johnchester, actress, and singer. won Model of the Year at the Martinez and Sons in 2013 and 1999. Andrew was at Seanland's secret for a long time after that they started acting and played in a couple of good movies. Andrew is generous and funny. Tehy best acting was I believe in the movie called Fuchsia Significant. As they models, they was in a lot of commercials which made people buy the products as they looked so beautiful. Andrew sings plays the drums and the guitar. Andrew has they own albums. Andrew had they first modelling job at age 9 which is quite early for someone. Andrew has been in a busy schedule

Figure 6. The original description of Delevigne (top) and the description with falsified information (below). We use colour coding to show the original identified potential PII (blue) and the surrogate (red).

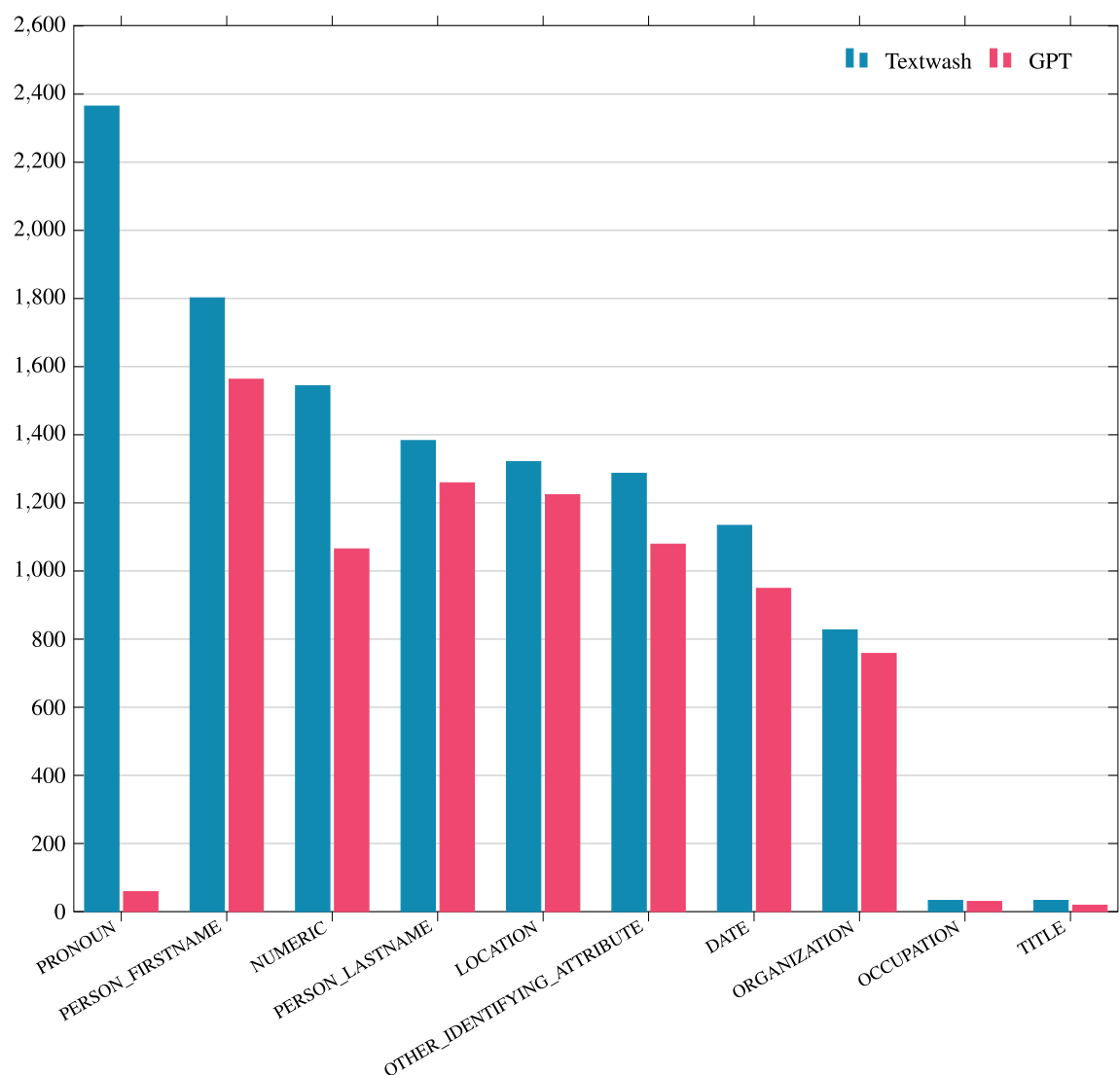


Figure 7. Removed terms.

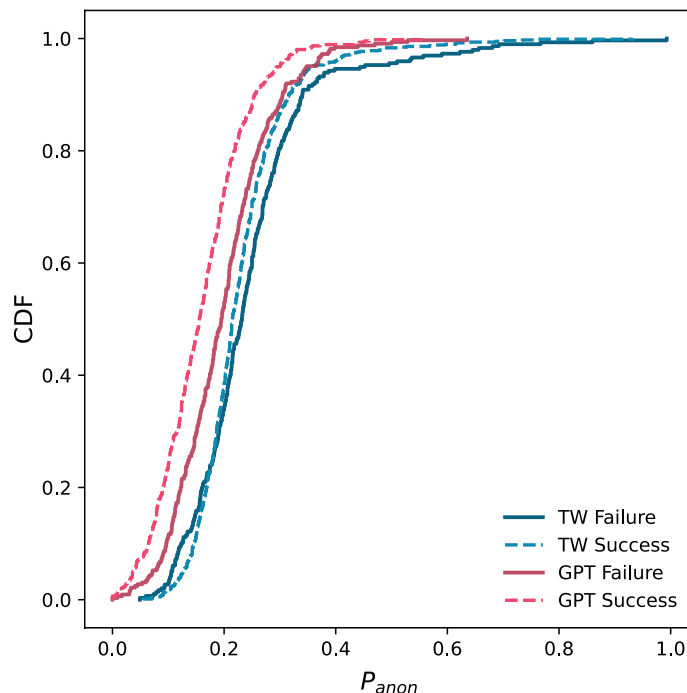


Figure 8. Proportion of text removed by each method for the cases where the motivated intruder test of person identification succeeded or failed.

GPT motivated intruder test. We observe that in both cases of successful and failed de-anonymisation, $P_{anon}(TW)$ acts as an upper bound to $P_{anon}(GPT)$, meaning that GPT is consistently more efficient than Textwash in terms of the proportion of tokens per text identified as sensitive. Expectedly, for both methods, the P_{anon} of the texts that were successfully deanonymised by GPT is lower than the cases where the motivated intruder test failed. The biggest difference is observed for $P_{anon}(GPT)$, around 5% between the instances of successful and unsuccessful de-anonymisation.

The superior performance of GPT to capture sensitive tokens motivated us to assess the impact on text anonymisation of the tokens exclusively returned by GPT compared to those captured by both methods. For this, we calculate the fraction of tokens (per text) as follows. Let \mathcal{T}_{TW} denote the set of tokens redacted by Textwash and \mathcal{T}_{GPT} denote the set of sensitive tokens returned by GPT. Then, the fraction of identified tokens that are attributed to Textwash is:

$$\mathcal{F}(TW) = \frac{|\mathcal{T}_{TW}|}{|\mathcal{T}_{TW} \cup \mathcal{T}_{GPT}|}$$

Similarly, the fraction of tokens that are exclusively attributed to GPT is:

$$\mathcal{F}(GPT) = \frac{|\mathcal{T}_{GPT} \setminus \mathcal{T}_{TW}|}{|\mathcal{T}_{TW} \cup \mathcal{T}_{GPT}|}$$

We plot the CDFs of these fractions for the texts w.r.t the success of the GPT-motivated intruder test in Fig. 9. We observe that where the test failed (anonymisation was successful), $\mathcal{F}(GPT)$ is consistently higher than $\mathcal{F}(TW)$, by a margin of 20% on average (for the cases where the identification test failed, on average the 59.5% of total sensitive tokens were exclusively identified by GPT, while Textwash captured the 40.5%). For the cases where the test succeeded, the difference between $\mathcal{F}(GPT)$ and $\mathcal{F}(TW)$ is considerably lower, with average values of 47.2% and 52.3% for Textwash and GPT, respectively. These observations further prove the capacity of GPT to capture salient tokens encompassing sensitive information (PSI), surpassing even the results of state-of-the-art anonymisation tools like Textwash, specifically fine-tuned for this task. To better explore this, in Table 4, we present the top terms lemmatised using spaCy⁴³ and bi-grams that GPT exclusively captured. Indeed, their majority comprises specific features capable of revealing the identity of each individual, as well as false negatives for Textwash, such as surname. Similar tokens were also identified by Kleinberg et al. as being responsible for the information leakage leading to successful deanonymisation (by humans). Nevertheless, we observe the existence of terms that are rather generic but provided the context they appear in, their removal obstructs the overall understanding of a reader.

Comparison with existing work

Recently, some researchers have started investigating the potential leakage of private sensitive information from language models. Lehman et al.⁴⁴ considered the case of BERT, leaking privacy data. The authors conclude that

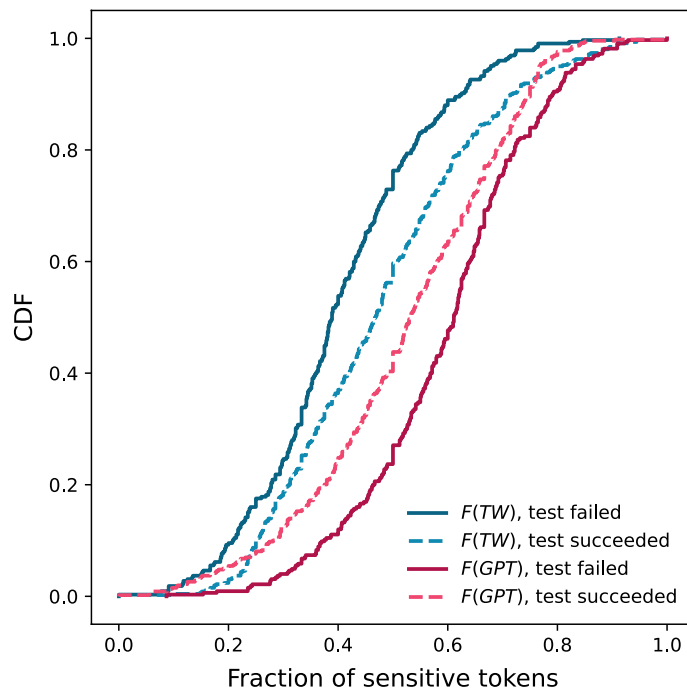


Figure 9. Fractions of the sensitive tokens exclusively captured by each method for the cases where the motivated intruder test of person identification succeeded or failed (only for the 784 descriptions successfully deanonymised in the first trial of the motivated intruder test).

while some associations can be made, they are not enough to be considered a threat. Note that in one of their attacks, the names of patients were kept in the training samples, yet the associations did not significantly increase.

Brown et al.⁴⁵ acknowledge the potential risks of large language models leaking private information and therefore highlight the need for training the models with data intended to be public and not simply publicly available data. The latter is aligned with Zimmer's approach⁴⁶, as people who may have published something do not make it public for every possible use and cannot be aware of the implications that it could have when correlated with other information. More importantly, Brown et al. stress that traditional and effective methods like differential privacy do not apply to such language models. Undoubtedly, the closest work to ours is that of Lukas et al.⁴⁷, who used GPT-2 on three datasets and successfully managed to extract PII from anonymised texts. Nevertheless, the attacks refer to generic PII and not full deanonymisation, and their threat scenario is completely different to ours. Moreover, there is no actual comparison of what a human could achieve.

Based on the above, we claim that our work is the first one to leverage LLMs to fully deanonymise a dataset of anonymised texts with such accuracy and illustrate the risks implied by the use of fine-tuned LLMs. Moreover, the comparison with human capacity as well as the identification of which pieces of information may lead to the deanonymisation of individuals has not been performed in the related literature. Finally, we should note that using other prompts, the deanonymisation not only of the person but of all other redacted information is revealed, see Table 5. As it can be observed, GPT, correctly replaced 20 redacted tags out of the total 23. Notably, the three mistakes are one for the replacement of a pronoun with the name and the two advertised companies which in the original text were Vogue and Chanel. However, a simple web search shows that the result is correct as she has advertised the products of both Burberry and Lancôme. Therefore, the misses cannot be claimed as being wrong.

The latter implies that even when the deanonymisation might not have been successful, other PII could have been extracted; however, they were not considered in this work and are left for future work. Therefore, a comparison in terms of the percentage of extracted PII to⁴⁷ would clearly show that this approach is far more efficient. Indeed, our claims are justified can be justified by the misclassifications of Fig. 5a where one can observe obvious extraction of some PII (e.g., ethnicity, age group, gender) when the deanonymisation task failed.

Conclusions

Our work puts more water to the mill in the discussion of the ethical use of AI. As common practice shows, companies will exploit the data that they have legally and ethically collected from their clients and the internet to infer more data for them. The reason is that this abundance of information can be used to personalise their products and services. As a result, they increase their quality, market penetration, and of course, their income. In fact, many companies are already touting that they can enable such functionalities while the cost for creating a trained GPT alternative, e.g. BLOOM⁴⁸ is in the scale of \$7 million, and the cost is expected to significantly drop⁴⁹. In this regard, we believe that it is only a matter of time before organisations start using LLMs on their

Celebrity	Sensitive terms exclusively captured by GPT
Elton John	singer (40), songwriter (21), marry (17), pianist (17), composer (17), award (14), song (13), piano (11), musician (9), film (8), record (8), child (7)
JK Rowling	author (22), book (15), film (10), female (10), writer (8), charity (6), benefit (6), twitter (6), billionaire (5), best-selling (4), publisher (4), controversy (4)
Christian Bale	actor (22), marry (10), award (8), batman (7), role (6), academy (4), dark (4), film (4), movie (3), act (3), weight (3), golden_globe (3)
Sam Smith	singer (39), award (24), songwriter (22), gay (18), non_binary (18), they (12), they_them (11), song (11), gender-queer (10), music (10), pronoun (9), single (8)
Daniel Radcliffe	actor (24), film (12), franchise (7), movie (5), stage (5), theatre (4), young (4), act (4), play (3), charity (3), acco-lade (3), relationship (3)
Kate Middleton	child (13), marry (13), prince (12), charity (10), mental_health (10), university (7), royal_family (7), sport (6), charity_work (6), art_history (5), mother (5), royal (5)
Naomi Campbell	model (24), supermodel (19), singer (11), actress (8), father (6), businesswoman (5), dancer (5), rehab (5), assault (5), singe (4), music_videos (4), mother (4)
Emma Watson	actress (29), activist (14), women_rights (13), model (13), feminist (11), harry_potter (8), gender_equality (6), film (6), woman (6), un_women_goodwill_ambassador (5), ambassador (5), potter (4)
Judi Dench	actress (20), film (15), award (11), marry (5), academy (5), husband (4), bafta (4), year (4), best (4), play (4), sup-port (4), love (3)
Adele	singer (31), award (20), album (18), songwriter (16), grammy (13), child (13), marry (11), divorced (9), voice (9), sell (9), someone_like_you (8), artist (7)
David Beckham	footballer (21), marry (15), model (9), retire (9), tattoo (8), professional (8), football_player (8), football (7), beckham (7), league (6), posh_spice (5), co_owner (5)
Cara Delevingne	model (37), actress (23), singer (15), eyebrow (11), sister (9), bisexual (8), act (7), guitar (7), film (7), drum (6), fashion (6), pansexual (5)
Lewis Hamilton	driver (16), race (15), racing_driver (11), win (10), podium_finishes (9), vegan (8), racism (8), black_driver (7), championship (7), father (7), motorsport (6), pole_positions (6)
Kate Moss	supermodel (24), model (22), daughter (15), drug_use (8), businesswoman (7), party_lifestyle (7), drug_allega-tions (7), drug (7), fashion (7), heroin_chic (6), moss (4), catwalk (4)
Daniel Craig	actor (11), marry (8), film (6), play (3), international_fame (3), the_girl_with_the_dragon_tattoo (3), knives_out (3), stage (2), action (2), british (2), casino_royale (2), spectre (2)
Ricky Gervais	comedian (18), actor (17), writer (13), director (9), award (9), write (8), the_office (8), podcast (7), producer (6), pop_star (6), golden (6), globe (6)
Hugh Grant	actor (24), romantic_comedies (9), film (8), bafta (7), golden_globe (7), about_boy (5), golden (5), globe (5), child (5), funny (4), marry (4), love_actually (4)
Mick Jagger	singer (20), songwriter (13), band (11), actor (10), knight (8), relationship (7), film (7), child (7), lead_singer (6), producer (6), popular (6), roll (6)
Ed Sheeran	singer (19), marry (19), guitar (16), musician (14), singer_songwriter (13), song (12), actor (11), songwriter (11), ginger_hair (10), album (10), award (8), artist (7)
Benedict Cumberbatch	actor (18), award (11), emmy (5), marry (5), cbe (4), academy (4), bbc (4), screen (4), sherlock (3), tall (3), bafta (3), theatre (3)

Table 4. Most frequent terms (1- and 2-grams) appearing in person descriptions only captured by GPT ($\in \mathcal{T}_{GPT} \setminus \mathcal{T}_{TW}$).

System: You are an efficient assistant. Find which celebrity matches the following description and match the missing information	
User: PERSON_FIRSTNAME_1 PERSON_LASTNAME_2 is a NUMERIC_4 year old actress, born in LOCATION_2 but raised in the LOCATION_1. PRONOUN is most famously known for playing PERSON_FIRSTNAME_2 PERSON_LASTNAME_1 in the OTHER_IDENTIFYING_ATTRIBUTE_2 franchise. PRONOUN played this role for NUMERIC_3 years and was it was actually PRONOUN first professional role of PRONOUN acting career. Another note-worthy performance from PERSON_FIRSTNAME_1 PERSON_LASTNAME_2 was PRONOUN role in the DATE_1 live action OTHER_IDENTIFYING_ATTRIBUTE_1, where PRONOUN played the main character PERSON_FIRSTNAME_3. PRONOUN is also very present in the fashion world, becoming a model at a young age and working with companies such as ORGANIZATION_1 and ORGANIZATION_2. More importantly, PRONOUN is very involved in activism and is a very outspoken feminist.	
Assistant:	Emma Watson is a 6 year old actress, born in Paris but raised in the UK. She is most famously known for playing Hermione Granger in the HarryPotter franchise. Emma played this role for 10 years and it was actually her first professional role of her acting career. Another note-worthy performance from Emma Watson was her role in the 2017 live action Beauty and the Beast, where she played the main character Belle. Emma is also very present in the fashion world, becoming a model at a young age and working with companies such as Burberry and Lancôme. More importantly, she is very involved in activism and is a very outspoken feminist.

Table 5. An example of a fully deanonymised text produced by GPT.

documents and realise that this way, not only can they get more visibility about their customers, but they can also deanonymise documents revealing information that would be impossible for them to do so.

As illustrated through our experiments, the use of LLMs for document deanonymisation can have a devastating impact. In fact, using an off-the-shelf commercial solution, GPT, which is not explicitly trained on such data

nor for this goal, exhibited remarkable results in document deanonymisation of celebrities by outperforming humans almost three times. We argue that the above justifies the need to consider LLMs a significant threat in the era of big data and artificial intelligence. Moreover, the results challenge what we perceive as identifying information as humans and what a machine does. While important for us features, e.g. pronouns are disregarded, even minor linguistic or knowledge hints can lead to complete deanonymisation.

While we acknowledge that this threat might have always been in the back of the mind of several researchers when discussing privacy violations and possible abuse from artificial intelligence, the misuse of big data and background knowledge, this is the first work that practically illustrates this attack over anonymised texts. Indeed, the comparison of the human capacity to anonymise celebrities by synthesising the information from relatively short anonymised texts is by far exceeded by LLMs, even if they were not trained explicitly for this task. Thus, one can safely assume that if the texts were longer, the results of the LLM would be significantly better as more correlations could be extracted. We argue that the above not only validates our realistic threat scenario but also shows that the threat can be significantly augmented as visual data can be processed alongside text⁵⁰. The above contrasts to⁵¹, yet we attribute this change to the stronger association of the new models. Note that Huang et al.⁵¹ attribute the results of the deanonymisation to memorisation due to the low association capabilities of the LLM they used. Yet, our results indicate far more advanced association capabilities and a significantly higher risk.

To this end, provided the efficacy of our adaptation of the HIPS algorithm to deceive the GPT, we believe that a promising direction toward effective text anonymisation in the era of LLMs would be to consider a setting where hard-to-detect yet misleading cues and strategical replacement of PII is introduced to steer the responses of LLMs away from the true identifiers in the case of deanonymisation attacks. This way, the privacy of individuals would be effectively protected since LLM responses would appear valid, but would actually contain the falsified information that was intentionally inserted. Nevertheless, the above implies the need to develop new metrics that are LLM oriented so that one can assess whether a document is more prone to deanonymisation than others and which parts of the texts are more sensitive than others. For instance, nouns have less impact on LLMs than on humans and removing them significantly decreases the quality of the text. Thus, it might be wiser to leave the nouns and focus on other parts of the text. Similarly, randomly faked data might be easily spotted by the LLM justifying the high deanonymisation percentage, however, context-aware methods could be more efficient. For instance, rather than having random data, one could inject noise that could hint at another member of the dataset, making the adversary's task even harder. Furthermore, generative AI could be used to interweave these errors better in the text. Overall, we argue that LLMs will definitely play a central role in text anonymisation in the forthcoming years resulting in many revisions in the field for both attacking and defending methods.

Data availability

The datasets analysed during the current study are available from the corresponding author upon reasonable request.

Received: 22 March 2023; Accepted: 17 September 2023

Published online: 25 September 2023

References

1. 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). Official Journal of the European Union, Vol. L119 (4 May 2016) 1–88 (2016).
2. Samaratiy, P. & Sweeney, L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression (1998).
3. Kleinberg, B., Davies, T. & Mozes, M. *Textwash-Automated Open-source Text Anonymisation*. arXiv preprint [arXiv:2208.13081](https://arxiv.org/abs/2208.13081) (2022).
4. Narayanan, A. & Shmatikov, V. Myths and fallacies of “personally identifiable information”. *Commun. ACM* **53**, 24–26 (2010).
5. Agencia Española de Protección de Datos and European Data Protection Supervisor. 10 misunderstandings related to anonymisation. https://edps.europa.eu/data-protection/our-work/publications/papers/aepd-edps-joint-paper-10-misunderstandings-related_en (2021).
6. Zigomitos, A., Casino, F., Solanas, A. & Patsakis, C. A survey on privacy properties for data publishing of relational data. *IEEE Access* **8**, 51071–51099 (2020).
7. Mamede, N., Baptista, J. & Dias, F. Automated anonymization of text documents. In *2016 IEEE congress on evolutionary computation (CEC)* 1287–1294 (IEEE, 2016).
8. Hassan, F., Domingo-Ferrer, J. & Soria-Comas, J. Anonymization of unstructured data via named-entity recognition. In *Modeling Decisions for Artificial Intelligence* (eds Torra, V. et al.) 296–305 (Springer International Publishing, Cham, 2018).
9. Kleinberg, B. & Mozes, M. Web-based text anonymization with node.js: Introducing NETANOS (named entity-based text anonymization for open science). *J. Open Source Softw.* **2**, 293. <https://doi.org/10.21105/joss.00293> (2017).
10. Sweeney, L. Replacing personally-identifying information in medical records, the scrub system. In *Proceedings of the AMIA annual fall symposium* 333 (American Medical Informatics Association, 1996).
11. Neamatullah, I. et al. Automated de-identification of free-text medical records. *BMC Med. Inform. Decis. Mak.* **8**, 1–17 (2008).
12. Ruch, P., Baud, R. H., Rassinaux, A.-M., Bouillon, P. & Robert, G. Medical document anonymization with a semantic lexicon. In *Proceedings of the AMIA Symposium* 729 (American Medical Informatics Association, 2000).
13. Derroncourt, F., Lee, J. Y., Uzuner, O. & Szolovits, P. De-identification of patient notes with recurrent neural networks. *J. Am. Med. Inform. Assoc.* **24**, 596–606 (2017).
14. Manning, C. D. et al. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* 55–60 (2014).
15. Liu, Z., Tang, B., Wang, X. & Chen, Q. De-identification of clinical notes via recurrent neural network and conditional random field. *J. Biomed. Inform.* **75**, S34–S42. <https://doi.org/10.1016/j.jbi.2017.05.023> (2017).
16. Sánchez, D., Batet, M. & Viejo, A. Detecting sensitive information from textual documents: An information-theoretic approach. In *Modeling Decisions for Artificial Intelligence: 9th International Conference, MDAI 2012, Girona, Catalonia, Spain, November 21–23, 2012. Proceedings* 9 173–184 (Springer, 2012).

17. Sánchez, D., Batet, M. & Viejo, A. Automatic general-purpose sanitization of textual documents. *IEEE Trans. Inf. Forensics Secur.* **8**, 853–862. <https://doi.org/10.1109/TIFS.2013.2239641> (2013).
18. Sánchez, D., Batet, M. & Viejo, A. Utility-preserving privacy protection of textual healthcare documents. *J. Biomed. Inform.* **52**, 189–198 (2014).
19. Özlem Uzuner, Sibanda, T. C., Luo, Y. & Szolovits, P. A de-identifier for medical discharge summaries. *Artif. Intell. Med.* **42**, 13–35. <https://doi.org/10.1016/j.artmed.2007.10.001> (2008).
20. Carrell, D. *et al.* Hiding in plain sight: Use of realistic surrogates to reduce exposure of protected health information in clinical text. *J. Am. Med. Inform. Assoc.* **20**, 342–348 (2013).
21. Meystre, S. M., Friedlin, F. J., South, B. R., Shen, S. & Samore, M. H. Automatic de-identification of textual documents in the electronic health record: A review of recent research. *BMC Med. Res. Methodol.* **10**, 1–16 (2010).
22. Walsh, C. G. *et al.* Enabling open-science initiatives in clinical psychology and psychiatry without sacrificing patients' privacy: Current practices and future challenges. *Adv. Methods Pract. Psychol. Sci.* **1**, 104–114 (2018).
23. Lison, P., Pilán, I., Sánchez, D., Batet, M. & Øvrelid, L. Anonymisation models for text data: State of the art, challenges and future directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* 4188–4203 (2021).
24. Devlin, J., Chang, M., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C. & Solorio, T. (eds.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. <https://doi.org/10.18653/v1/n19-1423> (Association for Computational Linguistics, 2019).
25. Liu, Y. *et al.* Roberta: A robustly optimized bert pretraining approach. arXiv preprint [arXiv:1907.11692](https://arxiv.org/abs/1907.11692) (2019).
26. Xia, P., Wu, S. & Van Durme, B. Which* bert? a survey organizing contextualized encoders. arXiv preprint [arXiv:2010.00854](https://arxiv.org/abs/2010.00854) (2020).
27. Brants, T., Popat, A. C., Xu, P., Och, F. J. & Dean, J. Large language models in machine translation. In *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (ed. Eisner, J.) 858–867 (ACL, 2007).
28. Vaswani, A. *et al.* Attention is all you need. In Guyon, I. *et al.* (eds.) *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA 5998–6008* (2017).
29. Brown, T. *et al.* Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
30. Selbst, A. D. Negligence and AI's human users. *BUL Rev.* **100**, 1315 (2020).
31. Smith, H. & Fotheringham, K. Artificial intelligence in clinical decision-making: Rethinking liability. *Med. Law Int.* **20**, 131–154 (2020).
32. Brundage, M. *et al.* Toward trustworthy AI development: Mechanisms for supporting verifiable claims. arXiv preprint [arXiv:2004.07213](https://arxiv.org/abs/2004.07213) (2020).
33. European Commission. Proposal for a regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206> (2021).
34. GitHub. Copilot. <https://github.com/features/copilot> (2023).
35. OpenAI. Codex. <https://openai.com/blog/openai-codex> (2023).
36. Toulas, B. GitHub Copilot update stops AI model from revealing secrets. <https://www.bleepingcomputer.com/news/security/github-copilot-update-stops-ai-model-from-revealing-secrets/> (2023).
37. OpenAI. API data usage policies. <https://openai.com/policies/api-data-usage-policies> (2023).
38. Ouyang, L. *et al.* Training language models to follow instructions with human feedback. arXiv preprint [arXiv:2203.02155](https://arxiv.org/abs/2203.02155) (2022).
39. Textwash source code, misclassified sample. https://github.com/maximilianmozes/textwash/blob/d347f40cab948dbcb522e3f8829c2a9b05bd7fbc/paper/study2/person_descriptions/orig/rowling_69.txt (2023).
40. Faker. <https://github.com/joke2k/faker>.
41. Osborne, J. D. *et al.* Bratsynthetic: Text de-identification using a markov chain replacement strategy for surrogate personal identifying information. arXiv preprint [arXiv:2210.16125](https://arxiv.org/abs/2210.16125) (2022).
42. Mackey, E., Elliot, M. & O'Hara, K. *The Anonymisation Decision-Making Framework* (UKAN Publications, 2016).
43. spaCy. <https://spacy.io> (2023).
44. Lehman, E., Jain, S., Pichotta, K., Goldberg, Y. & Wallace, B. C. Does BERT pretrained on clinical notes reveal sensitive data? In Toutanova, K. *et al.* (eds.) *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6–11, 2021 946–959*. <https://doi.org/10.18653/v1/2021.naacl-main.73> (Association for Computational Linguistics, 2021).
45. Brown, H., Lee, K., Mireshghallah, F., Shokri, R. & Tramèr, F. What does it mean for a language model to preserve privacy? In *2022 ACM Conference on Fairness, Accountability, and Transparency* 2280–2292 (2022).
46. Zimmer, M. "but the data is already public": On the ethics of research in facebook. *Ethics Inf. Technol.* **12**, 313–325 (2010).
47. Lukas, N. *et al.* Analyzing leakage of personally identifiable information in language models. *CoRR*. <https://doi.org/10.48550/arXiv.2302.00539> (2023). 2302.00539.
48. BLOOM. <https://huggingface.co/bigscience/bloom-7b1> (2023).
49. Wiggers, K. A year in the making, BigScience's AI language model is finally available. <https://techcrunch.com/2022/07/12/a-year-in-the-making-bigsciences-ai-language-model-is-finally-available/> (2022).
50. Suris, D., Menon, S. & Vondrick, C. ViperGPT: Visual inference via python execution for reasoning. *CoRR*. <https://doi.org/10.48550/arXiv.2303.08128> (2023). 2303.08128.
51. Huang, J., Shao, H. & Chang, K. C. Are large pre-trained language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7–11, 2022* (eds. Goldberg, Y., *et al.*) 2038–2047 (Association for Computational Linguistics, 2022).

Acknowledgements

This work was supported by the European Commission under the Horizon 2020 Programme (H2020), as part of the project HEROES (<https://heroes-fct.eu/>) (Grant Agreement no. 101021801) and under the Horizon Europe Programme, as part of the project LAZARUS (<https://lazarus-he.eu/>) (Grant Agreement no. 101070303). The content of this article does not reflect the official opinion of the European Union. Responsibility for the information and views expressed therein lies entirely with the authors.

Author contributions

C.P. conceived of the presented idea. N.L. conducted the experiments. All authors wrote and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to C.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023