De-Identifying Student Personally Identifying Information with GPT-4

Shreya Singhal[1], Andres Felipe Zambrano[2], Maciej Pankiewicz[3], Xiner Liu[2], Chelsea Porter[2], Ryan S. Baker[2]
[1]Indian Institute of Technology, Madras
[2]University of Pennsylvania
[3]Warsaw University of Life Sciences
{singhalshreya201, maciekpankiewicz, ryanshaunbaker}@gmail.com,
{azamb13, xiner, cfporter}@upenn.edu

ABSTRACT

Education is increasingly taking place in learning environments mediated by technology. This transition has made it easier to collect student-generated data including comments in discussion forums and chats. Although this data is extremely valuable to researchers, it often contains sensitive information like names, locations, social media links, and other personally identifying information (PII) that must be carefully redacted before utilizing the data for research to protect their privacy. Historically, this task of redacting PII has been painstakingly conducted by humans; more recently, some researchers have attempted to use regular expressions and supervised machine-learning methods. Nowadays, with the recent high performance shown by Large Language Models in a wide range of tasks, they have become another alternative to be explored for de-identifying educational data. In this work, we assess GPT-4's performance in de-identifying data from discussion forums in 9 Massive Open Online Courses. Our results show an average recall of 0.958 for identifying PII that needs to be redacted, suggesting that it is an appropriate tool for this purpose. Our tool is also successful at identifying cases missed by humans when redacting data. These findings indicate that GPT-4 can not only increase the efficiency but also enhance the quality of the redaction process. However, the precision of such redaction is considerably worse (0.526), over-redacting names and locations that do not represent PII, showing a need for further improvement.

Keywords

De-identification, Anonymization, GPT-4, Large Language Models, Privacy, Massively Open Online Course.

1. INTRODUCTION

In the digital era, the proliferation of new educational technologies and platforms has increased the availability of data resources for researchers. There has been a considerable expansion in research involving textual data, collected from discussion forums, chat sessions, transcripts of classroom and human-tutor dialogues, and other sources as well. Although this data holds immense potential for insights into student behavior, pedagogical effectiveness, and communication patterns, the presence of Personally Identifying Information (PII) in such data sets introduces critical ethical and legal challenges. For instance, many countries and regions have strict data protection laws and regulations, such as the General Data Protection Regulation (GDPR) in the European Union [2]. Therefore, to protect the privacy of students and participants, this data needs to be de-identified before it can be shared with other researchers. This step becomes even more critical when datasets are released publicly. The public release of data is valuable for open science, enabling other researchers to replicate past studies or explore new questions, but requires steps to ensure the privacy and confidentiality of the participants involved.

De-identifying data presents multiple challenges due to the varied nature of PII. While certain types of PII, like mail and email addresses, phone numbers, or personal webpage links, can be readily identified using methods such as regular expressions or supervised machine learning natural language processing (NLP), others pose greater difficulties. For instance, nicknames or terms that can function as both names and common dictionary terms [5] are harder to detect and redact. Additionally, not all instances of names or locations constitute PII, such as when mentioning the author of an article, a political leader, or a city where some historical event discussed in class happened. Moreover, students can also make misspellings, grammatical errors, or variations in punctuation and spacing or use words from different languages that do not correspond to the predominant language used in the training process of the supervised learning technique.

Despite these challenges, supervised learning techniques have demonstrated promising results in de-identifying data, achieving recall rates above 0.95 in redacting student names (refer to the related work section for details; [1]). However, these models require a ground truth dataset for training and may not generalize to cases where that ground truth dataset is not applicable. The creation of such a dataset, along with the time needed for training and developing the tool, represents a significant time investment. Additionally, manual de-identification by humans is not infallible, as coders may inadvertently overlook instances of PII or misclassify data, thus reducing the effectiveness of supervised learning models in real-life applications. Having a second pass by a different human coder can reduce this risk but increases the time cost substantially.

An alternative approach for de-identification that could be easier to implement is using a Large Language Model (LLM) such as GPT-4. Some research has shown that GPT-4 can identify personal names within data sets [7, 10]. Given GPT-4’s success in this analogous task, this paper investigates employing GPT-4 as a tool for de-identifying text data generated by students. We specifically explore the capabilities and limitations of GPT-4 in de-identifying discussion forum posts from students in 9 Massive Open Online Courses (MOOCs), looking at GPT-4's precision and recall in identifying PII, comparing this performance with current benchmark models, and investigating whether it can detect cases of PII missed by human coders.

2. RELATED WORK

2.1 De-identification with Pattern Matching and Supervised Learning

The two main approaches that have been explored for identifying and redacting PII in text are pattern matching and supervised machine learning techniques. For example, Kyaalp et al. [5] employed a database of 3.8 million names collected from Social Security to de-identify personal names from medical reports. The challenge observed by the authors was that many names coincide with common dictionary terms. A similar approach, employing a list of first and last names of students, has been employed in the education field for de-identifying STEM laboratory reports. This approach achieved a high precision of 0.79 and recall of 0.75 [12], but would be difficult to scale, as the approach requires having a list of student names.

Regular expressions are another pattern matching method employed for de-identification. Farrow et al. [3] compared the performance of regular expressions with the previous approach of employing the list of first and last names of students, showing an improvement from 0.515 to 0.876 in recall but a reduction from 0.879 to 0.567 in precision, using data collected across 6 sessions of a Master distance-learning course. Although the magnitude of the improvement in recall is comparable with the drop in precision, the authors argued that recall is more important than precision due to the high risk of unredacted PII. The authors also examined a hybrid approach combining class lists and regular expressions, which slightly increased recall to 0.905, at a further cost to precision (0.550).

Supervised machine learning techniques have also been employed for redacting PII. Bosch et al. [1] used the extra-tree variant of random forest and deep neural networks for de-identifying discussion forum text data from 2 online courses offered by a public university in the United States. The authors incorporated features that considered the position of each word within a sentence, its presence in U.S. census lists, and its appearance on lists of cities, political regions, or countries worldwide. They also considered the occurrence of each word among standard dictionary terms, including words that were within one or two edits of difference to accommodate potential misspellings. Their results demonstrated better performance than earlier pattern-matching techniques. When averaging the outputs of both machine learning models, the study achieved a recall of 0.970 and a precision of 0.827. Additionally, they reached a Cohen's Kappa of 0.794, which closely approaches the original Kappa of 0.864 reported for the agreement between two human coders in their study.

2.2 LLM-based De-identification

Transformer models have shown recall over 0.99 for de-identifying medical datasets [9]. Inspired by this high performance in the medical field, Holmes et al. [4] used two fine-tuned transformer models based on RoBERTa (a pre-trained large language model; [6]) to de-identify data from essays submitted by students in a MOOC, focusing on names exclusively. This approach obtained a recall of 0.84 and a precision of 0.68. They compared these models with a rule-based student name labeling system based on a general-purpose Name Entity Recognition (NER) model, which obtained a lower performance (recall of 0.81 and precision of 0.33).

Recently, researchers have begun using the GPT family of Large Language Models for related tasks. Qin et al. [10] used GPT-3.5 to identify names in news articles, achieving an F1-score of 0.532 for general name entity recognition using GPT 3.5 and an F1-score of 0.872 for identifying personal names (precision and recall were not reported). Liu et al. [7] then investigated the use of GPT-4 for de-identifying medical reports, reporting that GPT-4, using a zero-shot prompt, achieved an accuracy of 0.99, surpassing the 0.947 accuracy of a fine-tuned RoBERTa model on the same dataset. Although the authors did not provide precision and recall metrics for these models, this comparative improvement suggests GPT-4's potential as an effective tool for de-identifying data. Within this study, we investigate whether GPT-4 can also be successfully used as a tool for de-identifying data collected from educational contexts.

3. METHODS

3.1 Data

The dataset used in this study consists of a collection of forum posts from students enrolled in nine different Massive Open Online Courses (MOOCs) at the University of Pennsylvania between 2012 and 2015. These courses covered a variety of subjects: accounting, calculus, design, gamification, business trends, poetry, mythology, probability, and vaccines. This diversity in course topics was strategically chosen to mitigate bias towards any specific domain and to enhance the generalizability of our findings across different courses.

Our initial dataset was compiled by randomly sampling 500 forum posts from each of the nine courses, 4500 posts in total. To ensure relevance and uniformity, posts written in languages other than English or that consisted only of special symbols, characters, website links or mathematical formulas were excluded. Our final dataset comprised 3,505 forum posts from 2,882 unique students. The number of posts was approximately equally distributed across all the 9 courses (379 to 399 posts per course). We maintained the original text of the posts, including any typographical or grammatical errors, as these elements are intrinsic to the natural language processing challenges we aimed to investigate.

3.2 Human De-identification Process: First Iteration

Three human reviewers were tasked with manually redacting any PII from the posts. This redaction process involved removing names, contact details, city or country of origin or residence, links to any personal website, and any other information that could potentially reveal the identity of a post's author. The two redactors were Masters students trained by a faculty member in the process of redaction. The workload was evenly distributed between two of the three reviewers, with each handling approximately half of the posts from each course (about 195 posts per course per reviewer). The third coder, an experienced professional consultant with around 20 years of experience in document editing and transcription, including redaction, conducted an additional review to address any potential errors made by the first two coders.

As in past work in the EDM community (e.g. [1]), we gave the redactors instructions as to what types of names to retain (names of famous people and authors) and what types of names to redact (names of instructors and students), but asked them to use their judgment in selecting which was which. In many past papers in other fields (see review in [8]), human redactors are given a list of known student names for redaction, but doing so here would have biased in favor of human redactors by giving them information the LLM did not have. In total, human coders redacted 2134 words, with 1282 posts containing at least one redaction (36.6% of the total posts). This process formed the first ground truth dataset, against which the efficacy of LLMs in performing similar redactions was evaluated.

3.3 De-identification Process with GPT-4

For the de-identification process, we employed the OpenAI gpt-4-0613 model, accessed through its API. Data was sent to OpenAI under its privacy policy at the time of the work and this writing, which guarantees that data will only be used for the user's intended purpose and to verify reports of abuse, and that data will only be retained for 30 days before it is deleted. The max_tokens parameter, which defines the maximum length of the model's response, was set to 1000. This limit was sufficient to accommodate the length of each post in our dataset. All other parameters were set to default. Each forum post from the original set of posts was sent individually in a request to the model using the following prompt. (Please note that the PII in this prompt was modified to prevent the disclosure of any student's personal details. Throughout this paper, we employ fictitious links and pseudonyms in all examples.)

Prompt:

Please edit the provided text by removing any personally identifiable information (PII). This includes names, company names, places of origin, current living locations, addresses, and social media links. Replace all removed PII with '[REDACTED]'. Ensure that the rest of the text remains unchanged, word for word. Maintain the original punctuation, quotation marks, spaces, and line breaks. If the text does not contain any PII, return it as is.

For example, if the input is:

Dear All,You'll find below a Linkedin group I just create.http://www.linkedin.com/groups?home=12345&abcd=789hmIt will be hold in English (as a second language) and is open to any student of Coursera for the 'Introduction to Financial Accounting'.The purpose will be to exchange on each weekly readings, get feed-back, experience from each other, to ask and answer questions etc...Link you soon!Let's team work!

The output must be:

Dear All,You'll find below a Linkedin group I just create.[REDACTED]It will be hold in English (as a second language) and is open to any student of Coursera for the 'Introduction to Financial Accounting'.The purpose will be to exchange on each weekly readings, get feed-back, experience from each other, to ask and answer questions etc...Link you soon!Let's team work!

Please repeat this process with the following post:

[POST TO BE DE-IDENTIFIED]

Our prompt specifies the exact types of PII to be redacted, such as names, company names, places of origin, current living locations, addresses, and social media links. In an early draft of our prompt, we also explicitly instructed GPT not to redact the names of public figures such as artists or politicians. However, this clarification caused GPT to have lower overall performance for both precision and recall. We also requested GPT-4 to maintain the integrity of the original text in terms of structure and formatting (word by word) because, without this, GPT-4 corrected grammar or punctuation mistakes from the original posts, and even changed some words to enhance the clarity of the original post.

3.4 Second Iteration of Human De-identification and GPT Evaluation

To evaluate the performance of the GPT-based de-identification process, we assessed the level of agreement between the outputs from the GPT-4 model and the human de-identified posts by comparing them word-by-word. Two types of discrepancies were observed during this evaluation: disagreements and cases where the two approaches agreed but redacted in different ways. For example, in a LinkedIn URL, the human redacted code was *"Connect with me at: [REDACTED]"*, while the GPT code was *"Connect with me at: LinkedIn: [REDACTED]"*. Here, GPT-4 included *"LinkedIn"* in the redaction, whereas the human coder treated the entire phrase as PII. GPT repeated this addition of the social network of the corresponding link in several instances. Similarly, one of the messages that human coders redacted as *"Thanks Mr [REDACTED]. It is a very interesting class..."*, was redacted by GPT-4 redacted as *"Thanks [REDACTED]. It is a very interesting class..."*, also removing the title. Both humans and GPT-4 were inconsistent at including titles such as *"Mr"*, *"Mrs"*, *"Prof"*, and others in the redacted version of the posts.

To ensure comparability between GPT and human-de-identified posts, we manually corrected the above-mentioned differences before calculating agreement/disagreement. We removed all articles, non-alphanumeric words, and punctuation and then compared each of them word by word to also address instances where GPT corrected non-alphanumeric symbols or grammar, spelling, or punctuation symbols (even after requesting in the prompt to avoid these corrections).

After correcting for these low-level differences, the remaining 45 discrepancies corresponded to disagreements between human and GPT-based de-identification. We manually checked each of them, rectifying each case where human coders failed to detect PII. Following these corrections, we evaluated GPT's performance using the updated gold standard (human-based ground truth with corrections from GPT). The distribution of redacted elements in each file for each course and the distribution of redactions per post (both after correction) are given in Table 1. For those interested in using our approach, the code developed for the GPT-based de-identification process can be accessed at https://github.com/pcla-code/llm-de-identification.

The data presented in the table reveals that the *Design* course had the highest percentage of posts with redactions (51.1%), while also being one of the courses with the shortest posts (44.4 words per post). *Poetry* had a distinctively higher average word count per post at 241 words. *Poetry* also had the smallest number of redactions per post (0.31), pointing to more in-depth discussions or the presence of extracts from other poems or self-creations without necessarily adding any PII.

3.5 Evaluation

After correcting the human-based de-identification, we assessed the performance of the GPT-4 model in redacting PII, using precision, recall, and Cohen's Kappa. The confusion matrix was defined as:

- True Positive (TP): Words identified as PII by human coders as well as GPT-4.
- True Negative (TN): Words that were not identified as PII by either human coders or GPT-4.

Table 1. Distribution of posts, words by post and redacted elements across all the courses after corrections. The percentage of posts with at least one redacted element, the percentage of redacted words for each course, and the percentage of redacted words initially missed by humans are shown in parentheses.

Course Topic	Total Posts	Posts with Redactions	Words per Post	Redacted Words	PII Initially Missed by Human coders
Accounting	387	165 (42.6%)	47.0	251 (1.4%)	6 (3.6%)
Calculus	396	114 (28.8%)	40.0	162 (1.0%)	3 (2.6%)
Design	380	194 (51.1%)	44.4	283 (1.7%)	8 (4.1%)
Gamification	379	124 (32.7%)	63.5	237 (1.0%)	7 (5.6%)
Business trends	387	143 (37.0%)	81.1	291 (0.9%)	15 (10.5%)
Poetry	399	85 (21.3%)	241.0	124 (0.1%)	2 (2.4%)
Mythology	390	138 (35.4%)	67.9	196 (0.7%)	1 (0.7%)
Probability	396	117 (29.5%)	56.9	177 (0.8%)	0 (0%)
Vaccines	391	159 (40.7%)	71.2	287 (1.0%)	3 (1.9%)

- False Positive (FP): Words that were identified as PII by GPT-4 but not by human coders.
- False Negative (FN): Words that were identified as PII by human coders but not by GPT-4.

To calculate the metrics, we first identified the TP, TN, FP, and FN at the word-level following the above-mentioned definition. Then we calculated precision, recall, and Cohen's Kappa at the course-level. Finally, we calculated the overall average of each metric across the 9 courses. To mitigate potential inconsistencies within the GPT-based de-identification process, we sent the data to GPT-4 three times and evaluated the performance metrics for each iteration. We then averaged across iterations.

4. RESULTS

4.1 Human Mistakes

In inspecting our results, we found that after the first iteration of human-based de-identification, human coders failed to redact 45 words that involve PII distributed across all the courses (See table 2). For example, in a post from a *Business Trends* course, the human-coded version was:

"I would think that the replacement of retired workers with young ones is more complicated than I proposed. You're so right about the technology factor, Laura. That is a major point."

In this case, the human coder failed to remove *"Laura"*, who is clearly the name of the author of the post. By contrast, GPT-4 correctly identified this as PII and redacted it. In another instance from a different course, the post obtained after human redaction was:

"interesting that the LinkedIn example came up during the third set of lectures :)has anyone seen the Fun Theory site Michael mentioned???"

In this case as well, the coder did not remove *"Michael,"* who appears to be someone the author of the post is addressing within the course. Human-coders also missed some personal webpages. For example, the post

"...What I do is make this;http://www.personawebpage.com/blogs; once in a while...It makes really colorful bowls which i store tiny things in :);"

was not redacted by the human coders, despite the link leading to a personal webpage that disclosed personal information of one student. Although these examples of human errors represented only a small part of the overall disagreements, they demonstrate that humans can make errors, indicating that even a fairly thorough process such as the one used here might be insufficient to reliably fully de-identify this type of data. This potential risk of human mistakes has also been observed by Bosch et al. [1], who found 37 disagreements between 2 human coders across 600 possible names in their dataset (6.1%). Our results show that the analysis of disagreements between human-based and automated de-identification can contribute to mitigating this issue and improving the quality of de-identification.

4.2 Performance

Table 2 summarizes the average of 3 runs of the de-identification process for each metric, for each course included in this study, considering human redaction as the ground truth (after correcting the human mistakes). The recall rate was consistently over 0.85 for all courses examined. However, precision was lower than 0.75 in all cases. These results show that GPT identified almost all PII. However, it often failed to recognize names, locations, or links that are not PII (such as famous people). While not ideal, as GPT over-redacts some information, it keeps sensitive information protected, which may be an acceptable trade-off for enhanced privacy [4].

Cohen's Kappa, which assesses the agreement between the GPT-4 model and human coders considering the distribution of both classes (PII and no PII words), varied significantly across courses. The highest Kappas were observed for *Design* (Kappa=0.843), *Accounting* (Kappa=0.824), *Gamification* (Kappa=0.780), *Calculus* (Kappa=0.771), and *Probability* (Kappa=0.763). In contrast, for the *Poetry* (Kappa=0.267), *Business Trends* (Kappa=0.527), *Mythology* (Kappa=0.530), and *Vaccines* (Kappa=0.612) courses, GPT and human-based redactions demonstrated considerably lower agreement. GPT's performance appears to be lower in courses characterized by longer average post lengths, all of them exceeding 65 words (Poetry has a much higher average of 241 words per post). Additionally, in contrast to courses where GPT obtains better performance (many of them related to mathematics), these courses typically involve more qualitative discussions where names or locations—such as those of artists or leaders—should not necessarily be redacted. This distinction between PII and names or locations that do not need to be redacted could lead to confusion for GPT, still resulting in high recall but showing a low precision in the de-identification process. This limitation is addressed in earlier list-based approaches (i.e. where the algorithm is provided with a list of names to redact), but that approach makes generalization much more difficult.

Table 2. Performance metrics of GPT-based de-identification process considering human redaction as our ground truth.

Course	Precision	Recall	Kappa
Accounting	0.716	0.975	0.823
Calculus	0.666	0.922	0.771
Design	0.742	0.984	0.843
Gamification	0.658	0.967	0.780
Business trends	0.366	0.983	0.527
Poetry	0.158	0.895	0.267
Mythology	0.374	0.937	0.530
Probability	0.602	0.988	0.763
Vaccines	0.454	0.966	0.612
Average	**0.526**	**0.958**	**0.657**

Table 3. Comparison with other state of the art de-identification methods.

Paper	Bosch et al. [1]	Farrow et al. [3]	Holmes et al. [4]	This Paper
PII	Names	Names	Names	Names, Locations and Links
Method	Extra-trees + Deep Neural Nets	Reg Ex	Fine-tuned RoBERTa	GPT-4
Names Required	Yes	Yes	Yes	No
Precision	0.827	0.550	0.680	0.526
Recall	0.970	0.905	0.840	0.958

Comparing our results with previous literature (see Table 3), we observed that GPT-based de-identification offered a higher recall (0.958) than the results observed by Farrow et al. [3] when combining class lists and regular expressions (0.905). However, the average precision of GPT (0.526) was poorer than the precision observed by Farrow et al. [3]. Although the improvement in recall was relatively small compared to the drop in precision, recall is arguably the most important metric, as it directly reflects the frequency of instances where student privacy was not adequately protected. A similar outcome is observed when comparing our results with previous transformed-based de-identification processes [4], which showed a recall of 0.84 with a precision of 0.68 for redacting names. Compared to these previous results, GPT-4 demonstrates substantially better recall, but with a greater reduction in precision. However, beyond these promising results in terms of recall, the current best approach using supervised machine learning algorithms [1] still outperforms GPT-4 in this task, with a recall of 0.970 and a precision of 0.827.

4.3 Over-redaction of Names, Locations and Links

One of the main issues observed for the GPT-based de-identification process was that GPT was not always able to successfully differentiate the names of artists, scientists, or political leaders from the names of students. For instance, for the *Poetry* course, which had the lowest precision, the content often includes extensive essays discussing the works of various poets. In those cases, human coders did not consider the names of poets and artists as PII in literary discussions, recognizing the educational context. However, the GPT-4 model failed to make this distinction, leading to an example where names such as *"John Latouche"* and *"Jackson Pollock"* were inappropriately redacted, significantly impacting the precision score for this course:

Human redacted text:

"... in the poem 'A Step Away From Them,' the mention of 'Bunny,' 'John Latouche,' and 'Jackson Pollock' contextualizes the poet's friends' deaths..."

GPT-4 redacted text:

"... in the poem '[REDACTED],' the mention of '[REDACTED],' '[REDACTED],' and '[REDACTED]' contextualizes the poet's friends' deaths..."

In addition, GPT also treated almost all locations as PII. For example, in the *Business Trends* course, discussions involve analyzing country-specific economic trends. Names of countries and institutions are essential for these discussions, but the GPT-4 model redacted these as well:

Human redacted text:

"As the UK is not a part of the EZ, it was not directly affected by the Euro Crisis and did not contribute to the bailout of Greece..."

GPT-4 redacted text:

"As [REDACTED] is not a part of the [REDACTED], it was not directly affected by the Euro Crisis and did not contribute to the bailout of [REDACTED]..."

Finally, in the forum posts of the *Mythology* class, GPT incorrectly redacted the names of mythological creatures. For example, in the post *"Here you are, our friend Cyclops,"* humans did not redact any words. However, GPT redacted the word *Cyclops*, considering that it was the name or nickname of a student. Although there could be a case where *Cyclops* is a name or a nickname, knowing that this post appeared in the forum of a *Mythology* class, the student was probably referring to the mythological creature rather than another student. These examples highlight the challenges GPT faces in distinguishing between names and locations that are public information and those that actually involve PII.

5. DISCUSSION AND CONCLUSION

This research aimed to assess the effectiveness of GPT-4 in redacting personally identifying information (PII) from a diverse dataset of forum posts from nine academic courses. The primary objective was to understand the model's capabilities and limitations in handling sensitive data, helping us to evaluate whether GPT-4 can be part of the solution for protecting privacy and data security in digital environments.

We utilized OpenAI's GPT-4 model to process 3,505 forum posts from nine Massive Open Online Courses (MOOCs) at the University of Pennsylvania and compared its results with human-based redactions. Our results show that GPT-4 achieves high recall, consistently over 0.85 across all courses, indicating its efficiency in identifying PII. However, the precision was often lower than 0.7, indicating that GPT-4 incorrectly over-redacts names and locations that are not PII. This pattern of higher recall and lower precision in de-identification aligns with findings from previous studies [1, 3, 4] that employed a range of methods for de-identifying student data. While the gap between recall and precision is wider with GPT, its enhanced recall over methods that combine class lists with regular expressions [3] and transformer models [4] indicates that GPT could be a preferable choice for maximizing privacy, particularly when the data set does not have large numbers of mentions of non-PII names and locations. However, in research contexts where such information is essential, the reduced precision of GPT-4 might represent an important drawback.

The current best-performing approach, which uses supervised machine learning algorithms, still outperforms GPT-4 in both precision and recall [1]. However, it is not clear that both studies can be directly compared. For one thing, the higher performance in [1] might be partly because the authors focused exclusively on student names, potentially simplifying the de-identification process. The difference might also be due to differences in content between these studies. We note that performance indicators were substantially better in courses with a strong mathematical component, while they were lower in classes where public figures or historical locations were more frequently mentioned. This suggests that the context and content of the data being processed play a crucial role in the effectiveness of different de-identification approaches.

Most past models use additional information (such as a list of known student names, see [3, 4]); in this work, we tackled a harder problem, conducting redaction solely using the text itself. In addition, Bosch et al.'s [1] approach's success is also likely to stem from the thorough feature engineering process they used, producing features such as word appearance in the U.S. census and in the dictionary, while also considering all possible spelling mistakes in one or 2 characters when checking appearance in such lists. GPT-4 or other LLMs might benefit from these specific details being explicitly included in the prompt.

Our approach in using the GPT-4 model was general in nature, focused on developing a single approach that could work for many contexts. We did not customize or tailor the prompts to account for specific course content or the nature of the names appearing in the texts. This general approach was adopted to maintain a consistent methodology across all courses, that could be applied as-is to new courses. However, this approach also likely represents a lower bar for how well LLM-based de-identification can eventually perform. Future studies could explore more nuanced prompt engineering and course-specific model training (or fine-tuning) to enhance the precision of PII redaction without compromising recall. Additionally, it may be worth investigating a hybrid approach combining human and AI redactions, or a hybrid approach combining LLMs and other AI methods. It is also important for future work to consider whether algorithmic bias [11] impacts the performance of this approach, for instance if GPT performs more poorly for PII from less well-represented groups of learners.

As noted by Zambrano et al. [13] in the context of qualitative coding, one of the main advantages of using GPT is not only the automation itself but also the possibility of having an additional layer of verification to mitigate potential mistakes made by humans. Even though the GPT redactions are not perfect, the analysis of their disagreements with human redactions can help us identify our own mistakes in the process. In this case, GPT-4 also identified many examples of PII that human coders overlooked. This suggests that large language models (LLMs), even with its over-redaction, are not only useful for automating data de-identification but also for enhancing the accuracy of human-performed de-identification, which can be prone to errors and omissions.

Ensuring the security of the data provided to any LLM for de-identification remains a crucial concern. In the case of the GPT-4 API, OpenAI states that the data may be retained for a maximum of 30 days and will only be accessed or reviewed if necessary to monitor for abuse. We chose OpenAI's model due to its proven high performance in similar tasks [7, 10] and their public commitment to not misuse data. Nevertheless, future work in this area may choose to use alternative open source large language models that can be run fully locally, such as LLaMA. Such developments would help mitigate potential security concerns associated with the transmission of data to external providers like OpenAI.

In conclusion, our study demonstrates the significant potential of GPT-4 in processing and redacting sensitive information from large datasets. While the model shows high recall rates, its tendency for over-redaction highlights a key area for improvement. This research contributes to the evolving narrative on AI's role in ensuring data privacy, particularly when open science can benefit from data sharing but where doing so involves some risk of disclosure of PII. Much of the recent discourse on AI is about its risks to privacy and learners; with this application, AI may be able to reduce those risks.

6. ACKNOWLEDGMENTS

We thank the New Venture Fund for their support through grant "RPPL Research Infrastructure for Professional Development: Year One".

7. REFERENCES

[1] Bosch, N., Crues, R., Shaik, N. and Paquette, L. 2020. " Hello,[REDACTED]": Protecting Student Privacy in Analyses of Online Discussion Forums. *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)* (2020), 29–39.

[2] European Parliament and the Council of the European Union 2016. 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.

[3] Farrow, E., Moore, J.D. and Gasevic, D. 2023. Names, Nicknames, and Spelling Errors: Protecting Participant Identity in Learning Analytics of Online Discussions. *LAK23: 13th International Learning Analytics and Knowledge Conference* (2023), 145–155.

[4] Holmes, L., Crossley, S.A., Morris, W., Sikka, H. and Trumbore, A. 2023. Deidentifying Student Writing with Rules and Transformers. *International Conference on Artificial Intelligence in Education* (2023), 708–713.

[5] Kayaalp, M., Browne, A.C., Callaghan, F.M., Dodd, Z.A., Divita, G., Ozturk, S. and McDonald, C.J. 2014. The pattern

of name tokens in narrative clinical text and a comparison of five systems for redacting them. *Journal of the American Medical Informatics Association*. 21, 3 (2014), 423–431.

[6] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*. (2019).

[7] Liu, Z., Yu, X., Zhang, L., Wu, Z., Cao, C., Dai, H., Zhao, L., Liu, W., Shen, D., Li, Q., and others 2023. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*. (2023).

[8] Meystre, S.M., Friedlin, F.J., South, B.R., Shen, S. and Samore, M.H. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Medical Research Methodology*. 10, 1 (Aug. 2010), 70. DOI:https://doi.org/10.1186/1471-2288-10-70.

[9] Murugadoss, K., Rajasekharan, A., Malin, B., Agarwal, V., Bade, S., Anderson, J.R., Ross, J.L., Faubion, W.A., Halamka, J.D., Soundararajan, V., and others 2021. Building a best-in-class automated de-identification tool for electronic health records through ensemble learning. *Patterns*. 2, 6 (2021).

[10] Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M. and Yang, D. 2023. Is ChatGPT a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*. (2023).

[11] Ray, P.P. 2023. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*. 3 (2023), 121-154.

[12] Rudniy, A. 2018. De-identification of laboratory reports in STEM. *Journal of Writing Analytics*. 2, (2018), 176–202.

[13] Zambrano, A.F., Liu, X., Barany, A., Baker, R.S., Kim, J. and Nasiar, N. 2023. From ncoder to chatgpt: From automated coding to refining human coding. *International Conference on Quantitative Ethnography* (2023), 470–485.