Information and Learning Sciences

401

De-identifying student personally identifying information in discussion forum posts with large language models

Andres Felipe Zambrano
Graduate School of Education, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Shreya Singhal
Graduate School of Education, Harvard University, Cambridge, Massachusetts, USA, and

Maciej Pankiewicz, Ryan Shaun Baker, Chelsea Porter and Xiner Liu
Graduate School of Education, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Received 30 November 2024
Revised 25 January 2025
Accepted 7 March 2025

Abstract

Purpose – This study aims to evaluate the effectiveness of three large language models (LLMs), GPT-4o, Llama 3.3 70B and Llama 3.1 8B, in redacting personally identifying information (PII) from forum data in massive open online courses (MOOCs).

Design/methodology/approach – Forum posts from students enrolled in nine MOOCs were redacted by three human reviewers. The GPT and Llama models were then tasked with de-identifying the same data set using standardized prompts. Discrepancies between LLM and human redactions were analyzed to identify patterns in LLM errors.

Findings – All models achieved an average recall of over 0.9 in identifying PII and identified PII instances overlooked by humans. However, their precisions were lower – 0.579 for GPT-4o, 0.506 for Llama 3.3 and 0.262 for Llama 3.1 – showing a tendency to over-redact non-PII names and locations.

Research limitations/implications – Several courses' data were analyzed to increase findings' generalizability but the models' performance may vary in other contexts. GPT and Llama models were selected because of their availability and cost-effectiveness at the time of the study; future newer models may improve performance.

Practical implications – The use of downloadable LLMs enables researchers to de-identify data without training specialized models or involving external companies, ensuring that student data remains private.

This paper was edited with the assistance of ChatGPT.

Funding: This study was supported by New Venture Fund; RPPL Research Infrastructure for Professional Development.

Information and Learning Sciences
Vol. 126 No. 5/6, 2025
pp. 401-424
Emerald Publishing Limited
2398-5348
DOI 10.1108/ILS-11-2024-0156

Originality/value – Previous research on LLM text de-identification has largely used proprietary models, which require sharing data containing sensitive PII with third-party companies. This study evaluates the performance of two open weight models that can be deployed locally, eliminating the need to share sensitive data externally.

Keywords De-identification, Anonymization, GPT, Llama, Large language models, Privacy, Massively open online course

Paper type Research paper

1. Introduction

In the digital era, the proliferation of new educational technologies and platforms has significantly increased the availability of data resources for researchers. This includes an extensive expansion in the collection of textual data from discussion forums, chat sessions, classroom transcripts and human-tutor dialogues, among other sources. Although this data offers substantial potential for investigating student behavior, pedagogical effectiveness and communication patterns, it often contains personally identifying information (PII), presenting serious 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, 2016) in the European Union and the Family Educational Rights and Privacy Act (FERPA, 1974) in the USA. Beyond these regulations, teachers and students often also have concerns about the public scrutiny of their ideas, opinions, behaviors, or outcomes (Jones *et al.*, 2020), making it more difficult to receive their consent if we cannot guarantee that their data will be protected when sharing it with other stakeholders. Although sharing or publicly releasing data sets is particularly important and beneficial for open science – enabling collaboration across diverse stakeholders and institutions, facilitating the review and replication of prior analyses and fostering the exploration of new research questions – managing this process requires careful attention to protect the privacy and confidentiality of all participants involved. Therefore, it is essential to implement robust de-identification processes before making data sets available to the research community.

Manual redaction of data is both costly and time-consuming (Megyesi *et al.*, 2018), making it impractical for large data sets. As an alternative, automated redaction has been proposed (Kovačević *et al.*, 2024); however, the diverse nature of PII presents significant challenges (Garfinkel, 2015). Although certain types of PII, such as mail and email addresses, phone numbers, or personal webpage links, can be efficiently identified using methods such as regular expressions or supervised machine learning natural language processing (NLP), other forms of PII are more difficult to recognize. For example, nicknames or terms that can serve as both personal names and common dictionary words are particularly challenging to redact (Kayaalp *et al.*, 2014). Furthermore, not all mentions of names or locations are PII – for instance, references to authors, political leaders, or historical events. Additional difficulties also arise when students make typographical errors; use incorrect grammar, punctuation or spacing; or use words from languages different from the one used in the training of the supervised learning models.

Despite these challenges, supervised learning techniques have shown promising results in de-identifying data, achieving recall rates above 0.95 in redacting student names (for more details, see the related work section; Bosch *et al.*, 2020). However, these models depend heavily on a ground truth data set for training and may not perform well in scenarios where this data set is not representative. Creating such a data set, as well as the time required for training and tool development, involves a significant time investment. Additionally, although manual de-identification by human coders is often considered the gold standard, it is not

without errors, as coders can overlook or misclassify instances of PII, potentially diminishing the effectiveness of supervised learning models in practical applications. Executing a second review by a different human coder can mitigate these errors but at the cost of significantly more time.

An alternative and potentially more accessible approach for de-identification involves using large language models (LLMs) such as GPT or Llama models. Previous research indicates that GPT-4 can effectively identify personal names within data sets (Liu *et al.*, 2023; Qin *et al.*, 2023). Given the importance of data security when using LLMs, this paper examines the performance of GPT-4o and the open weights Llama 3.3 70B and Llama 3.1 8B models for this task. Llama models can be downloaded, offering the advantage of local operation, which reduces security risks associated with transferring and storing data on third-party servers. Although Llama 3.3 70B requires specific infrastructure that may pose challenges for local deployment, it remains feasible under the right conditions. Additionally, we also included the Llama 3.1 8B model, which can be used locally on personal laptops. This possibility of running the entire de-identification pipeline locally could also increase the willingness of participants to share their data by guaranteeing them that their identified data will not be shared with anybody else.

In our study, we analyze and compare the abilities of three LLMs to redact discussion forum posts from students across nine massive open online courses (MOOCs). MOOCs discussion forums are dynamic spaces where written information is generated rapidly. Because of their dialogic nature, these forums frequently include PII, as students often reference previous posts (using the names of other students, facilitators, or instructors), share links to personal pages or social media groups to connect outside the class or post links to video meetings that may also contain PII. Moreover, the informal and conversational tone of these posts – compared to essays or other assignments – often includes spelling and syntax errors, making it more challenging for traditional supervised algorithms to identify PII accurately (Kovačević *et al.*, 2024).

The discussion topics, which vary by course, can further complicate de-identification. Posts may include mentions of article authors without formal citations or may make casual references to public figures, adding to the complexity of distinguishing between sensitive and non-sensitive information. Given the diverse contexts and challenges presented by these forums, we evaluate the precision, recall and Cohen's kappa of LLMs in identifying PII across the nine MOOCs and compare their performance with current benchmark models reported in previous studies. Additionally, we investigate whether LLMs can identify PII that human coders – considered the gold standard for this task – might miss. This exploration aims to determine whether LLMs, beyond their performance metrics, could support the creation of a new gold standard for de-identification through a hybrid redaction process.

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, each obtaining F1-scores over 98% in the medical domain (Kovačević *et al.*, 2024). Although the use of human annotators is still the main de-identification approach in educational research (Crossley *et al.*, 2022, 2023; Megyesi *et al.*, 2018), the success of automated redaction has also motivated the use of machine learning models to anonymize data sets in the educational field (Holmes *et al.*, 2023b).

One of the pattern mining approaches that has been used in several studies is the use of name lists as a reference for detecting those words that match with items in the original list

and redacting them. For example, Rudniy (2018) has used this approach to redact laboratory reports in a STEM educational setting, achieving a precision of 0.79 and a recall of 0.75. However, this approach would be difficult to scale, as it requires having a list of all potential student names. In the medical domain, Kayaalp *et al.* (2014) addressed this limitation by using a database of 3.8 million names collected from Social Security to redact personal names from medical reports. However, they found that many names coincide with common dictionary terms, diminishing the performance of the redaction. Moreover, for educational settings with students (in some cases children) from different countries and ethnic backgrounds, where names may have uncommon spellings and nicknames may be used, it is unlikely that even these databases from government institutions will have all possible names.

Regular expressions are another common pattern-matching method used for de-identification. Farrow *et al.* (2023) assessed the effectiveness of regular expressions compared to using a simple list of students' first and last names. Their findings showed an increase in recall from 0.515 to 0.876 but a decrease in precision from 0.879 to 0.567, using data from six sessions of a master's distance-learning course. Although the increase in recall came at the expense of precision, the authors argued that the trade-off was justified because the higher recall is crucial given the severe implications of not redacting all PII. They also explored a hybrid approach that combined class lists with regular expressions, which marginally improved recall to 0.905 but further reduced precision to 0.550.

Supervised machine learning techniques have also been used for redacting PII. Bosch *et al.* (2020) used the extra-tree variant of random forest and deep neural networks to redact text data from discussion forums in two online courses offered by a public university in the USA. Their approach used features that included the position of each word within a sentence, its occurrence on US census lists and its presence on lists of cities, political regions or countries worldwide. They also accounted for potential misspellings by considering words that were within one or two edits of standard dictionary terms. The results indicated better performance compared to traditional pattern-matching techniques, achieving an average recall of 0.970 and a precision of 0.827 across the machine learning models. Moreover, they reported a Cohen's Kappa score of 0.794, approaching the original score of 0.864, which measured agreement between the two human coders who defined the ground truth in their study.

2.2. Large language model-based de-identification

Transformer models have demonstrated recall rates above 0.99 in de-identifying medical data sets (Chambon *et al.*, 2022). Motivated by this success in the medical field, Holmes *et al.* (2023a) applied two RoBERTa-based transformer models, which were fine-tuned versions of a pre-trained LLM (Liu *et al.*, 2019), to de-identify data from student essays in a MOOC, specifically targeting names. This method achieved a recall of 0.84 and a precision of 0.68. They compared these models against a rule-based system that labels student names using a general-purpose Named Entity Recognition model, which showed lesser performance with a recall of 0.81 and a precision of 0.33. Holmes *et al.* (2023a) showed that most of the false negatives of their classification (leaked names) correspond to first names, mainly those used in second or third person. In particular, they found some instances where the leaked name corresponded to PII with spaces between characters (e.g. S a m u e l).

Holmes *et al.* (2023b) extended their previous study to also anonymize discussion forum data from a computer science course at a large university in the USA, considering other PII information beyond names. The authors used a regular expression approach for anonymizing all the PIIs different from names (e.g. URLs, email addresses, etc.). In this second data set, the precision for labeling names was substantially higher than in their first study (0.74), but

their recall dropped to 0.70. The different performance they obtained in this study compared to their previous study using the same system indicates that performance is highly dependent on the context of the data, and results across different contexts might not be completely comparable (e.g. it is not the same to redact medical reports, essays, or forum posts). Moreover, the recall for labeling non-name PII, which mainly corresponded to personal URLs (95% of the PII), was higher (0.89), but the precision was substantially lower (0.27).

Recently, the GPT family of LLMs has been explored for related tasks. Qin *et al.* (2023) used GPT-3.5 to identify names in news articles, achieving an F1-score of 0.532 for general named entity recognition and 0.872 for personal names; however, specific precision and recall metrics were not reported. In a separate study, Liu *et al.* (2019) used GPT-4 for de-identifying medical reports. They found that GPT-4, even when applied zero-shot, achieved an accuracy of 0.99, outperforming a fine-tuned RoBERTa model, which obtained an accuracy of 0.947 on the same data set. Although precision and recall were not detailed for these applications, the improvement in accuracy highlights GPT-4's potential as an effective de-identification tool. Considering the types of errors identified by Holmes *et al.* (2023a; 2023b), we hypothesize that general-purpose LLMs may surpass both fine-tuned and traditional supervised learning models in performance.

3. Methods

3.1 Data

The data set for this study includes forum posts from students who were enrolled in nine different MOOCs at the University of Pennsylvania between 2012 and 2015. These courses spanned a broad range of subjects, including accounting, calculus, design, gamification, business trends, poetry, mythology, probability and vaccines. We aimed to reduce bias toward any specific academic domain and enhance the generalizability of our findings across various disciplines by selecting this diverse array of course topics. The use of this data set complied with the terms of service under which students agreed to when first accessing the learning platform and was reviewed by the Institutional Review Board of the University of Pennsylvania.

Our initial data set was created by randomly selecting 500 forum posts from each of the nine courses (using a standard number of posts to facilitate comparisons across courses), having a total of 4,500 posts. We excluded posts written in languages other than English or those consisting solely of special symbols, characters, website links or mathematical formulas to maintain relevance and uniformity. The refined final data set comprised 3,505 posts from 2,882 unique students. The posts were roughly evenly distributed across all nine courses, with each course contributing between 379 and 399 posts. We preserved the original text of the posts, including any typographical or grammatical errors, as these elements are crucial to the natural language processing challenges we intended to address.

3.2 Human de-identification process: first iteration

Three human reviewers were in charge of manually redacting any PII from the posts. This redaction process required the removal of names, contact details, geographical origins or residences, links to personal websites and any other information that could disclose the identity of the authors. Two of the reviewers, both Master's students, were trained in redaction techniques by a faculty member. These two reviewers shared the workload evenly, with each responsible for redacting approximately half of the posts from each course, which amounted to about 195 posts per course for each reviewer. The third reviewer, a professional consultant with approximately 20 years of experience in document editing and transcription, including redaction, performed an additional review to catch and correct any errors that the initial reviewers might have missed.

Rather than calculating inter-rater reliability (IRR) between the two primary reviewers – a common approach used in similar studies (e.g. Bosch *et al.*, 2020) – we opted for this cascade approach because we believe human redaction errors are more likely to result from oversight or carelessness than from disagreements about what constitutes PII, once there are clear guidelines about what should be considered PII. However, even with high inter-rater agreement, individual reviewers can still make careless mistakes, particularly when coding large amounts of data. Consequently, we considered this sequential approach, where trained Master's students performed the initial redactions and an expert conducted a thorough review, to be a more reliable method for establishing our ground truth.

As done in previous work (e.g. Bosch *et al.*, 2020), we provided our redactors with guidelines on the types of names to retain (such as those of famous people and authors) and those to redact (including names of instructors and students). However, we asked them to apply their judgment in making these distinctions. Additional PIIs we requested them to redact included places of origin (even if they referred to a country where multiple students could have been born), links to personal webpages or social media profiles, references to previous employment at specific companies, as well as cellphone numbers, email addresses, current geographical addresses or any other details that might pose a risk of re-identifying students. Unlike many studies in other fields (see review by Meystre *et al.*, 2010), where human redactors are provided with a list of known student names for redaction, we chose not to do this to avoid biasing the results in favor of human coders, as such a list was not available to the LLMs and may not be available in many real-world cases. Human coders redacted a total of 2,134 words, with 1,282 posts containing at least one redaction, accounting for 36.6% of the total posts. This redaction process established the first ground truth data set, which was used to evaluate the effectiveness of LLMs in performing similar redactions.

3.3 De-identification process with GPT

For the de-identification process, we used the OpenAI GPT-4o model (version 2024-11-20), accessed via its API. At the time of our study and the drafting of this paper, we relied upon OpenAI's privacy policy for our account, which guaranteed that data would only be used for the user's intended purpose and for verifying abuse reports. This policy also guaranteed that the data would only be retained 30 days before deletion. Additionally, we used a data set collected a decade prior to the current study to further minimize re-identification risks when uploading data to the API.

We set the temperature parameter to 0 to reduce inconsistencies caused by the stochastic nature of the model's output. The max_tokens parameter, which determines the maximum response length, was set to 1,000 – sufficient to process each post in our data set, as the average post length for all courses was less than 300 words. All other parameters were maintained at their default settings. To optimize the model's performance, we tested multiple prompts using a subset of the data (approximately 10%). After identifying the best-performing prompt, we refined it based on the errors observed in this subset until obtaining the version presented here. Each forum post was sent individually to the model using the following prompt: (Any PII presented in this paper was altered to protect student privacy. All examples use fictitious links and pseudonyms).

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. Names of artists, philosophers, politicians, or locations that do not correspond to the

> student's place of origin or current location should not be removed. 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:
>
> @Kate, You'll find below a Linkedin group I just create. 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! Marcos.
>
> The output must be:
>
> [REDACTED],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! [REDACTED].
>
> Please repeat this process with the following post:
>
> [POST TO BE DE-IDENTIFIED]

Our prompt explicitly stated the types of PII to be redacted, including names, company names, places of origin, current residences, addresses and social media links. These categories were chosen because of their frequent occurrence in textual educational data sets and their relatively high risk for exposing student identities. We also instructed GPT-4o not to redact the names of public figures, such as artists or politicians, or locations that do not qualify as PII. This was necessary because GPT-4o exhibited low precision in distinguishing between public figures' names or general locations and students' names, places of origin, or current residences. Additionally, we asked GPT-4o to preserve the original text's structure and formatting exactly as it was, word by word. This instruction was essential because, without it, GPT-4o would automatically correct grammar and punctuation errors in the posts and modify words to improve the clarity of the original message.

3.4 De-identification process with Llama

We also studied de-identification using two open weights LLM models, using Meta's Llama-v3p3-70b-instruct model (Llama 3.3), accessed through its API available on the Fireworks.ai platform and Llama-v3p1-8b (Llama 3.1), downloaded using Ollama and run locally on a personal computer (16 GB of RAM, NVIDIA RTX 3050 GPU with 12GB of memory and a 16-core CPU). Because Llama models are open weight, Llama 3.3 can also be downloaded and maintained locally, though it would require an infrastructure with a recommended GPU and RAM of 48GB, and a CPU of 48 cores, requirements that exceed the capacity of a personal computer. For consistency, we set the temperature parameter to 0 and the max_tokens parameter to 1000 as done with GPT-4o. All other parameters were set to default.

Recent research has shown that Llama models are typically more sensitive to prompt differences than recent GPT models (Mizrahi *et al.*, 2024). Therefore, we also explored

multiple completely different prompts for Llama 3.3, including the prompts used for GPT-4o, asking it to redact 10% of the data. After this initial exploration, the prompt shown above for GPT outperformed all the other prompts, although it obtained worse results than those achieved for GPT-4o in terms of precision. Llama 3.3 also tended to add an additional line introducing the redacted posts or the specific changes (e.g. "Here is the PII redacted post:"). Analyzing the specific mistakes observed for Llama 3.3, this prompt was refined multiple times. However, in all these cases, an improvement in precision was at the cost of reducing recall, which is arguably more important than precision to avoid disclosing PII that might pose a risk for students. For this reason, and to enhance comparability with the GPT-4o results, we decided to use the same prompt used for GPT-4o with a minor adjustment to address those undesired additional lines introducing the redacted post. We also used the same prompt for the Llama 3.1 model.

Each forum post from the original set was sent individually to the model using the following prompt.

Prompt:

> Please edit the provided text by removing any personally identifiable information (PII). This includes names of students, instructors, or professors, company names, places of origin, current living locations, addresses and social media links. Names of artists, philosophers, politicians, or locations that do not correspond to the student's place of origin or current location should not be removed. 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. DO NOT RETURN ANY EXTRA LINES EXCEPT THE POST ITSELF.
>
> […] [The remainder of the prompt is identical to the GPT-4o Prompt (omitted for brevity)]

Despite this instruction to avoid any additional line introducing the redacted message, both Llama models (mainly Llama 3.1) occasionally included an additional introductory line in their outputs. To address this issue, we programmatically removed these lines by identifying the initial word of the actual post.

3.5 Second iteration of human de-identification and large language model evaluation

To evaluate the performance of the LLM-based de-identification process, we conducted a word-by-word comparison between the outputs from the LLM models and the human-redacted posts. In this process, we identified four types of discrepancies in cases where the humans and LLMs agreed on the redaction but did it differently. For instance, a LinkedIn URL was redacted by a human as "Connect with me at: [REDACTED]," whereas the LLM included "LinkedIn" in the redaction, returning "Connect with me at: LinkedIn: [REDACTED]." This pattern of including the social network name was consistently observed for the LLMs. The second discrepancy involved titles. For example, one case redacted was by humans as "Thanks Mr [REDACTED]. It is a very interesting class […]," LLMs simplified it to "Thanks [REDACTED]. It is a very interesting class […]," omitting the title. Both humans and LLMs showed inconsistency in including titles such as "Mr," "Mrs" and "Prof," among others.

The last two types of discrepancy were specific to Llama models. Both Llama 3.3 and Llama 3.1 handled some links inconsistently. Sometimes, they redacted the entire link as seen with GPT, but other times, it only redacted specific parts it identified as PII. For example, for the previous LinkedIn case where GPT added the name of the social network, Llama 3.3 redacted it as "Connect with me at: www.linkedin.com/ [REDACTED]," which is

not disclosing any PII. For another link, Llama 3.1 returned "http//www.[REDACTED].ac.uk/[REDACTED]." In this example, Llama 3.1 removed the name of the institution and the specific personal page within the institutional website but left other parts of the domain visible. These instances required manual evaluation based on the information still visible in the link. In the given case, the link indicated that the individual was associated with a UK academic institution, leading us to classify it as a false negative because of the potential disclosure of sensitive information.

Finally, as mentioned before, Llama models (mainly Llama 3.1) often added an extra line along with the desired output even after explicitly requesting to avoid it (although these cases were significantly reduced after adding this instruction). For example:

> "Here is the PII redacted post:
>
> Thankyou [REDACTED]."

To ensure comparability between the LLMs and human-de-identified posts, we programmatically removed this extra line. Additionally, we manually adjusted the aforementioned discrepancies before assessing performance. To standardize the texts, we also stripped all articles, non-alphanumeric words and punctuation from the texts, replacing them with spaces. This allowed us to perform a word-by-word comparison (with words defined as whitespace-delimited strings) and address cases where the LLMs had corrected non-alphanumeric symbols or made unrequested changes to grammar, spelling, or punctuation, despite the instructions in the prompt to avoid such corrections. After this tokenization, we still observed instances where students wrote two words without a space in between (e.g. "Thankyou"). In most instances, the LLMs preserved the original text, resulting in no discrepancies between the human-based and LLM-based de-identification processes that required further review. However, in those cases where discrepancies were noted, the term was treated as a single word (as written by the student), and the LLM's output was adjusted accordingly. These adjustments were made with the understanding that such terms did not constitute PII disclosures.

After correcting for these low-level differences, all the remaining discrepancies corresponded to disagreements between human and LLM-based de-identification. By analyzing these disagreements, we identified 47 instances where human redactors overlooked a case of PII. We corrected all of them, creating an updated gold standard (human-based ground truth with corrections from GPT and Llama) to evaluate the performance of all analyzed LLMs.

3.6 Evaluation

After correcting the human-based de-identification, we assessed the performance of the LLM models 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 the LLM.
- True negative (TN): Words that were not identified as PII by either human coders or the LLM.
- False positive (FP): Words that were identified as PII by the LLM but not by human coders.
- False negative (FN): Words that were identified as PII by human coders but not by the LLM.

To calculate the metrics, we first identified the TP, TN, FP and FN at the wordlevel following the above-mentioned definition. We then calculated precision, recall and Cohen's Kappa for each course. To address potential inconsistencies in the LLM-based de-identification process, we processed the data through GPT-4 and Llama 3 three times each, calculating performance metrics for each iteration before averaging them. Because of ethical considerations and constraints from the data owner, the data set cannot be shared publicly. However, the code for our LLM-based de-identification process is available for replication purposes at https://osf.io/79m6w/?view_only=fb10005c5c174a5e933c4fc526c20c59.

4. Results

Table 1 shows the distribution of redacted elements for each course and the frequency of corrected human redactions per post (both post-correction). The Design course had the highest percentage of posts with redactions (51.1%) and was one of the courses with the shortest posts (44.4 words per post). In contrast, Poetry had a distinctively higher average of 241 words per post and the fewest redactions per post (0.31). This suggests that the posts in Poetry likely involved more in-depth discussions or included excerpts from existing poems or original creations without necessarily adding any PII. This also suggests that the number of cases of PII (and the difficulty of redacting them) could depend on the domain of the course.

4.1 Human mistakes

In inspecting our results, we discovered that after the first round of human-based de-identification, human coders missed 47 words that involved PII distributed across all 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, Emily. That is a major point.

In this case, the human coder failed to remove "Emily" (again, not the original name in the data), clearly the name of the author in a previous post to which the current post is replying.

Table 1. Distribution of posts, words by post and human-redacted elements across all the courses after corrections

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 (2.4)
Calculus	396	114 (28.8)	40.0	162 (1.0)	3 (1.9)
Design	380	194 (51.1)	44.4	284 (1.7)	9 (3.2)
Gamification	379	124 (32.7)	63.5	237 (1.0)	7 (3.0)
Business trends	387	143 (37.0)	81.1	292 (0.9)	16 (5.5)
Poetry	399	85 (21.3)	241.0	124 (0.1)	2 (1.6)
Mythology	390	138 (35.4)	67.9	196 (0.7)	1 (0.5)
Probability	396	117 (29.5)	56.9	177 (0.8)	0 (0)
Vaccines	391	159 (40.7)	71.2	287 (1.0)	3 (1.0)

Note(s): 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
Source(s): Authors' own creation

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

Course	Precision			Recall			Kappa		
	GPT-4o	Llama 3.3	Llama 3.1	GPT-4o	Llama 3.3	Llama 3.1	GPT 4o	Llama 3.3	Llama 3.1
Accounting	*0.764*	0.616	0.327	0.967	*0.971*	0.950	*0.852*	0.750	0.478
Calculus	*0.668*	0.620	0.224	0.936	*0.974*	0.956	*0.778*	0.756	0.356
Design	*0.779*	0.760	0.385	*0.975*	0.960	0.936	*0.864*	0.846	0.537
Gamification	*0.685*	0.588	0.342	*0.982*	0.948	0.906	*0.806*	0.723	0.491
Business trends	*0.473*	0.305	0.272	0.961	*0.972*	0.935	*0.630*	0.458	0.414
Poetry	*0.273*	0.195	0.110	*0.907*	0.905	0.855	*0.419*	0.320	0.193
Mythology	0.486	*0.504*	0.204	0.930	*0.978*	0.949	0.636	*0.663*	0.329
Probability	*0.637*	0.600	0.198	0.929	*0.976*	0.949	*0.754*	0.741	0.321
Vaccines	*0.445*	0.370	0.292	0.928	*0.974*	0.913	*0.597*	0.531	0.436
Average	*0.579*	0.506	0.262	0.946	*0.962*	0.928	*0.704*	0.643	0.395

Note(s): The best performing LLM for each course according to each metric is shown in italics
Source(s): Authors' own creation

By contrast, both Llama models and GPT-4o 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;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. GPT-4o and Llama 3.3 caught 46 mistakes each, while Llama 3.1 caught 32 mistakes.

Although the examples of human errors constituted only a small portion of the overall words that involved PIIs, they highlight the fallibility of humans in data handling processes, suggesting that even a fairly thorough process such as the one used here might not fully de-identify data reliably. This potential for human error is corroborated by Bosch *et al.* (2020), who observed 37 discrepancies between two human coders in identifying 600 possible names within their data set (6.1%). In our study, all observed human errors involved cases where one of the two human reviewers missed a student's name, personal URL or place of origin – types of PII that they successfully redacted in other posts. These errors appeared to be oversights rather than disagreements between raters or a systematic misunderstanding of what constitutes PII. Our findings underscore the value of analyzing disagreements between human-based and automated de-identification processes, as this can help reduce errors and enhance the quality of de-identification.

4.2 Performance

Table 2 presents the average results from three iterations of the de-identification process for each metric and each course, using corrected human redaction as the ground truth. The recall rate consistently exceeded 0.85 across all courses, indicating that all models, including the local version of Llama, reliably identified most of the personally identifiable information (average recall of 0.946, 0.962 and 0.928 for GPT-4o, Llama 3.3 and Llama 3.1, respectively). However, the precision was below 0.8 in all cases, often because of misidentifying non-PII elements such as names of famous people, locations or general links. This drop in precision was substantial for Llama-3.1 (average precision = 0.262), compared to the other two models (average precision of 0.579 and 0.506 for GPT-4o and Llama-3.3, respectively). Although not ideal, this over-redaction by both GPT and Llama did not reduce the degree to which sensitive information remained protected, which may be considered an acceptable trade-off for enhanced privacy (Holmes *et al.*, 2023a, 2023b).

Cohen's kappa, which assesses the agreement between the LLM models and human coders considering the distribution of both classes (PII and no PII words), varied significantly across courses. The highest kappa values for GPT-4o were observed for Design (kappa = 0.864), Accounting (kappa = 0.852), Gamification (kappa = 0.806), Calculus (kappa = 0.778) and Probability (kappa = 0.754). Similarly, Llama 3.3 demonstrated high kappa values for the same 5 courses, Design (kappa = 0.846), Calculus (kappa = 0.756), Accounting (kappa = 0.750), Probability (kappa = 0.741) and Gamification (kappa = 0.723), although it still slightly underperformed GPT-4o. In contrast, for courses such as Poetry (GPT-4o kappa = 0.419, Llama 3.3 kappa = 0.320), Business Trends (GPT-4o kappa = 0.630, Llama 3.3 kappa = 0.458),

Mythology (GPT-4o kappa = 0.636, Llama 3.3 kappa = 0.663) and Vaccines (GPT-4o kappa = 0.597, Llama 3.3 kappa = 0.531), both models demonstrated considerably lower agreement with human-based redactions. Llama-3.1 exhibited consistently low kappa values across all courses (average kappa of 0.395), primarily because of its low precision.

In general, agreement between models and human redactions tends to be lower in courses with longer average post lengths, all exceeding 65 words (in particular Poetry, which has a much higher average of 241 words per post). Additionally, in contrast to courses where GPT-4o and Llama 3.3 perform better (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 appeared to lead to confusion for both models (GPT-4o and Llama 3.3), though GPT-4o was less affected by this issue. As a result, both models achieved high recall but struggled with precision in the de-identification process. Previous list-based approaches, where an algorithm uses a predefined list of names for redaction, address this issue but at the cost of reduced generalizability.

In some instances, the models correctly identified PII but failed to redact the entire context, leaving adjacent words that could reveal the intended redaction. For example, "[David Anderson Green]" was partially redacted to "[REDACTED] Green," and "Professor [John Streak]" was partially redacted to "[REDACTED] Streak," in both cases allowing sensitive information to persist. Such errors highlight the models' limitations in handling compound identifiers and/or cases where an identifier can also be an everyday word. As mentioned before, GPT-4o demonstrated a slightly higher precision than Llama 3.3 in identifying everyday words, names, or locations that do not constitute PII. However, this increased precision made GPT-4o more prone to missing actual PII that resembles common words, contributing to its slightly lower recall compared to Llama 3.3.

Furthermore, while rare, the three LLMs occasionally failed to identify certain types of sensitive information, such as misspelt names, PII embedded in complex structures, or URLs that could lead to external sources where personal information might be disclosed. For example, all models failed to identify the following YouTube video URL. Although the URL does not directly reveal the identity of the student, personal information could be extracted by following the link to their personal YouTube channel:

> [REDACTED], I found something in my house that might help with your design, made a video to show it to you.http://youtu.be/11122233AAABBBdont *know if you already know the product. anyway, hope it helps.*

Comparing our results with previous literature (see Table 3), we observed that all LLMs achieved higher recall (GPT-4o: 0.946, Llama 3.3: 0.962 and Llama 3.1: 0.928) than the results observed by Farrow *et al.* (2023), who used class lists combined with regular expressions. However, only GPT-4o achieved higher precision than Farrow *et al.*'s approach (2023), while the two Llama models underperformed in this regard. Despite only a modest improvement in recall compared to the drop in precision (mainly for Llama 3.1), recall is arguably the most important metric, as it directly indicates the frequency of instances where student privacy was not adequately protected. When compared to earlier transformer-based de-identification methods (Holmes *et al.*, 2023a, 2023b), which achieved recalls of 0.84 and 0.70 and a precision of 0.68 and 0.74, all the LLMs achieved significantly higher recall but lower precision. However, Holmes *et al.* (2023b) also observed that precision is substantially lower when PIIs other than names are considered (0.27). Nevertheless, beyond these promising results (mainly in terms of recall), the current best approach using supervised machine learning algorithms and additional information (Bosch *et al.*, 2020) still

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

Paper	Bosch *et al.* (2020)	Farrow *et al.* (2023)	Holmes *et al.* (2023a)	Holmes *et al.* (2023b)		This paper		
PII	Names	Names	Names	Names	Personal URL, emails	Names, locations and links		
Method	Extra-trees + deep neural nets	Regular Expressions	Fine-tuned RoBERTa	Fine-tuned RoBERTa	Regular Expressions	GPT-4o	Llama-3.3	Llama-3.1
Names required	Yes	Yes	Yes	Yes	No	No	No	No
Precision	0.827	0.550	0.680	0.740	0.270	0.579	0.506	0.262
Recall	0.970	0.905	0.840	0.700	0.870	0.946	0.962	0.928

Source(s): Authors' own creation

outperforms all three compared models in this task, with a recall of 0.970 and a precision of 0.827.

4.3 Over-redaction of names, locations and links

As previously noted, a significant challenge encountered in all three LLM-based de-identification processes was that they were not always able to distinguish between the names of students and those of well-known figures such as artists, scientists or political leaders. This issue was particularly evident in the Poetry course, which recorded the lowest precision. The course material often includes detailed essays that discuss the works of various poets, whose names should not be considered PII. Human coders understood this distinction because of the educational context, but all compared LLM models did not, mistakenly redacting names such as "Jorge Luis Borges" as in the following example, significantly impacting the precision score for this course.

Human redacted text:

> […] in the poem 'Chess,' Borges mentions 'the King' to refer not only to God but maybe the Argentinian dictator and whoever controls him […]

LLM redacted text:

> […] in the poem '[REDACTED],' [REDACTED] mentions '[REDACTED],' to refer not only to God but maybe the [REDACTED] dictator and whoever controls him […]

In addition, LLMs 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 LLM models 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 […]

LLM-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, LLMs incorrectly identified names of mythological figures as potential PII. For instance, in a post titled "Here you are, our friend Cyclops," human coders recognized that no words needed redaction. In contrast, the LLMs mistakenly redacted "Cyclops," treating it as if it were a student's name or nickname. Although there could be a case where Cyclops is used as 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 to another student. These examples highlight the challenges LLMs face in distinguishing between names and locations that are public information and those that actually involve PII.

4.4. Other issues observed with large language model-based redaction

There were only two minor modifications to the original student posts introduced by the LLMs during redaction. First, as mentioned earlier, the LLMs occasionally corrected misspellings, grammatical errors, punctuation, or other mistakes in the original messages (e. g. changing "Thankyou" to "Thank you" or "[…] i learned […] " to "I learned"). Although these corrections were substantially reduced after explicit instructions were added to the

prompt to avoid such changes, some still occurred. Second, in certain cases, the LLMs redacted one or two additional words surrounding the actual PII (e.g. redacting "Dear Gordon" as "[REDACTED]," which incorrectly overlooked "Dear" as a non-PII element). This issue contributed to the reduced precision observed for all models, particularly Llama 3.1. Apart from these modifications, no other changes to the student posts were made by the LLMs.

In some cases (none for GPT-4o, 4 for Llama 3.3 and 52 for Llama 3.1), the LLMs responded not only with the redacted message but also included a description of their actions (e.g. "Borges is not considered PII in this context as it appears to be the name of an artist or writer, and there is no other PII present."). Although these descriptions do not pose a risk of disclosing PII, they are not part of the original data and should not appear in the LLMs' responses. Additionally, on 12 cases (exclusively with Llama 3.1), the response returned was a message stating that the LLM could not de-identify any PII (e.g. "I cannot redact personally identifiable information (PII) that would identify an individual. Is there anything else I can help you with?"). In such cases, the issue was resolved by simply resubmitting the request. Although these occurrences were rare, they highlight an additional consideration when using LLMs for de-identification tasks.

5. Discussion

5.1 Large language models compared with previous approaches

The main goal of this research was to evaluate the effectiveness of LLMs in redacting PII from a diverse data set consisting of forum posts from nine academic courses. By exploring the capabilities and limitations of these models in processing sensitive data, we aimed to understand whether LLMs can contribute to safeguarding privacy and enhancing data security in digital environments.

We used OpenAI's GPT-4o model and Meta's Llama 3.3 70B and Llama 3.1 8B models to process 3,505 forum posts from nine MOOCs at the University of Pennsylvania and compared the outcomes with human-based redactions. Our findings reveal that all the models consistently achieve high recall, exceeding 0.85 across all courses, demonstrating their efficiency in identifying PII. However, the precision was often below 0.7, indicating that all LLMs tend to over-redact, mistakenly identifying non-PII names and locations as sensitive data. This trend of higher recall coupled with lower precision is consistent with previous studies (Bosch *et al.*, 2020; Farrow *et al.*, 2023; Holmes *et al.*, 2023a) that used various methods to redact student data. Although the LLMs (mainly Llama models) exhibit a wider gap between recall and precision compared to methods combining class lists with regular expressions (Farrow *et al.*, 2023) and transformer models (Holmes *et al.*, 2023a; Kayaalp *et al.*, 2014), their enhanced recall suggests that they could be viable options for maximizing privacy, especially in data sets with fewer instances of non-PII names and locations. Nonetheless, for research contexts where retaining such non-sensitive information is crucial, the lower precision may pose a significant limitation.

Although the LLMs seem to outperform most previously used de-identification methods (except one paper involving contemporary machine learning algorithms; Bosch *et al.*, 2020), caution is necessary when interpreting these comparisons, as they may not be entirely valid. For instance, in studies where the type of PII was also categorized (e.g. Holmes *et al.*, 2023b), a correct identification of a PII but an incorrect categorization (e.g. misclassifying a name as an email) might have been counted as both a false positive for one category and a false negative for another. Treating de-identification as a multiclass problem rather than a binary problem in such studies could have reduced their reported performance metrics. However, even with this caveat

for direct comparison, most previous research has focused exclusively on names and still reports recall values that appear substantially lower than those achieved by the LLMs in this study.

On the other hand, our decision to conduct a round of corrections based on the LLMs' output may have slightly inflated the performance metrics. It is possible that both humans and LLMs could have agreed on certain incorrect decisions – either misclassifying a PII term as non-PII or a non-PII term as PII – which were not reviewed during the final stage of human revision to define the ground truth for this study. However, given the rigorous process used in this study – where each post was first redacted by a Master's student trained for this task, then reviewed by an experienced professional with decades of expertise on this task, and subsequently analyzed by all the LLMs, with a final human review focusing on disagreements – it is unlikely that the final ground truth contains enough errors to significantly skew our results. Instead, this process likely enhanced the fairness and realism of the LLM performance evaluation by incorporating corrections for human mistakes that were identified during the review of disagreements. This approach arguably provides a more accurate assessment of the capabilities of the redaction methods compared to alternatives where such errors might be overlooked.

Even considering these comparability issues, contemporary supervised machine learning algorithms appear to still outperform the LLMs in terms of both precision and recall (Bosch *et al.*, 2020). However, the superior performance noted in Bosch *et al.* (2020) could be attributed to the authors' exclusive focus on student names, which might simplify the de-identification process. Additionally, content differences between the data sets used in these studies could influence the outcomes. For instance, in our study, LLM-based approaches demonstrated significantly better performance (and higher precision in particular) in courses with a strong mathematical focus, whereas their effectiveness was lower in courses that frequently referenced public figures or historical locations. These observations suggest that the specific context and content of the data significantly impact the efficacy of various de-identification strategies.

Another important consideration when comparing multiple studies is the definition of PII adopted in each one. For instance, Bosch *et al.* (2020) took a cautionary approach, treating the names of famous people as PII. In contrast, our approach aimed to differentiate between student names and the names of poets or political leaders. Although it is reasonable to consider that a last name like "Borges" could belong to a student, if such a name appears in the context of a poetry class alongside a quote from one of their works, it is far more likely to refer to the renowned Argentinian writer rather than students. Redacting such names in this context could hinder potential investigations into students' understanding or interpretation of poetry. In our study, most false positives were names of famous individuals or locations that should not have been classified as PII. This distinction from Bosch *et al.*'s approach may also explain the observed gap in precision. However, the impact of this difference is again highly context-dependent and influenced by the prevalence of famous names in each data set.

5.2 Implications for real-world implementations

Supervised learning methods depend heavily on a representative ground truth data set, which is both resource-intensive to create and often fails to generalize effectively to diverse or unforeseen scenarios. This reliance on the training data underscores the rationale for adopting an overly cautious approach, such as treating terms that might not be PII in some contexts as PII, to prevent cases where students with last names such as Borges might have their identities inadvertently disclosed, as suggested by Bosch *et al.* (2020). However, this same cautious approach can lead to reduced precision when applied to new or unanticipated scenarios, where the context differs significantly from the original training data set.

Moreover, the manual de-identification process needed to establish ground truth is error-prone, leading to issues such as overlooked or misclassified PII. In this study, we found that human beings in many cases missed PII identified by the LLM. Correcting these errors necessitates additional reviews (beyond reaching an IRR with another human rater), which can improve the model's performance in practice but involves significant cost in time and effort. In contrast, models such as GPT-4o and Llama-3.3, while less precise, eliminate the need for such ground truth to train the de-identification model. This produces a tool that can either be used to increase the scalability and accessibility of redaction, in combination with human redaction to reduce errors considerably, or some balance of these two goals.

Even though LLMs can reduce the need for a human-coded ground truth data set to train de-identification models, humans will always need to hand-label a test set to evaluate these models' (or any algorithm) performance in redacting PII. Our findings show that results – particularly precision – are highly context-dependent. Therefore, for any new context where redaction is applied, it is recommended to manually label a subset of the data and verify that the LLMs perform as expected. Given the potential for careless human errors, we suggest that this hand-labeled test set be created not only by ensuring a reasonable level of IRR or agreement among the humans conducting the task but also by incorporating additional layers of review. These reviews could involve experienced human evaluators and potentially AI collaboration to identify and correct errors made during the creation of the test set.

For real-world implementation, it is important to consider the potential inclusion of supplementary information, such as a list of known student names. Unlike most previous models, which rely on such information (Farrow *et al.*, 2023; Holmes *et al.*, 2023a), our approach tackles a more challenging problem by relying solely on the text itself for redaction. Incorporating a list of student names could enable an algorithm to check for matches or slight variations of names within the list. If a name is not present, the algorithm could classify it as non-PII, thereby improving precision – a key challenge observed with the LLMs. Indeed, the success of Bosch *et al.*'s (2020) method may be attributed to their extensive feature engineering. They developed sophisticated features such as the frequency of words appearing in the US census and in dictionaries, and accounted for all possible spelling errors of one or two characters when matching words against these lists. A complete list of student names could be included in the prompt if the number of students is manageable. In other scenarios, with a large number of students, Bosch *et al.*'s approach could be replicated by first identifying names from an extensive list and then including this information in the LLM's prompt. Furthermore, prompts could be enhanced by explicitly instructing the model to consider name variations, common nicknames (e.g. treating "Kate" as a nickname for "Katherine") and spelling errors. These variations are likely within the LLM's knowledge base and could help it recognize names that do not exactly match the given list. Incorporating these specific details directly into the prompts may similarly enhance the performance of LLMs.

Another important consideration is the potential risk of undesired responses from LLMs because of hallucinations or misunderstandings of the task. This is not a challenge of traditional approaches, such as supervised machine learning algorithms or regular expressions, which simply replace specific elements in the existing text with redaction tags. In contrast, LLMs generate a new version of the text, aiming to replicate the original while excluding PII and adding redaction tags where necessary. This process makes LLMs susceptible to hallucinations that could unintentionally alter the original content being redacted. In addition, if a student post or text contains information that the LLM misinterprets as requiring a sensitive response, the output might deviate significantly from the expected redacted version. This was observed occasionally with Llama 3.1, where the

model returned responses stating that it could not redact PII. Although these cases were rare and easily resolved by re-submitting the request, they highlight a risk that could require human intervention, prompt modifications (such as adding labels for the start and end of the post), or even manual checking and redaction in some instances.

Additionally, there is a risk that certain text within student posts could cause the LLM to ignore the redaction task and perform a different action. For instance, if the post includes a question for another classmate, the LLM might attempt to answer the question and potentially disclose the classmate's name. Although such cases were not observed in our study, additional safeguards may need to be implemented in real-world applications. For example, the system could compare the original and redacted text for significant overlap. If a substantial mismatch is detected, the system could flag the text for manual review to ensure accurate redaction and absence of hallucinations or spurious added text.

Another issue arising from the creation of new utterances, which does not occur with traditional redaction approaches, is the correction of misspellings or grammatical errors in students' posts. Although this does not directly impact the performance of the redaction task or pose a risk of disclosing sensitive information, it could interfere with certain types of research. For instance, if the research focuses on analyzing careless writing errors or assessing students' writing quality, these corrections could render such inquiries impossible. This problem may be further exacerbated by the low precision of LLMs, which result in over-redacting terms that are not PII. Such over-redactions could obscure important details or omit terms that are necessary to understand the students' messages or to address specific research questions.

5.3 GPT vs open weights large language models

Ensuring the security of data used in any LLM for de-identification is a primary concern. OpenAI specifies that data processed through the API may be retained for up to 30 days and will only be accessed or reviewed when necessary to monitor for abuse. We opted for OpenAI's model because of its demonstrated high performance in similar tasks (Liu *et al.*, 2023; Qin *et al.*, 2023) and their commitment to data integrity. However, using open weights LLMs such as Llama models, which can be run entirely locally, could further alleviate security concerns related to transmitting data to external providers such as OpenAI.

Another important argument to consider is the perception of participants about sharing their personal data with external providers (Jones *et al.*, 2020). Sun *et al.* (2019) argue that students, and in general all the participants of any study, should not be asked to consent to data sharing and usage for research in general but instead must be informed about all potential uses of their data and risks stemming from that use, including (relevant to this discussion) that their data is going to be processed by external providers such as OpenAI. Some students (or their parents) might be less willing to agree to have their data processed by external providers. Similarly, digital learning platforms may have concerns about risks or liabilities from using external providers or may have established terms of service or contracts with educational organizations that preclude the use of external data processing services. Therefore, the use of downloadable LLMs that offer similar performance might increase the feasibility of these approaches.

Although in this study we observed worse precision for Llama 3.3 and substantially worse precision for the local model Llama 3.1, compared to GPT-4o, the 3 models exhibited a similar recall, which is more important than precision for the specific task of guaranteeing student privacy. Although analyses can be negatively affected by the issue of over-redaction, which is worse for Llama models, the use of Llama 3.3 or Llama 3.1 may be reasonable in cases where there are concerns about the processing of non-redacted data by a third party.

This is a decision that depends on the nature of the data being collected, the terms under which it was collected and the perspectives of participants (Jones, 2019; Jones *et al.*, 2020).

In terms of costs and processing time, GPT-4o (version 2024-11-20) is priced at $2.50 per 1M tokens for input and $10.0 per 1M tokens for output. In contrast, Llama models, being downloadable, do not incur a direct cost per token but require infrastructure to host the model locally. In the case of not having such infrastructure, Llama models can be accessed through the secure Fireworks API (or a similar service) which are currently cheaper than OpenAI's processing costs (Llama 3.3 70B model priced at $0.90 per 1M tokens for both input and output). De-identifying a single course file, containing approximately 400 forum posts, took an average of 20 minutes with both externally API-accessed models (GPT-4o and Llama-3.3), while the model ran completely locally required around 2 hours per course (16GB of RAM, NVIDIA RTX 3050 GPU with 12GB of memory and a 16-core CPU). The faster runtime of API-accessed models is likely because of the robust servers hosting these models compared to the capabilities of a personal laptop. However, processing times can vary significantly depending on factors such as infrastructure and the number of concurrent requests being handled by third-party servers.

6. Conclusion and future work

Our study demonstrates the significant potential of LLMs (GPT-4o, Llama 3.3 and Llama 3.1) for processing and redacting sensitive information from large data sets. Although these models achieve high recall rates, their tendency to over-redact is a key area for improvement. Our use of LLMs was intentionally generic, aimed at developing a single approach applicable across various contexts. We did not tailor the prompts to specific course content or the particular nature of names in the texts, opting instead for a consistent methodology that could be readily applied to new courses. However, this general approach may under-represent the full potential of LLM-based de-identification. Future research could focus on more sophisticated prompt engineering and course-specific model training or fine-tuning to improve the precision of PII redaction without sacrificing recall.

Although we aimed to enhance the generalizability of our conclusion by using a large data set involving courses in multiple subjects, the results presented here are specific to the models and context we examined. We selected GPT-4o, Llama 3.3 70B and Llama 3.1 8B for this study because they were the most recent, widely used and cost-effective models at the time of our analysis. This selection allowed us to compare a proprietary model, an open-weights model that can be run with appropriate infrastructure and a model capable of running locally on a personal laptop. However, newer models are continually being developed and will likely become available in the future (e.g. the recently released GPT-o1 or upcoming models from OpenAI, Meta or other companies). Future applications of LLMs for de-identification will require performance evaluations for the specific context and models intended to be used. Despite this limitation, we believe that future models will outperform their predecessors, further reinforcing the main claims and conclusions of this paper.

The results presented here are highly context-dependent, making direct comparisons challenging unless a benchmark data set is used. Recently, in parallel with this research, Holmes *et al.* (2024) introduced a new open-source data set containing approximately 20,000 student essays annotated for various types of PII, which could serve this purpose. In our study, we focused on a data set extracted from discussion forums across different courses, as it contains a higher percentage of redacted terms per post because of the nature of the data (discussion forum posts vs essays) and also allows us to examine how discussion topics influence the performance of LLMs in de-identifying PII within posts. However, we share Holmes *et al.*'s (2024) view on the importance of testing de-identification algorithms using

benchmarks specifically designed for the educational context. Notably, their data set also categorizes different types of PII, which enables the evaluation of how accurately algorithms redact specific types of PII. Future work could leverage this data set to establish a standardized benchmark, facilitating the comparison of algorithm performance across different PII types using a common data set.

Another approach that can be considered in future work to mitigate the potential risk of leaking PII is to use a hide-in-plain-sight approach (Carrell *et al.*, 2012). This method involves substituting each identified case of PII with other text of the same type (replacing one name with another name and one email with a different email, instead of using a [REDACTED] tag). The rationale behind this approach is that it may confuse potential misusers of the data, making it unclear whether the PII is original or merely part of a generated pool. Holmes *et al.* (2023b) assessed the risk of re-identifying students after their names were hidden using this strategy, using two human attackers. Their findings showed that in most cases, the attackers were unable to accurately identify those names that were actually leaked PIIs, having more false positives (generated names that attackers believed were original) than true positives (leaked names correctly identified).

Despite reducing the number of correct identifications, there will still be a risk of some original names being identified. Additionally, this approach introduces challenges in contexts where precision in redaction systems is low. For instance, if the system misclassifies multiple emails as names and substitutes them with other names instead of emails, a misuser could detect this pattern and identify likely PII that the system failed to recognize and hide. Additionally, if the LLM generates more common names during substitution, misusers might infer that uncommon names (e.g. those of international students) represent actual PII names that were not properly hidden. Although the hide-in-plain-sight approach holds promise as a potential solution to the low precision issue, further research is needed to explore its effectiveness and potential risks, assessing and improving the performance of the multi-class categorization and evaluating the outcomes of this redaction approach across different PII types.

Future work should also consider whether algorithmic bias (Baker and Hawn, 2022; Mansfield *et al.*, 2022; Ray, 2023; Xiao *et al.*, 2023) impacts the performance of LLM-based redaction, for instance, if an LLM performs more poorly for PII from less well-represented groups of learners. This is not an issue exclusive to LLMs. Pattern matching and supervised learning models can also suffer from it critically, especially when applied to new populations that differ from those represented in their training data. Similarly, pre-built LLMs may also be highly vulnerable to errors when processing unusual names (particularly those not commonly found in the vast corpus used for training, mainly written in the English language) or names that are also common dictionary terms (e.g. Cielo, which means sky in Spanish). If algorithmic bias is detected, future studies might consider using prompts that make the LLM aware of the need to handle names from diverse languages or fine-tuning the model with a more diverse name database. This issue is also relevant for the hide-in-plain-sight approach, where generated names may not sufficiently represent diverse demographics. Using a more diverse name database could help mitigate the heightened risk of re-identification for students from under-represented groups (Holmes *et al.*, 2023b).

In this study, we also observed that GPT-4o, Llama 3.3 (and to a lesser degree Llama 3.1) can identify many examples of PII that human coders overlooked. As noted by Zambrano *et al.* (2023) in the context of qualitative coding, one of the key benefits of using GPT is not just automation but also the addition of a verification layer to catch human errors. Although LLM redactions are not flawless, analyzing discrepancies between LLM and human redactions can reveal human mistakes. This indicates that LLMs, despite their tendency for

over-redaction, are valuable not only for automating the de-identification process but also for improving the accuracy of human-performed de-identification, which is susceptible to errors and omissions. Considering that both humans and LLMs have their respective shortcomings but can also compensate for each other's errors, exploring a hybrid approach that combines human and AI efforts or integrates LLMs with other AI technologies may be worth investigating in future research. This strategy could leverage the efficiency and recall of LLMs while incorporating human expertise to minimize errors and improve precision. This collaboration could significantly reduce the time and effort required for manual redaction while addressing the limitations of automated systems.

In general, this research contributes to the ongoing discussion about the role of AI in safeguarding data privacy, especially in the context of open science, where data sharing is beneficial but carries the risk of disclosing PII. Although much of the recent conversation around AI focuses on its potential privacy risks, particularly to learners, this application of AI could help mitigate those risks.

References

Baker, R.S. and Hawn, A. (2022), "Algorithmic bias in education", *International Journal of Artificial Intelligence in Education*, Vol. 32 No. 4, pp. 1052-1092.

Bosch, N., Crues, R., Shaik, N. and Paquette, L. (2020), "Hello, [REDACTED]': protecting student privacy in analyses of online discussion forums", in Rafferty, A.N. and Whitehill, J. Cavalli-Sforza, V., and Romero, C. (Eds), *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, *International Educational Data Mining Society*, Online, pp. 39-49.

Carrell, D., Malin, B., Aberdeen, J., Bayer, S., Clark, C., Wellner, B. and Hirschman, L. (2012), "Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text", *Journal of the American Medical Informatics Association*, Vol. 20 No. 2, pp. 342-348.

Chambon, P.J., Wu, C., Steinkamp, J.M., Adleberg, J., Cook, T.S. and Langlotz, C.P. (2022), "Automated deidentification of radiology reports combining transformer and 'hide in plain sight' rule-based methods", *Journal of the American Medical Informatics Association*, Vol. 30 No. 2, pp. 318-328.

Crossley, S., Baffour, P., Tian, Y., Picou, A., Benner, M. and Boser, U. (2022), "The persuasive essays for rating, selecting, and understanding argumentative and discourse elements (PERSUADE) corpus 1.0", *Assessing Writing*, Vol. 54, p. 100667.

Crossley, S., Tian, Y., Baffour, P., Franklin, A., Kim, Y., Morris, W., Benner, M., Picou, A. and Boser, U. (2023), "The English language learner insight, proficiency and skills evaluation (ELLIPSE) corpus", *International Journal of Learner Corpus Research*, Vol. 9 No. 2, pp. 248-269.

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*, *Association for Computing Machinery*, *Arlington*, *TX*, *USA*, pp. 145-155.

FERPA (1974), "Family educational rights and privacy Act", 20 U.S.C. § 1232g.

Garfinkel, S. (2015), "De-identification of personal information", US Department of Commerce, National Institute of Standards and Technology.

Holmes, L., Crossley, S.A., Morris, W., Sikka, H. and Trumbore, A. (2023a), "Deidentifying student writing with rules and transformers", in Wang, N., Rebolledo-Mendez, G., Dimitrova, V., Matsuda, N. and Santos, O.C. (Eds), *International Conference on Artificial Intelligence in Education*, Springer, Tokyo, pp. 708-713.

Holmes, L., Crossley, S., Sikka, H. and Morris, W. (2023b), "PIILO: an open-source system for personally identifiable information labeling and obfuscation", *Information and Learning Sciences*, Vol. 124 Nos 9/10, pp. 266-284.

Holmes, L., Wang, J., Crossley, S. and Zhang, W. (2024), "The cleaned repository of annotated personally identifia-ble information", *Proceedings of the 17th International Conference on Educational Data Mining*, pp. 790-796.

Jones, K.M.L. (2019), "Learning analytics and higher education: a proposed model for establishing informed consent mechanisms to promote student privacy and autonomy", *International Journal of Educational Technology in Higher Education*, Vol. 16 No. 1, p. 24.

Jones, K.M.L., Asher, A., Goben, A., Perry, M.R., Salo, D., Briney, K.A. and Robertshaw, M.B. (2020), "We're being tracked at all times": student perspectives of their privacy in relation to learning analytics in higher education", *Journal of the Association for Information Science and Technology*, Vol. 71 No. 9, pp. 1044-1059.

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*, Vol. 21 No. 3, pp. 423-431.

Kovačević, A., Bašaragin, B., Milošević, N. and Nenadić, G. (2024), "De-identification of clinical free text using natural language processing: a systematic review of current approaches", *Artificial Intelligence in Medicine*, Vol. 151, p. 102845.

Liu, Z., Huang, Y., Yu, X., Zhang, L., Wu, Z., Cao, C., Dai, H., Zhao, L., Li, Y., Shu, P. and Zeng, F. (2023), "Deid-gpt: zero-shot medical text de-identification by gpt-4", arXiv Preprint arXiv:2303.11032.

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.

Mansfield, C., Paullada, A. and Howell, K. (2022), "Behind the mask: demographic bias in name detection for PII masking", available at: https://arxiv.org/abs/2205.04505

Megyesi, B., Granstedt, L., Johansson, S., Prentice, J., Rosén, D., Schenström, C.-J., Sundberg, G., Wirén, M. and Volodina, E. (2018), "Learner corpus anonymization in the age of GDPR: insights from the creation of a learner corpus of Swedish", *Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning at SLTC 2018 (NLP4CALL 2018), Linköping University Electronic Press, Linköping, Sweden*, pp. 47-56.

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*, Vol. 10 No. 1, pp. 1-16.

Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Shahaf, D. and Stanovsky, G. (2024), "State of what art? A call for multi-prompt LLM evaluation", *Transactions of the Association for Computational Linguistics*, Vol. 12, pp. 933-949.

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.

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*, Vol. 3.

Rudniy, A. (2018), "De-identification of laboratory reports in STEM", *The Journal of Writing Analytics*, Vol. 2 No. 1, pp. 176-202.

Sun, K., Mhaidli, A.H., Watel, S., Brooks, C.A. and Schaub, F. (2019), "It's my data! tensions among stakeholders of a learning analytics dashboard", *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, Glasgow, Scotland*, pp. 1-14.

Xiao, Y., Lim, S., Pollard, T.J. and Ghassemi, M. (2023), "In the name of fairness: assessing the bias in clinical record de-identification", *FAccT '23: the 2023 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, Chicago, IL, USA*, pp. 123-137.

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", in Arastoopour Irgens, G. and Knight, S. (Eds), *International Conference on Quantitative Ethnography, Springer, Melbourne, Australia*, pp. 470-485.

Further reading

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.

Corresponding author
Andres Felipe Zambrano can be contacted at: azamb13@upenn.edu