Data Anonymization for Privacy-Preserving Large Language Model Fine-Tuning on Call Transcripts

Nathan Zhang, Anne Paling, †Preston Thomas, Tania Habib,
Mahsa Azizi, Shayna Gardiner, Kevin Humphreys, †Frederic Mailhot*
Dialpad Canada Inc., †Dialpad Inc.
{nzhang, anne, preston, tania.habib, mahsa.azizi,
sgardiner, kevin.humphreys, fred.mailhot}@dialpad.com

Abstract

Large language models in public-facing industrial applications must accurately process data for the domain in which they are deployed, but they must not leak sensitive or confidential information when used. We present a process for anonymizing training data, a framework for quantitatively and qualitatively assessing the effectiveness of this process, and an assessment of the effectiveness of models fine-tuned on anonymized data in comparison with commercially available LLM APIs.

1 Data Privacy in the era of LLMs

Recent progress in the capabilities of large language models (LLMs) (Devlin et al., 2019; Brown et al., 2020; Zhao et al., 2023), has led to their widespread adoption as the foundation for a variety of tasks in industrial and academic NLP (Bommasani et al., 2021). With parameter counts in the tens and hundreds of billions, these models require vast amounts of data to train and fine-tune (Hoffmann et al., 2022). At the same time, this overparameterization enables the memorization and potential leakage or extraction of large portions of LLMs' training data (Biderman et al., 2023; Carlini et al., 2023; Hartmann et al., 2023). Taken together, the required volume of training data and memorization capabilities of LLMs raise substantial issues concerning data privacy (Li et al., 2023). This risk is compounded because LLMs, like all supervised learners, perform best on test sets that have similar distributions to their training data. Thus, organizations seeking to deploy practically effective LLMs must train them with data that reflect the distribution of their deployment, with specific, sensitive data such as medical records or call transcripts leading to improved performance, but correspondingly greater risk of exposing that data to breaches or adversarial attacks (Nasr et al., 2023).

Furthermore, the lack of predictability and difficulty in constraining the outputs of LLMs means that including personal information (PI) in a training or fine-tuning data split runs the risk of this data being exposed in output generated by the model — even in the absence of adversarial attacks and when the task does not call for such data. Maximal mitigation of this risk requires removing all instances of PI from the training data, for example by excising any sentences that contain PI, or redacting any PI tokens. This kind of full exclusion leads to the challenge discussed above: depending on its use-case or deployment environment, a model may need to process and respond to PI at inference time. Suppressing all instances of PI, effectively removing the entire entity, is an approach seen when undertaking anonymization of structured data, however with unstructured text as in this context, this is not a realistic option due to resulting in training data that will be distributionally and semantically (Hassan et al., 2023) different from the input. Additionally, these types of data perturbations have been shown to negatively impact model performance (Malle et al., 2016, 2017). A more targeted approach to PI token redaction, tagging a set of candidate PI tokens with tags from a pre-defined taxonomy, is offered by some companies as a publicly-available anonymization service.

In this paper we leverage and modify such an anonymization service, proposing a nuanced approach to token redaction and risk assessment, showing that these measures can address the standard trade-off between privacy protection and performance. Our specific contributions are:

- Modifications to the taxonomy of PI categories defined by Google's Cloud Data Loss Prevention service[1] that serve to increase the

*Corresponding author. We would like to thank our anonymous reviewers for detailed and helpful feedback, and our colleagues Mel Andersen and Tere Roldán for their assistance with data annotation.

[1] https://cloud.google.com/security/products/

Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization
CALD-pseudo 2024, pages 64–75

accuracy of anonymization of call transcripts generated by a proprietary *automatic speech recognition* (ASR) system.

- A framework for evaluating our modified anonymization pipeline with respect to *residual risk*: a measure encompassing both the likelihood of identifying an individual from residual PI that persists after anonymization, and the relative magnitude of harm based on the sensitivity of the remaining data. When properly calibrated, residual risk scoring for arbitrary combinations of PI or partial PI should closely align with the potential real-world impact of their exposure.
- A demonstration that a model fine-tuned with data that has been anonymized in accordance with our approach shows comparable F_1 and ROUGE scores to other popular LLMs on four in-domain tasks, with acceptable levels of residual risk.

1.1 Related Work

Data anonymization Elliot et al. (2020) present a framework for data anonymization, including a taxonomy of identifiers with different risk/exposure profiles. The framework's purpose is to furnish practical understanding of anonymization for use in business or organizational contexts. It is designed to control the risk of unintended re-identification and disclosure.

The problem of automated data anonymization specifically in the context of textual data is investigated by Lison et al. (2021). They draw links between work done in this area in the fields of NLP and privacy-preserving data publishing, and highlight some general challenges, including the trade-off between data utility and residual risk, and how to assess the quality of anonymization.

Privacy-preserving LLM/ML training Xu et al. (2021) provide a systematic review of existing privacy-preserving machine learning (PPML) approaches. They propose a Phase, Guarantee, and Utility based model to understand and guide the evaluation of various PPML solutions by decomposing their privacy-preserving functionalities.

Plant et al. (2022) empirically investigate the extent to which personal information is encoded in the representations of a variety of widely-available pre-trained LLMs. They demonstrate a positive correlation between the complexity of a model, the data volume used in pre-training, and data leakage. In addition, they present an evaluation and comparison of some popular privacy-preserving algorithms on a large multi-lingual sentiment analysis data set annotated with demographic information (location, age and gender). Their results show that larger and more complex models are more prone to leaking private information, and hence that the use of privacy-preserving methods is necessary. In addition to the preceding domain-general investigations, Yin and Habernal (2022) and Guerra-Manzanares et al. (2023) investigate some of the challenges of privacy-preserving training for machine learning and language modeling in the legal and healthcare domains, including increased resource needs to address the high computational complexity of some methods (e.g. homomorphic encryption), and privacy/accuracy trade-offs for methods with strong guarantees (e.g. differential privacy).

2 Data

The data set to be anonymized consists of transcripts generated by an internal proprietary ASR system. Raw transcripts are passed through an inverse text normalization module to generate final formatted transcripts. The transcripts in the data set include phone and video conference conversations between at least one and usually two or more speakers in business contexts, such as voicemails (single speaker), call center conversations (typically two speakers) and internal company meetings (two or more).

Transcripts generated from an ASR system are imperfect due to characteristics common to businesses, such as noisy environments, fast or quiet speakers, and poor-quality microphones. Recognition errors propagate to the final transcription, which can create difficulties in applying and evaluating the anonymization process.

3 Anonymization Process

Mindful of the ongoing discussion over the appropriate terminology for such processes (Garfinkey, 2015), we use the term "anonymization" herein because the intended outcome of our method is that no individual can be identified from the resulting text. Additionally, specifying anonymization distinguishes our method from *pseudonymization*, which appears superficially similar in that it includes replacing PI with tokens, e.g. *[PERSON_NAME_1]*.

`dlp`

The difference is that pseudonymization maintains a consistent mapping of the replacement token across conversations, potentially permitting later reidentification, whereas our process reuses these de-identified tokens across conversations, functionally eliminating the possibility of using them for re-identification purposes.

3.1 PI identification

There are several commercial offerings for PI identification and anonymization of text data. We surveyed services by Amazon,[2] Microsoft,[3] and Google.[4] We selected Google's Cloud Data Loss Prevention (DLP) service due to its broader coverage of PI categories. The DLP service defines a taxonomy of *information types*, or *infoTypes*; kinds of sensitive data such as names, email addresses, and telephone numbers.[5] An additional advantage of using the DLP service was the in-house access to data stored in BigQuery[6] and the ease of creating a configuration template to set up asynchronous jobs for large volumes of data, which was well suited for our use case.

In the PI identification process, we included most of the global infoTypes from the available taxonomy, as well as those infoTypes which are specific to the US and Canada (e.g. social security or social insurance numbers). A preliminary analysis suggested that the *ETHNIC_GROUP*, *GENDER*, *DATE*, and *TIME* infoTypes had a much higher rate of false positives (FPs) in our data sets, and so we excluded them. The taxonomy also includes categories for human names. The categories *FEMALE_NAME*, *MALE_NAME*, *FIRST_NAME*, and *LAST_NAME* are individually and collectively subsets of the *PERSON_NAME* infoType, and so we retain the latter while excluding all of the former.

We made the following modifications to DLP's PI identification to improve its performance on our data set:

1. **Exclusion List:** On the basis of the most frequent FPs seen in the masked transcripts we created an exclusion list for the *PERSON_NAME*, *ORGANIZATION_NAME*, and *LOCATION* infoTypes.

2. **Custom dictionary:** A custom dictionary was added to the PI detection configuration for the two infoTypes of *PERSON_NAME* and *ORGANIZATION_NAME* to reduce the number of false negatives (FNs) and increase the chance of correctly detecting names of organizations and people in the transcripts. Both of these resources were developed from a proprietary database of company and user names.

3. **Letters and digits**: After preliminary evaluation of the DLP API on our data, two additional infoTypes are created and added to the identification configurations:

    - *Spelled words:* Our transcription engine transcribes and formats letter sequences, for example verbally spelled-out words, with hyphens as separators e.g. A-L-P-H-A. The DLP API fails to detect and mask these instances, leaving potential PI in the anonymized data. A regular expression pattern to detect such groups in the transcript was added to the custom dictionary.
    - *Numbers:* Although our transcription engine can successfully decode and format digit sequences such as phone numbers, if a user repeats digits, or there is a transcription error such as "four" mistranscribed as "for", there is the potential for an unformatted sequence of digits to appear in the transcripts, which may not be detected by the DLP API. We therefore added a regular expression to detect numeric sequences of length 3, reducing the risk of missing potentially identifiable data due to mistranscription.

4. **Usernames:** Although they fall within the scope of DLP's built-in *GENERIC_ID* infoType, usernames such as *enigma52* or *Mrbigchef* were consistently not tagged as potential PI by the DLP service. We identified instances where these unmasked usernames were used across multiple social media platforms, or were some combination of multiple pieces of PI such as *first initial + last name*, *first name + last name* or *first name + birth*

[2] https://docs.aws.amazon.com/transcribe/latest/dg/pii-redaction.html

[3] https://learn.microsoft.com/en-us/azure/ai-services/language-service/personally-identifiable-information/how-to-call

[4] https://cloud.google.com/dlp/docs/sensitive-data-protection-overview

[5] For a complete list, see https://cloud.google.com/dlp/docs/infotypes-reference.

[6] https://cloud.google.com/bigquery

Original Transcript
Pam: This is Pam calling from Dunder Mifflin, may I speak to Jim?
Anonymized Transcript
[PERSON_NAME_1]: This is *[PERSON_NAME_1]* calling from *[ORGANIZATION_NAME_1]*, may I speak to *[PERSON_NAME_2]*?

Table 1: Example of context-aware anonymization

year, and in each of these instances could be used to identify the user.

The DLP service can detect domain or data-specific entities via the creation of a *hotword regex* (regular expression). We improved the detection accuracy of *GENERIC_ID* and *PERSON_NAME* in two ways. Firstly, we created a hotword regex for mentions of the word *user name* in our data, e.g. `(username|user name|Username|user ID)` and defined a context window of 100 characters around the hotword regex as an area of higher likelihood username detection. Secondly, we added a custom regex to the *GENERIC_ID* infoType to detect alphanumeric sequences of a certain length and commonly-used conditions for creating a username. Together, these approaches increased the hit rates for usernames up to 66% in our data set.

5. **Context-aware anonymization:** DLP does not offer means of differentiating tokens or instances of identified infoTypes, thus losing semantic information in the application of the anonymized text. In order to preserve context for later analysis, each masked span is assigned a unique numeric ID within the call. Multiple instances of the same masked information are assigned the same ID. See Table 1 for an example. [7]

4 Residual Risk Analysis

4.1 Identifying and annotating residual risks

Transcripts that were redacted using Google's DLP were subsequently annotated by humans to identify any residual PI that had not been detected, with a subset being subject to a second pass for verification.[8] Annotation of residual PI proved to be challenging, requiring multiple iterations of the guidelines with our annotators. Table 2 shows the output of post-anonymization annotation on a fictitious example.

Annotators did not tag instances of undetected PI that were not relevant to personal identification, even if there was an associated infoType. For example, DLP redacts generic ID numbers, but missed instances of these were only tagged if they could contribute to identifying an individual — for instance, organization-internal order numbers were not tagged, while a business registration license number would be. This choice was made because our goal was not to evaluate the accuracy of the PI tagging *per se*, but rather to quantify the risk of residual PI after anonymization. In Table 2, a transcription error results in partial detection, hence only partial anonymization, of the order number. It is not annotated, however, because the residual partial information of an internal order number is not usable for identifying the speaker.

Given the unstructured nature of transcript data and potential transcription errors, PI may be imperfectly formatted, so it may occur that only a portion of a span of PI is detected and tagged. To account for such cases, we associate with any given tag *[TAG]* a *TAG_PARTIAL* tag to be used by the annotators when only part of the PI is not anonymized. Table 2 demonstrates such cases; Person 1's last name and Person 2's email domain name are marked as *_PARTIAL*.

We found that the most common infoTypes missed by the de-identification process are *PRODUCT* and *ORGANIZATION*. There are two scenarios in which *PRODUCT* and *ORGANIZATION* are mentioned in a conversation: a common product or organization that can be used as a conversation topic, and the specific product or organization the speaker associates with. There is quite a difference between a person discussing using an iPhone from Apple and a person selling a product for their own company — the former does not provide much insight into identifying the speaker, but the latter does. However, DLP PI identification doesn't differentiate and doesn't have a consistent pattern in identifying the product and organization in the two scenarios. Therefore, in our evaluation, we only consider the missed product or organization to contain residual risk if they are closely related to the speaker. To illustrate, in Table 2, *Dunder Mifflin*

[7] Note that numeric IDs do not persist across calls, which would cross into pseudonymization and raise a reidentification risk.

[8] While we made some effort to mitigate false positives, this is not an issue that impacts our discussion here, which is concerned with preventing the leakage of PI.

Original Transcripts
Person 1: Dunder Mifflin, this is Rachel green speaking. Person 2: Hi, this is mark from ABC Trust Fund, we just ordered a set of paper and they have worse quality than staples. We would like to return and get refund. Person 1: Okay, what is the order number? Person 2: It's B. 231 C. for A. two. Person 1: And the email for that order? Person 2: It's M-K two one @abc.com
Anonymized Transcripts
Person 1: Dunder Mifflin, this is ***[PERSON_NAME_1]*** green speaking. Person 2: Hi, this is ***[PERSON_NAME_2]*** from ***[ORGANIZATION_NAME_1]***, we just ordered a set of paper and they have worse quality than staples. We would like to return and get refund. Person 1: Okay, what is the order number? Person 2: It's B. ***[NUMERIC]*** C. for A. two. Person 1: And the email for that order? Person 2: It's M-K two one ***[EMAIL_1]***
Anonymized + Annotated Transcripts
Person 1: ***(Dunder Mifflin)[MISSED_ORGANIZATION_NAME_SPEAKER]***, this is ***[PERSON_NAME_1] (Green)[MISSED_PERSON_NAME_PARTIAL]*** speaking. Person 2: Hi, this is ***[PERSON_NAME_2]*** from ***[ORGANIZATION_NAME_1]***, we just ordered a set of paper and they have worse quality than ***(staples)[MISSED_ORGANIZATION_NAME]***. We would like to return and get refund. Person 1: Okay, what is the order number? Person 2: It's B. ***[NUMERIC]*** C. for A. two. Person 1: And the email for that order? Person 2: It's ***(M-K two one)[MISSED_EMAIL_PARTIAL] [EMAIL_1]***

Table 2: Example of post-anonymization annotation of residual PI. Missed PI is enclosed in parentheses and assigned a tag derived from the associated infoType.

was marked with the tag *_SPEAKER* to denote the risk associated with the missed PI, while *Staples* was not as it does not associate with any speakers in the conversation.

4.2 Quantifying residual risk

To assess residual risk for a conversation, we must first quantify the risk for each infoType. We begin by distinguishing *direct* and *indirect* identifiers, following Elliot et al. (2020):

- **Direct Identifier:** A variable or set of variables specific to an individual (e.g. name, address, phone number, bank account) that are explicitly or commonly used for the purpose of identification. These identifiers have a comparatively higher risk profile.
- **Indirect Identifier:** Information that in isolation does not enable identification (e.g. gender, nationality, city of residence), but may do so in combination with other indirect identifiers and/or background knowledge. These identifiers have a reduced but non-zero risk profile.

Residual risk is assigned an integer score ranging from 0 to 5. As stated above, for direct identifiers such as a person's name, credit card number, passport number, or social security/insurance number, we assign a maximal risk score of 5, to account for both the specificity of the identifier (a proxy for the likelihood of re-identification) and the impact of potential misuse. For indirect identifiers like company name or city of residence, we assign a risk score of 2 or 3. These risk categorizations for different infoTypes were developed in collaboration with privacy counsel. For the full list of categorizations and scores, see Table 6 in Appendix A.

For partially-redacted PI, tagged with the *_PARTIAL* annotation, the risk score of the associated tag is halved and rounded up or down to the next higher or lower integer, depending on the risk profile. For example, if the tag *[MISSED_EMAIL]* has a score of 3, then the score for *[MISSED_EMAIL_PARTIAL]* becomes $\lfloor 3/2 \rfloor = 1$. In the case of *[PERSON_NAME]*, which has a base score of 5, we round the score for *[MISSED_PERSON_NAME] up* to 3 to account for the wide variety of circumstances in which *[MISSED_PERSON_NAME_PARTIAL]* can occur. As noted above, we consider both first names and last names as *_PARTIAL* because the *[PERSON_NAME]* infoType is a superset of the other

[_NAME] infoTypes and using both created inconsistencies.

To calculate the residual risk score for an entire conversation, we sum the scores of the *[MISSED_]* tags, avoiding double-counting of multiple instances of a given token of missed PI. For example, in a conversation with four instances of *(Marc)([MISSED_PERSON_NAME_PARTIAL])*, the risk score contributed by this tag would be 3 rather than 12 ($= 4 \times 3$). Note that there can be variations in the spelling and formatting of a given token of PI due to ASR transcription error e.g. *Mark*, *Marc*, *M-A-R-K*. In this case, we consider them as a single piece of PI in three instances instead of different PIs if the annotators determine it is most likely a reference to the same entity.[9]

For the example in Table 2, the total residual risk is calculated as follows (tag names are shortened for consideration of space):

$$\begin{aligned} MISSED\_ORG\_NAME\_SPEAKER &= 2 \\ MISSED\_PERS\_NAME\_PARTIAL &= \lceil 5/2 \rceil = 3 \\ MISSED\_EMAIL\_PARTIAL &= \lfloor 3/2 \rfloor = 1 \\ \text{Total\_Risk\_Score} &= 2 + 3 + 1 = 6 \end{aligned} \tag{1}$$

The score assigned to $MISSED\_ORG\_NAME\_SPEAKER$ is 2 (see Appendix A, whereas the $PERSON$ and $EMAIL$ identifiers, being tagged $PARTIAL$ are halved and round up (down, resp), as discussed in Section 4.2. Note that the $MISSED\_ORGANIZATION\_NAME$ tag is not included in the calculation above as it is assigned a score of zero, because it does not represent a conversational participant, but is simply the name of a company.

4.3 Successful anonymization at the population level

We wish to know what proportion of a corpus of anonymized conversational transcripts carry an unacceptable residual risk profile. Pursuant to some preliminary data analysis, and in the absence of strong arguments to the contrary, we make the simplifying assumption that residual risks scores are well modeled by a normal distribution, $\mathcal{N} \sim (\mu, \sigma)$.

Given the previously-defined risk scores for each category, and our assumption of residual risk score normality, we define the following simple criterion as a measure of "successful" anonymization of a given conversational transcript:

$$\mu + \sigma < 5 \tag{2}$$

That is, we want the distribution of residual risk scores in our corpus of anonymized transcripts to be such that their mean plus one standard deviation is less than 5. We select 5 as our threshold of acceptability for the following reasons: (i) it is the risk score for a single occurrence of a direct identifier, which carries a maximal residual risk profile (high likelihood of re-identification and high impact of misuse), and (ii) it is equal to a combination of two complementary indirect identifiers such as company name + person's first name. Thus, 5 represents an easily-administerable target for assessing whether PI in the output of automated processes is sufficiently reduced to warrant more detailed review (see 4.5, below). We manually reviewed a set of high-scoring (above criterion) transcripts to ensure that this threshold met our needs.

Our assumption that residual risk is normally distributed implies that approximately 16% of our corpus of anonymized conversations carry a residual risk greater than 5.[10] Upon review of sample conversations, we find that anonymized transcripts with risk scores that are above the threshold—but do not have direct identifiers as part of the score—do not in practice enable re-identification. This is because as the indirect identifiers found in the masked text do not in general have a compounding effect. While we cannot *guarantee* the impossibility of re-identification in such, the risk after review was deemed acceptable. Table 3 provides an illustrative example: the total residual risk score is 6, with three independent instances of PI that do not combine to increase the risk of identification of any individual in the conversation.

We assessed the strength of our criterion manually, with anonymized call transcripts sampled from our corpora of business conversations in customer support, sales, videoconferencing, and direct call contexts. As shown in Table 4, after anonymization, human annotation of residual PI, and risk score assignment, none of the sampled corpora carried un-

[9] There is a potential difficulty here for conversations including multiple participants with the same name. We hope to address this in a future iteration of this work.

[10] Recall that one standard deviation to each side of the mean of a normal distribution accounts for approximately 68% of the probability mass. Since we are only worried about one tail of the distribution, i.e. the proportion with score greater than 5, we have one half of the tails' probability mass included in our coverage, for a total of 84%.

Person 1: Hi *[PERSON_NAME_2]*. This is *[PERSON_NAME_3]* calling back from *(XYZ lawyer)(MISSED_ORGANIZATION_NAME_SPEAKER)*. Person 2: Oh, hi. Person 1: I am calling regarding your request to change your business name on *(IRS dot gov)(MISSED_URL)* website. Person 2: Oh, yes, I want it to be changed to *(ABC incorporated) (MISSED_ORGANIZATION_NAME_SPEAKER)*.

Table 3: Example conversation where residual risk score over-represents practical impact

acceptable risk profiles with our criterion (although several did so at $\mu + 2\sigma$).

We conclude that a target residual risk score of 5 represents a conservative but readily achievable level of assurance that the anonymization procedure is effective.

Context	Count	Mean	STD	P95	Max	$\mu + \sigma$
Customer Support 1	100	0.7	1.4	3.1	6	2.1
Customer Support 2	98	1.3	2.1	6.0	11	3.4
Meetings	99	1.2	3.1	6.1	20	4.3
Sales Calls	100	0.6	1.3	3.1	6	1.9
1-to-1 Phone Calls	100	1.0	1.7	5.0	8	2.7
Total	498	1.0	2.0	5.0	20	3.0

Table 4: Residual risk analysis

4.4 Results

We sampled 498 conversations across four business communication products to ensure the representation of different conversation contexts, such as video conferencing, customer support, and sales calls. Table 4 shows the residual risk statistics of the five data sets.

Conversations in the video conferencing data set tend to be longer than the other data sets, with word counts five to six times that in other data sets. [11] For the samples with high residual risks, the identified PIs are not compounding, i.e., they include multiple indirect identifiers that all refer to different people.

After the residual risk score passes the success criterion to demonstrate quantitatively that the anonymization process reliably reduces risk to an acceptable level, we conduct a red-team exercise to stress test the resulting output.

4.5 Red-Teaming

The term "red team" originates in the military context: a red team is a group that assumes the role of an adversary, simulating attacks to identify vulnerabilities so that they can be resolved before a real attacker can exploit them. In the context of anonymization, this means "attacking" the de-identified output using common internet resources (e.g. search engines) and creative thinking to attempt to re-identify participants. "Success" of the exercise in our context — PI protection — means that the adversary is unable to identify an individual based on remaining unmasked information in the data set.

The red team for this exercise consisted of data engineers, applied scientists, computational linguists, privacy counsel, and a security advisor.

We sampled 200 conversations across different conversation contexts for the red-teaming practice. The conversations were anonymized using the modified DLP method described above.[12] Our red team found that of 200 full conversations, 181 calls (90.5%) were fully anonymized (no PI identified by the red team) and 19 calls (9.5%) showed some residual PI. However, the team determined that even with creative research and inference, none of the remaining 19 calls contained enough PI to successfully identify any individual, meaning that the data set could be safely used to train an LLM with no risk of exposing identifiable PI in later generative tasks. The failure of the red team to achieve its goal is a strong indication of the success of our anonymization methods.

With the proposed anonymization workflow successfully passing both quantitative and qualitative evaluation, we conducted LLM fine-tuning experiments to demonstrate the usability of the anonymized data for downstream tasks.

5 Privacy-Preserving LLM Training

Given a successfully anonymized data set, it can be used in combination with training prompts to

[11] Meetings, the main source of video conferencing call data, are typically longer and have more speakers than audio-only calls.

[12] The conversations considered by the red team were not annotated with *MISSED_* labels, because this annotation step was only used during the calibration and quantitative evaluation of the automated de-identification method.

fine-tune an LLM. As the training prompts contain no PI either, the combined fine-tuning data set contains no PI. If the model remains suitably performant, this demonstrates the ability to benefit from highly relevant domain-specific (i.e. real-world) training data while substantially reducing or even eliminating the risk of leakage or extraction.

5.1 Model

We used the "Chat" version of the LLaMA-2 model (Touvron et al., 2023) with 7B parameters as our base model.[13] LLaMA-2 is an open-source LLM developed by Meta. We chose LLaMA-2-7B as it showed comparable performance to larger models with a reduced cost of deployment. This base model was fine-tuned with a Text-to-Text Transfer Transformer (Raffel et al., 2020) on 59000 external samples and 13000 in-domain conversations. In the following, we refer to our fine-tuned LLM as DialpadGPT.

5.2 Experiment

After the model was fine-tuned, we sampled 400 LLM outputs across four downstream tasks of interest (100 outputs per task), involving both generation and classification:

- **Action Item:** Generate a description of a well-defined task to be completed after the call conversation.
- **Summarization:** Generate a summary of the conversation.
- **Call Purpose:** Classify the call into one of a pre-defined group of broad conversational themes, and the speaker intention and/or attitude.
- **Call Outcome:** Classify the call into one of a pre-defined group of categories that specify the result of the call e.g. complaint resolved, callback requested.

The four tasks were included in the fine-tuning process. All inputs provided to the model to generate output samples were anonymized using the process described above (the process used on the fine-tuning data set).

5.3 Results

Human annotators manually reviewed the outputs for each task and found no instances of PI in any of the output samples — that is, each piece of PI remained anonymized in the output.

Model performance on the aforementioned test set is shown in Table 5, which compares ROUGE-1 scores (Lin, 2004) and F_1 scores of DialpadGPT to the following commercial LLMs:

- **GPT-3.5:** GPT-3.5 is the model behind OpenAI's[14] ChatGPT (Laskar et al., 2023). We use the *gpt-3.5-turbo-0613* model, which has a maximum context length of 4096 tokens.
- **GPT-4:** GPT-4 is the latest LLM released by OpenAI (OpenAI, 2023), which has a maximum context length of 8192 tokens. In this experiment, we evaluate two versions of the model: GPT-4 (*gpt-4-0613*) and GPT-4 Turbo (*gpt-4-1106-preview*).
- **PaLM-2:** PaLM-2 (Anil et al., 2023) is an LLM developed by Google. It leverages the mixture of objectives technique (Anil et al., 2023) and significantly outperforms the original PaLM (Chowdhery et al., 2023) model. We use the *text-bison@001 model*, which has an input context window length of 8192 tokens.[15]

Across two generative tasks and two classification tasks, DialpadGPT, fine-tuned with anonymized data, outperforms all four popular commercial models.

6 Conclusion

In this paper, we presented a method for improving data anonymization on transcripts of business conversations using a publicly available service. We proposed a framework for quantitative and qualitative criteria for anonymization (residual risk scoring plus red team review), and showed that an LLM fine-tuned with data anonymized by the proposed workflow on relevant tasks has superior performance compared to commercially available LLMs. This shows LLMs are still able to understand and leverage contextual information without access to those key entities. In practice, we found that having some key entities like user or company names is helpful for some downstream tasks. In future work we will assess the performance of LLMs

[13] https://huggingface.co/meta-llama/Llama-2-7b

[14] https://platform.openai.com/docs/models/

[15] Available via Google's *VertexAI* platform. https://cloud.google.com/vertex-ai/docs/generative-ai/model-reference/text

Models/Tasks	Summarization ROUGE-1	Action Items ROUGE-1	Call Purpose F_1	Call Outcome F_1
DialpadGPT	**0.6096**	**0.5532**	**0.6562**	**0.738**
GPT-3.5	0.4957	0.3918	0.5078	0.6638
GPT-4	0.5783	0.5483	0.5508	0.6114
GPT-4 Turbo	0.5243	0.4143	0.6289	0.6812
PaLM-2	0.4832	0.4629	0.4492	0.4803

Table 5: Comparison between LLMs on downstream tasks of interest.

fine-tuned on an augmented anonymized data set, with names substituted by gender-neutral names and companies substituted by synthetic companies.

7 Limitations

One limitation of relying on a commercial system for data anonymization is that it is not always clear how to improve the process when unexpected results are obtained. With the improvements we made to the system, person and company name are still the infoTypes most likely to have false negatives, especially in lexically ambiguous cases like the name *Mark*[16] or with uncommon or unusually formatted company names.

In addition, the use of proprietary data for evaluating the results of fine-tuning LLMs renders direct comparison to other organizations' models challenging. Despite the low residual risk and resulting high confidence in the anonymization of the data sets, privacy best practices nonetheless caution against publishing our resulting data sets (Narayanan and Shmatikov, 2007). That being said, the overall methodology described here is certainly replicable, using a publicly available anonymization API, with task or domain-specific modifications to the PI taxonomy, and with the residual risk threshold tuned appropriate to the use case.

Finally, in our evaluation we mainly focused on the recall/hit-rate of the PI tagging. The precision/recall trade-off in machine learning suggests that an anonymization system with very high recall, i.e. poor precision, will lose context and generate data that cannot be used in LLM training. In practice, however, we did not observe such cases and our experiment showed that LLMs are still able to capture enough context after the anonymization.

[16]The person's name *Mark* and verb *mark* are often confused in the formatting process in terms of casing and thus fail to be identified as PI in the anonymization process.

References

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. 2023. Emergent and predictable memorization in large language models.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models. *ArXiv*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mark Elliot, Elaine Mackey, and Kieron O'Hara. 2020. *The anonymisation decision-making framework, 2nd Edition: European practitioners' guide*. UK Anonymisation Network.

Simson L. Garfinkey. 2015. De-identification of personal information. Technical report, National Institute of Standards and Technology.

Alejandro Guerra-Manzanares, Leopoldo Julian Lechuga Lopez, Michail Maniatakos, and Farah Shamout. 2023. Privacy-preserving machine learning for healthcare: open challenges and future perspectives. In *ICLR 2023 Workshop on Trustworthy Machine Learning for Healthcare*.

Valentin Hartmann, Anshuman Suri, Vincent Bindschaedler, David Evans, Shruti Tople, and Robert West. 2023. Sok: Memorization in general-purpose large language models.

Fadi Hassan, David Sánchez, and Josep Domingo-Ferrer. 2023. Utility-preserving privacy protection of textual documents via word embeddings. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):1058–1071.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Xiangji Huang. 2023. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. *arXiv preprint arXiv:2305.18486*.

Haoran Li, Yulin Chen, Jinglong Luo, Yan Kang, Xiaojin Zhang, Qi Hu, Chunkit Chan, and Yangqiu Song. 2023. Privacy in large language models: Attacks, defenses and future directions.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation models for text data: State of the art, challenges and future directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4188–4203, Online. Association for Computational Linguistics.

Bernd Malle, Peter Kieseberg, and Andreas Holzinger. 2017. Do not disturb? classifier behavior on perturbed datasets. In *Machine Learning and Knowledge Extraction - First IFIP TC 5, WG 8.4, 8.9, 12.9 International Cross-Domain Conference, CD-MAKE 2017, Reggio di Calabria, Italy, August 29 - September 1, 2017, Proceedings*, volume 10410 of *Lecture Notes in Computer Science*, pages 155–173. Springer.

Bernd Malle, Peter Kieseberg, Edgar R. Weippl, and Andreas Holzinger. 2016. The right to be forgotten: Towards machine learning on perturbed knowledge bases. In *Availability, Reliability, and Security in Information Systems - IFIP WG 8.4, 8.9, TC 5 International Cross-Domain Conference, CD-ARES 2016, and Workshop on Privacy Aware Machine Learning for Health Data Science, PAML 2016, Salzburg, Austria, August 31 - September 2, 2016, Proceedings*, volume 9817 of *Lecture Notes in Computer Science*, pages 251–266. Springer.

Arvind Narayanan and Vitaly Shmatikov. 2007. How to break anonymity of the netflix prize dataset.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models.

OpenAI. 2023. Gpt-4 technical report.

Richard Plant, Valerio Giuffrida, and Dimitra Gkatzia. 2022. You are what you write: Preserving privacy in the era of large language models.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Runhua Xu, Nathalie Baracaldo, and James Joshi. 2021. Privacy-preserving machine learning: Methods, challenges and directions.

Ying Yin and Ivan Habernal. 2022. Privacy-preserving models for legal natural language processing. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 172–183, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

A Residual Risk Categorization

Type	Missing Occurrence Tag	Google Tag	Score
Contact Info	(MISSED_EMAIL)	EMAIL_ADDRESS	4
	(MISSED_LOCATION)	LOCATION	2
	(MISSED_LOCATION_COORD)	LOCATION_COORDINATES	4
	(MISSED_US_STATE)	US_STATE	1
	(MISSED_PERSON_NAME)	PERSON_NAME	5
	(MISSED_PHONE)	PHONE_NUMBER	4
	(MISSED_ADDRESS)	STREET_ADDRESS	4
	(MISSED_USER_NAME)	USER_NAME	3
Entities	(MISSED_DOMAIN)	DOMAIN_NAME	1
	(MISSED_HTTP_COOKIE)	HTTP_COOKIE	1
	(MISSED_ORGANIZATION_NAME)	ORGANIZATION_NAME	0
	(MISSED_ORGANIZATION_NAME_SPEAKER)	ORGANIZATION_NAME	2
	(MISSED_PRODUCT)	PRODUCT	0
	(MISSED_PRODUCT_SPEAKER)	PRODUCT	2
	(MISSED_STORAGE_SIGNED_POLICY)	STORAGE_SIGNED_POLICY_DOCUMENT	2
	(MISSED_STORAGE_SIGNED_URL)	STORAGE_SIGNED_URL	3
	(MISSED_URL)	URL	2
Demographic	(MISSED_AGE)	AGE	1
Health Info	(MISSED_DATE_OF_BIRTH)	DATE_OF_BIRTH	3
	(MISSED_ICD9_CODE)	ICD9_CODE	2
	(MISSED_ICD10_CODE)	ICD10_CODE	2
	(MISSED_MEDICAL_RECORD_NUMBER)	MEDICAL_RECORD_NUMBER	5
	(MISSED_MEDICAL_TERM)	MEDICAL_TERM	1
ID number	(MISSED_ADVERTISING_ID)	ADVERTISING_ID	3
	(MISSED_GENERIC_ID)	GENERIC_ID	4
	(MISSED_ICCID_NUMBER)	ICCID_NUMBER	4
	(MISSED_IMEI_HARDWARE_ID)	IMEI_HARDWARE_ID	4
	(MISSED_IMSI_ID)	IMSI_ID	4
	(MISSED_IP_ADDRESS)	IP_ADDRESS	3
	(MISSED_MAC_ADDRESS)	MAC_ADDRESS	3
	(MISSED_MAC_ADDRESS_LOCAL)	MAC_ADDRESS_LOCAL	3
	(MISSED_PASSPORT)	PASSPORT	5
	(MISSED_VAT_NUMBER)	VAT_NUMBER	2
	(MISSED_VEHICLE_IDENTIFICATION_NUMBER)	VEHICLE_IDENTIFICATION_NUMBER	5
Payment Info	(MISSED_CREDIT_CARD_NUMBER)	CREDIT_CARD_NUMBER	5
	(MISSED_CREDIT_CARD_TRACK_NUMBER)	CREDIT_CARD_TRACK_NUMBER	5
	(MISSED_IBAN_CODE)	IBAN_CODE	5
	(MISSED_SWIFT_CODE)	SWIFT_CODE	1
	(MISSED_ROUTING_NUMBER)	ROUTING_NUMBER	3
	(MISSED_SSN)	SSN	5

Table 6: Residual risk scores assigned to infoTypes