# Generalizing Clinical De-identification Models by Privacy-safe Data Augmentation using GPT-4

**Woojin Kim**[1, 2] **Sungeun Hahm**[2] **Jaejin Lee**[2, 3]

[1]Korean National Police Agency, Seoul, Republic of Korea
[2]Dept. of Data Science, Seoul National University, Seoul, Republic of Korea
[3]Dept. of Computer Science and Engineering, Seoul National University, Seoul, Republic of Korea
{woojinkim118, isungeuni, jaejin}@snu.ac.kr

## Abstract

De-identification (de-ID) refers to removing the association between a set of identifying data and the data subject. In clinical data management, the de-ID of Protected Health Information (PHI) is critical for patient confidentiality. However, state-of-the-art de-ID models show poor generalization on a new dataset. This is due to the difficulty of retaining training corpora. Additionally, labeling standards and the formats of patient records vary across different institutions. Our study addresses these issues by exploiting GPT-4 for data augmentation through one-shot and zero-shot prompts. Our approach effectively circumvents the problem of PHI leakage, ensuring privacy by redacting PHI before processing. To evaluate the effectiveness of our proposal, we conduct cross-dataset testing. The experimental result demonstrates significant improvements across three types of F1 scores.

## 1 Introduction

In the realm of healthcare data management, the protection of sensitive information embedded within medical records is essential. Protected Health Information (PHI) encompassing identifiers, such as names, addresses, and contacts, requires safeguards to prevent unauthorized access and ensure patient confidentiality. The U.S. Health Insurance Portability and Accountability Act (HIPAA) lists a comprehensive set of 18 categories of PHI that must be de-identified in Electronic Health Records (EHRs) for secondary use.

The advent of automated systems for the de-identification (de-ID) of PHI within medical data has been facilitated by advances in rule-based methods, machine learning, and deep learning models (Liu et al., 2023b). For deep-learning-based approaches, a de-ID task is basically a sequence tagging task of named entity recognition (NER). Currently, i2b2 2006 (Uzuner et al., 2006, 2007) and i2b2 2014 de-ID challenge datasets (Stubbs et al., 2015; Stubbs and Uzuner, 2015) are widely-used for training de-ID models. These datasets contain clinical notes with PHI replaced with surrogate PHI for privacy and annotated with multi-classes of PHI labels, such as PATIENT, DOCTOR, etc. LSTM-based models (Dernoncourt et al., 2017) and BERT-based models (Alsentzer et al., 2019) have achieved micro F1 scores exceeding 0.97, demonstrating the effectiveness of deep-learning-based models to address the critical task of PHI de-ID in medical records.

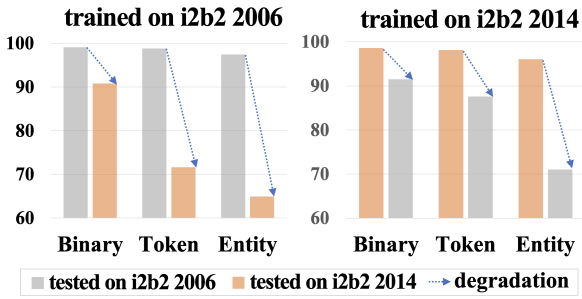


Figure 1: Performance degradation in cross-dataset settings using Bio-ClinicalBERT (Alsentzer et al., 2019). *Binary*, *token*, and *entity* refers to *binary token-level F1*, *token-level micro F1*, and *entity-level micro F1*. They are types of F1 scores used to evaluate sequence tagging of de-ID models. See Section 4.3 for the details regarding these metrics.

However, de-ID tasks present unique challenges compared to other NER tasks, primarily due to the difficulty in accessing suitable datasets. Moreover, making a publicly available dataset requires expertise and efforts to anonymize and annotate PHI (Yue et al., 2020). This leads to poor generalization of deep-learning-based models. While the i2b2 2006 and i2b2 2014 corpora offer clinical de-ID datasets both annotating PHI with surrogate PHI, the nuanced differences in annotation standards and the variability of institutes (e.g., hospitals) across datasets can significantly impact



*Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21204–21218
November 12-16, 2024 

cross-dataset test performance in cross-institute settings (Yang et al., 2019; Yue and Zhou, 2020).

Figure 1 shows the performance decrease in three types of F1 scores in a cross-dataset setting when training on the i2b2 2006 dataset and testing on the i2b2 2014 dataset, and *vice versa*. Data augmentation can be a solution to the problem of acquiring sensitive data and enhancing generalization, and some studies have explored this approach. Yue and Zhou (2020) propose rule-based EHR data augmentation with PHI replacement and context augmentation (i.e., synonym replacement). They improve token-level micro F1 on a cross-dataset setting of i2b2 datasets. However, rule-based augmentation lacks context diversity.

This paper proposes a data augmentation method using GPT-4 (Achiam et al., 2023) that can enrich the context of augmented datasets while ensuring privacy. As illustrated in Figure 2, our framework preemptively conducts PHI-scrubbing (i.e., replacing PHI with placeholders like "John" $\rightarrow$ "[PATIENT]," and "16th of November" $\rightarrow$ "[DATE]") before providing samples to GPT-4. This can effectively circumvent the problem of PHI leakage through APIs.

Our method includes zero-shot and one-shot prompt engineering for generating synthetic de-ID clinical corpus with i2b2 2006 and 2014 datasets. Zero-shot generation does not require any pre-obtained dataset. On the other hand, one-shot augmentation requires a pre-obtained dataset. With the one-shot augmented dataset, we conduct a cross-dataset test on i2b2 2006 and i2b2 2014 and achieve performance boosts up to 6.48%, 29.83%, and 36.50% across three types of F1 scores: binary token-level F1, token-level micro F1, and entity-level F1, respectively.

Overall, the contributions of this paper are summarized as follows:

- We propose a novel method for augmenting data for sequence tagging tasks using GPT-4 to generate diverse synthetic data even without any pre-obtained datasets.
- Our method overcomes the limitations of previous rule-based augmentation approaches in the clinical de-ID, which mainly focuses on entity replacements or minor textual modifications without significantly altering the context. By addressing data scarcity through augmentation, our method enhances generalization in cross-dataset settings.
- We propose a framework for handling sensitive data using online Large Language Models, which can be beneficial for applications in domains that manage sensitive information.

## 2 Related Work

This section briefly describes the related work to this paper.

### 2.1 Manual and Rule-based De-ID

Manual de-ID requires human experts or annotators who can access EHRs and redact sensitive information (Ubani et al., 2023). Manually redacting PHI is time-consuming and costly (Dorr et al., 2006; Douglass et al., 2005). Despite the high cost, existing reports demonstrate that manual removal of PHI is not that accurate, with recall ranging from 63% to 94% depending on the annotator (Neamatullah et al., 2008). Meanwhile, rule-based de-ID (Neamatullah et al., 2008; Meystre et al., 2010; Lison et al., 2021; Emelyanov, 2021) pre-defines regular expressions for word patterns and find the matching patterns in unstructured texts. Though it is faster than manual de-ID, it still needs fine-tuning the rules and patterns for each dataset (Ubani et al., 2023).

### 2.2 Learning-based De-ID

There are some prior deep-learning-based de-ID methods based on LSTM (Dernoncourt et al., 2017; Liu et al., 2017; Tang et al., 2019) and BERT (Alsentzer et al., 2019; Yue and Zhou, 2020). They fine-tune baseline models on PHI-labelled datasets, such as i2b2 2006 and i2b2 2014. Although these methods show promising performance on their test dataset, they show unstable performance when they are tested on the unseen test data in cross-institute settings (Stubbs et al., 2017; Yang et al., 2019; Yue and Zhou, 2020; Urbain et al., 2022). Other approaches include training a re-identification model in an adversarial setting and using it for de-identifying texts (Morris et al., 2022).

Recently, after the advent of large language models (LLMs), Liu et al. (2023b) employ ChatGPT (OpenAI, 2023a) and GPT-4 directly for de-identifying clinical data and show that the pre-trained LLMs can considerably well de-identify unseen texts. However, in real-life scenarios, this method poses a potential risk of breaching HIPAA privacy regulations, especially when using APIs

that transmit actual PHI-containing data to the client (Liu et al., 2023b,a). This highlights a critical challenge in balancing the innovative use of LLMs with the imperative of safeguarding sensitive information. Our work uses pretrained knowledge of GPT-4 to augment clinical data while maintaining privacy and safety.

### 2.3 Data Augmentation

Data augmentation (DA) includes methods of increasing the diversity of training data without gathering more data (Feng et al., 2021). DA can be performed at character, word, sentence, and document levels (Dai et al., 2023).

**DA in sequence tagging tasks.** Sequence tagging tasks like NER need DA for both entities and labels. Ding et al. (2020) train an LSTM-based model to learn linearized sequences of labeled sentences for generating augmented data. Dai and Adel (2020) modify sentence-level DA methods of synonym replacement for NER and show improvements in recurrent and transformer models. Yue and Zhou (2020) propose PHICON that conduct PHI and context augmentation on i2b2 2006 and i2b2 2014 datasets. The PHI augmentation replaces PHI entities in the original corpora with constructed surrogate-PHI candidate lists. Context augmentation of PHICON conducts synonym replacement and random insertion at word level (Wei and Zou, 2019). While there is a lack of previous studies employing synthetic data generation for NER tasks in the medical domain, PHICON uses rule-based augmentation on clinical notes and improves generalization performance. However, it still falls short during cross-testing. This shortfall is believed to be caused by its focus on entity replacements or minor textual modifications without significantly altering the context, making obtaining large and diverse corpora difficult.

**DA by prompting on LLMs.** Dai et al. (2023) propose few-shot prompting with ChatGPT for sentence classification datasets by rephrasing original sentences while preserving semantic consistency. Ubani et al. (2023) propose zero-shot prompting for DA, generating three text classification datasets without providing any sample to ChatGPT. These methods propose sentence-level augmentation on sentence classification datasets. They show that zero-shot and few-shot prompting-based DA outperforms existing rule- and model-based DA techniques. However, they focus on augmenting sentence classification datasets. Sequence tagging or token classification tasks involve a more complex dataset. To our knowledge, methods for augmenting sequence tagging tasks, such as NER, using prompting on LLMs have yet to be developed.

## 3 Proposed Methods

This section outlines our approach for enhancing de-ID sequence tagging datasets, such as i2b2 2006 and i2b2 2014, at the document level. We employ both one-shot and zero-shot prompting techniques with GPT-4-turbo (OpenAI, 2023b) via the OpenAI API. The distinction between zero-shot and one-shot prompts is based on their use of the original datasets; one-shot prompts incorporate sample data from these datasets to guide the model, whereas zero-shot prompts do not. One-shot is a type of few-shot that uses only one example to guide the model's response. In our case, we use the term one-shot as we provide one example (a clinical record) for each prompt. Figure 2 illustrates the overall pipeline of our one-shot augmentation. Excluding "i2b2 2006" in the figure also demonstrates the pipeline for our zero-shot augmentation. Our methodology uses the advanced NLP capabilities of GPT-4-turbo to augment the datasets, potentially improving the performance of models when fine-tuned on them for de-ID tasks.

### 3.1 One-shot Prompting

**PHI-scrubbing.** We use data samples from the original dataset for one-shot prompting. To prevent PHI leakage through API functions, we create a PHI-scrubbed corpus by replacing PHI in the original corpus with placeholders (e.g., "John" $\rightarrow$ "[PATIENT]" and "16th of November" $\rightarrow$ "[DATE]"). This process is designated by the blue arrow labeled *PHI-scrubbing* in Panel A of Figure 2.

**Guidelines and task descriptions.** The purple arrow labeled "Augmentation by GPT-4" in Panel A of Figure 2 indicates that the samples are provided to GPT-4 for dataset generation. The goal is to produce PHI-scrubbed patient reports that are similar in format to the original dataset but with different contexts. Prompts for one-shot augmentation include a sample, guidelines, and a task description. The task description instructs creating a synthetic patient report with PHI removed, following the specified guidelines. These guidelines, detailing the requirements for the desired output, emphasize the exclusion of actual PHI and instruct

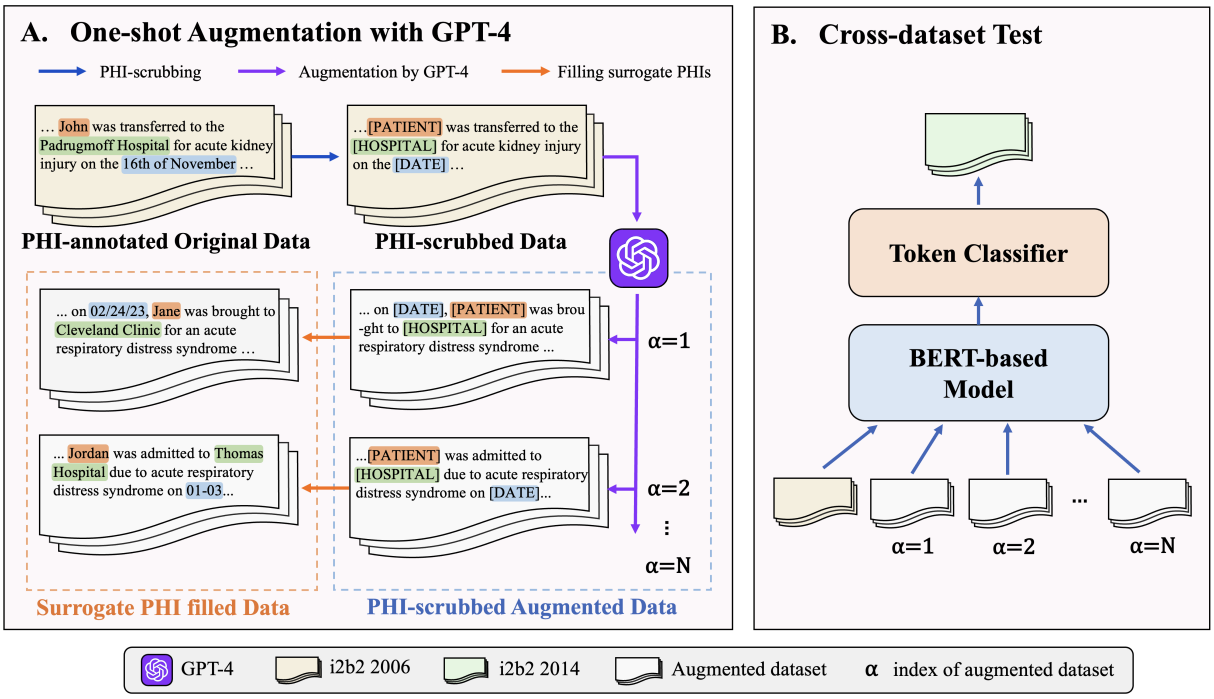


Figure 2: The process of one-shot data augmentation and cross-dataset test. Panel A outlines the augmentation process using ChatGPT, which begins with PHI-scrubbing the original data. It generates a new dataset using the training dataset of original data, repeating the augmentation process for $\alpha$ times to generate diverse datasets. $\alpha$ represents the sequential index of each augmented version of the original data. Panel B illustrates the framework for our cross-dataset test, where a BERT-based model uses the original and augmented datasets of i2b2 2006 to evaluate the model's cross-dataset performance on the i2b2 2014 dataset.

on using PHI placeholders, with explanations for each type. The guidelines are developed considering the annotation guidelines of i2b2 2014 by Stubbs and Uzuner (2015). With these elements, GPT-4 generates synthetic patient reports with PHI placeholders, referred to as "PHI-scrubbed Augmented Data" in Figure 2. See Appendix A for the detailed samples, guidelines, and task descriptions we used.

**PHI augmentation.** Since the generated data are PHI-scrubbed, we replace placeholders (e.g., "[PATIENT]" and "[DATE]") with appropriate surrogate PHI. Inspired by previous work (Yue and Zhou, 2020), surrogate PHI is randomly selected from candidate lists according to the placeholders, as shown in "Surrogate PHI filled data" in Panel A of Figure 2. The augmented datasets, along with the augmentation level $\alpha$ and the original data, are then used for cross-dataset testing, as shown in Panel B. See Appendix A for the detailed explanation of conducting PHI augmentation.

### 3.2 Zero-shot Prompting

Zero-shot prompting is tailored for scenarios where retaining any part of the training dataset for fine-tuning de-ID models is not feasible. As shown in Figure 4 of Appendix, zero-shot prompts consist only of guidelines and task descriptions without including data samples. The guidelines and task descriptions for zero-shot prompting mirror those used in one-shot prompting. With only the provided guidelines and task descriptions, GPT-4 generates synthetic patient reports with PHI replaced by placeholders. The process of PHI augmentation on the generated dataset follows the same procedure as that in one-shot augmentation.

## 4 Experimental Setup

This section describes our experimental methodology, which includes datasets, models, evaluation metrics, and testing methods.

### 4.1 Datasets

**Original datasets.** We utilize the i2b2 2006 and i2b2 2014 datasets, dividing them into training, validation, and testing sets in a 7:1:2 ratio based on the number of notes. The datasets adhere to different annotation standards. For instance, the i2b2 2014 datasets include 20 types of PHI labels, whereas the i2b2 2006 datasets feature only eight types. For the cross-dataset test, we follow the preprocessing steps outlined by Yue and Zhou (2020), which involve excluding PHI types that occur fewer than 20 times to eliminate low-frequency data and merging specific fine-grained PHI types into broader categories to ensure consistency across the datasets. This consolidation results in five main PHI categories: Name (encompassing Doctor, Patient, Username), Location (Hospital, Location, Zip, Organization), Date, ID (ID, Medical Record), and Contact (Phone).

**Augmented datasets.** By repeating the augmentation process from $\alpha = 1$ to 5, we develop one-shot augmented datasets derived from the i2b2 2006 and i2b2 2014 corpora, generating ten new datasets. Each augmented dataset is meticulously crafted to maintain the same number of patient records as its corresponding original dataset. In addition to these, we also produce five unique datasets using the zero-shot prompting technique by repeatedly performing the same augmentation. Each dataset from zero-shot prompting includes 912 records, aligning with the record count of the i2b2 2014 datasets. The preprocessing steps of augmented datasets are equivalent to those of original datasets. See Appendix B for the detailed statistics of the datasets.

| Hyperparameters | Value |
|---|---|
| Learning rate | 5e-5 |
| Batch size | 64 |
| Epochs | 20 |
| Seed | 1 |
| Maximum sequence length | 192 |

Table 1: Hyperparameters used in training the models.

### 4.2 Base Models and Hyperparameters

We select two BERT-based models: BioBERT (Lee et al., 2020) and Bio+ClinicalBERT (Alsentzer et al., 2019). They are pretrained with BERT (Devlin et al., 2018) on medical domains and show superior performance on clinical NLP tasks (Alsentzer et al., 2019; Turchin et al., 2023). Our study focuses on improving generalization performance through data augmentation. Thus, we conduct experiments using these models, demonstrating state-of-the-art performance on the i2b2 dataset. We train the models until the token-level micro F1 on the validation dataset saturates, employing early stopping with patience of 3 to prevent overfitting.

The hyperparameters to train the models on these datasets are shown in Table 1. Considering the practice of using sentence-level inputs and the need for reasonable comparisons, we set the maximum sequence length to 192 tokens. This was based on the standards for NER tasks set by BioBERT (Lee et al., 2020) as detailed at `https://github.com/dmis-lab/biobert-pytorch`, which served as a benchmark for our experiments. The hyperparameters of GPT-4 used to augment the datasets are explained in detail in Appendix C.

### 4.3 Evaluation Metrics

We assess our models' performance using three distinct metrics, each catering to different aspects of the classification task. These metrics include binary token-level F1, token-level micro F1, and entity-level micro F1. **Binary token-level F1** ("binary F1") is widely used to evaluate de-ID models (Dernoncourt et al., 2017) while handling multiple classes of PHI as a single class to assess model accuracy by classifying each token as part of a PHI entity or not. **Token-level micro F1** measures the model's overall performance on classifying each token across multiple PHI classes, aggregating precision and recall for all classes. **Entity-level micro F1** (or Exact F1) evaluates the model's accuracy in correctly identifying and classifying entire entities, including their exact spans. For example, if the model incorrectly identifies "John Doe" with the tags "B-Name, O" (where "B-Name" indicates the beginning of a name entity, and "O" signifies a non-entity), the entity-level F1 score for the "Name" category would be zero, highlighting the model's precision in delineating entire PHI entities.

### 4.4 Testing Methodology

**Cross-dataset test.** In the cross-dataset testing scenario, we train models using the i2b2 2006 dataset and assess their performance on the i2b2 2014 dataset ("2006 $\rightarrow$ 2014") and *vice versa*. For 2006 $\rightarrow$ 2014 at **one-shot augmentation** level $\alpha = N(\leq 5)$, we train, evaluate, and test the base

models as follows:

1. Merge original i2b2 2006 training set with augmented datasets from levels $\alpha = 1$ to $\alpha = N$ to obtain training dataset at augmentation level $\alpha = N$.
2. Throughout the training epochs, we continuously fine-tune our models. Evaluate the model on the original i2b2 2006 validation set.
3. Upon completion of training and fine-tuning, test the model on the comprehensive i2b2 2014 dataset (Train + Validation + Test).

For the reverse scenario, training on the i2b2 2014 and testing on the i2b2 2006 ("2014 $\rightarrow$ 2006"), the process above from step 1 to step 3 is applied analogously, merely replacing "2014" with "2006" and *vice versa* at each step. We compare the results to baselines where data augmentation is not applied. Instead, the rule-based augmentation "PHICON" — a rule-based augmentation method introduced by Yue and Zhou (2020) and applied to i2b2 datasets — is performed on the baselines.

**Performance test.** We use datasets generated through zero-shot augmentation in scenarios where no datasets can be obtained *a priori*. It involves exclusively training models on artificially generated data and testing their performance on i2b2 2006 and i2b2 2014 datasets. At **zero-shot augmentation** level $\alpha = N(\leq 5)$, we train, evaluate, and test the base models as follows:

1. Combine zero-shot augmented datasets from levels $\alpha = 1$ to $\alpha = N$ and divide it into training and validation datasets in an 85:15 ratio.
2. With the datasets prepared, models undergo training on the designated training set. Subsequent evaluation is conducted on the validation set derived from the same augmented pool.
3. Upon completion of training and fine-tuning, test the model on the comprehensive datasets of i2b2 2014 and i2b2 2006.

We do not compare the results to the rule-based data augmentation (i.e., PHICON) as a baseline in this scenario. For the one-shot prompting method, it is reasonable to compare because it generates clinical notes using samples from datasets obtained *a priori*. However, the zero-shot prompting method assumes scenarios with no available dataset, making it incomparable to PHICON.

## 5 Experimental Results

This section provides the experimental results for the proposed method to boost the generalization capability of the de-ID models.

### 5.1 One-shot Augmentation

The results of the cross-dataset tests demonstrate that datasets augmented through one-shot prompts across a range of augmentation levels from $\alpha = 1$ to 5 exhibit enhanced performance. The summarized outcomes, as presented in Table 2, alongside a comparative analysis with PHICON, highlight our one-shot augmentation's effectiveness.

A key observation is a more significant improvement in entity-level micro F1 score ("E") compared to the binary token-level and token-level micro F1 scores ("B" and "T"). Specifically, when compared to the no augmentation baseline ("No Aug"), one-shot augmentation results in maximum improvements of 6.48%, 29.83%, and 36.50%, and minimum improvements of 3.36%, 2.25%, and 17.69% for "B," "T," "E" scores, respectively. As entity-level F1 considers the exact span of each PHI entity, this highlights that our augmentation with GPT-4 boost models' preciseness in de-identifying PHI entities.

The standout performer in our analysis is BioBERT in the entity-level F1, particularly in "2006$\rightarrow$2014", which shows an improvement of 29.8% in token-level F1 compared to the baseline ("No Aug") and 10.5% improvement in entity-level F1 over PHICON. This trend is mirrored in the 2014 to 2006 setting, where the entity-level F1 of BioBERT incorporates the most substantial gains, with a 36.5% improvement over baseline and an 18.4% improvement over PHICON. The sudden performance improvement from $\alpha = 4$ to $\alpha = 5$ in "2014$\rightarrow$2006" appears to result from the model escaping local minima during the training process, despite not tuning hyperparameters and using the settings from the previous study (Lee et al., 2020).

Meanwhile, the impact is notably more substantial in the "2006$\rightarrow$2014" test scenario compared to "2014$\rightarrow$2006". This variation can be attributed to the different complexity levels of annotation standards between the datasets. The i2b2 2006 dataset has a more complex annotation standard than the i2b2 2014 dataset. For example, under the "DATE"

| Setting | Model | F1 | No Aug | +PHICON | + One-shot Augmentation | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | $\alpha = 1$ | $\alpha = 2$ | $\alpha = 3$ | $\alpha = 4$ | $\alpha = 5$ |
| 2006 → 2014 | Bio+Clinical BERT | B | 88.63 | 93.2 | 93.97 | 94.34 | **94.37** | 94.19 | 94.29 |
| | | T | 79.04 | 87.36 | 90.7 | 90.87 | 91.04 | 90.81 | **91.05** |
| | | E | 64.03 | 75.01 | 81.16 | 81.82 | 82.27 | 82.32 | **82.36** |
| | BioBERT | B | 90.84 | 93.03 | 93.89 | **94.27** | 93.6 | 94.01 | 94.19 |
| | | T | 70.19 | 85.63 | 90.33 | 91.07 | 90.29 | 91.10 | **91.13** |
| | | E | 66.79 | 74.45 | 80.69 | 81.86 | 80.86 | **82.29** | 81.70 |
| 2014 → 2006 | Bio+Clinical BERT | B | 90.23 | 92.78 | 93.15 | 92.90 | 92.87 | **93.26** | 93.06 |
| | | T | 86.51 | 87.85 | 88.31 | 88.04 | 87.85 | **88.46** | 87.98 |
| | | E | 64.38 | 72.73 | **75.77** | 74.24 | 74.41 | 75.45 | 73.82 |
| | BioBERT | B | 88.23 | 90.59 | 88.18 | 88.53 | 89.17 | 89.13 | **93.77** |
| | | T | 84.45 | 85.23 | 83.70 | 83.90 | 83.42 | 84.28 | **88.46** |
| | | E | 55.97 | 64.50 | 55.04 | 55.87 | 58.28 | 60.56 | **76.40** |

Table 2: Cross-dataset test results using One-shot augmented datasets. "B," "T," and "E" in the F1 column refer to binary token-level F1, token-level micro F1, and entity-level micro F1, respectively. The "No Aug" column shows the performance of the models in cross-dataset settings without any data augmentation. The "+PHICON" column represents the performance after conducting the PHICON (Yue and Zhou, 2020) data augmentation method. Finally, the "+ One-shot Augmentation" section evaluates the performance with incremental one-shot augmented datasets, where $\alpha \in \{1, 2, 3, 4, 5\}$ The bold and underlined results represent the highest and second-highest scores achieved for each model and setting combination.

| Test on | Model | F1 | Trained on | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Cross-dataset (06⇌14) | Zero-shot augmented dataset | | | | |
| | | | | $\alpha = 1$ | $\alpha = 2$ | $\alpha = 3$ | $\alpha = 4$ | $\alpha = 5$ |
| 2014 | Bio+Clinical BERT | B | 88.63 | 90.56 | 91.42 | **91.66** | 90.64 | 88.94 |
| | | T | 79.04 | 85.52 | 87.11 | **87.23** | 86.98 | 84.92 |
| | | E | 64.03 | 75.90 | 78.52 | **78.70** | 78.28 | 77.37 |
| | BioBERT | B | **90.84** | 87.17 | 87.72 | 87.21 | 89.33 | 84.79 |
| | | T | 70.19 | 82.35 | 83.03 | 82.97 | **85.52** | 82.12 |
| | | E | 66.79 | 72.24 | 73.87 | 74.89 | **76.71** | 73.05 |
| 2006 | Bio+Clinical BERT | B | 90.23 | 84.80 | 85.84 | 86.35 | **90.81** | 89.68 |
| | | T | **86.51** | 76.13 | 78.76 | 79.88 | 84.30 | 83.45 |
| | | E | 64.38 | 48.21 | 54.37 | 56.12 | 73.69 | **73.77** |
| | BioBERT | B | 88.23 | 84.34 | 84.64 | 84.52 | **90.32** | 90.18 |
| | | T | 84.45 | 76.93 | 77.52 | 78.23 | **84.79** | 83.34 |
| | | E | 55.97 | 48.60 | 49.87 | 51.27 | **75.05** | 72.77 |
| 2006 + 2014 | Bio+Clinical BERT | B | 89.43 | 87.68 | 88.63 | 89.01 | **90.73** | 89.31 |
| | | T | 82.78 | 80.83 | 82.94 | 83.56 | **85.64** | 84.19 |
| | | E | 64.21 | 62.06 | 66.45 | 67.41 | **75.99** | 75.57 |
| | BioBERT | B | 89.54 | 85.76 | 86.18 | 85.87 | **89.83** | 87.49 |
| | | T | 77.32 | 79.64 | 80.28 | 80.6 | **85.16** | 82.73 |
| | | E | 61.38 | 60.42 | 61.87 | 63.08 | **75.88** | 72.91 |

Table 3: Test results on i2b2 2006 and i2b2 2014 with training models only on zero-shot augmented datasets. "B," "T," and "E" in the F1 column refer to binary token-level F1, token-level micro F1, and entity-level micro F1, respectively. We compare the models with the baseline trained on cross-dataset (e.g., training on i2b2 2006 and testing on i2b2 2014). For test results on "2006 + 2014", we report arithmetic means of the test results on "2006" and "2014" as the number of test dataset samples differs. The bold and underlined results represent the highest and second-highest scores achieved for each model and test dataset combination.

category, i2b2 2014 encompasses both year and date, whereas i2b2 2006 annotations include dates but exclude the year. With the 2006 dataset's annotation standards being more intricate and the 2014 dataset providing a larger volume of patient reports, the effectiveness of data augmentation becomes more pronounced, especially in scenarios where training data is limited (Wei and Zou, 2019; Yue and Zhou, 2020).

### 5.2 Zero-shot Augmentation

In the context of zero-shot augmentation, where access to actual or surrogate PHI-containing clinical datasets is restricted, the potential of GPT-4 for zero-shot augmentation/generation is evident from the results in Table 3. Models trained on datasets augmented through zero-shot techniques showcase superior performance across the metrics compared to those trained on datasets obtained *a priori*. This underlines the ability of zero-shot augmentation with GPT-4 to generalize effectively across unseen datasets, a significant finding given that the approach does not require any synthetic sample.

When tested on the 2014 dataset, all metrics, except the binary token-level F1 in BioBERT, showcase improvements, reaching up to 22.9% increase in performance for the entity-level F1 of Bio+Clinical BERT. Similarly, when tested on the 2006 dataset, our method outperforms the baseline (trained on cross-dataset) in nearly all cases, with the BioBERT model's entity-level metric experiencing the highest leap at 34%.

Overall, the aggregated performance metrics from the "2006 + 2014" dataset report an increase in all metrics compared to training on cross-dataset. Moreover, the results indicate an upward trend in performance with increasing augmentation levels $\alpha$, reaching peak performance at $\alpha = 4$. This demonstrates the strategic advantage of leveraging higher augmentation levels to achieve optimal model performance.

### 5.3 Influence of Augmentation Level $\alpha$

At $\alpha = 5$, the model uses five times larger data than $\alpha = 1$. In the one-shot augmentation setting, saturation of performance is observed. In the zero-shot setting, the performance gradually improves but starts to degrade after level 4. In zero-shot augmentation, the approach leverages GPT-4's general knowledge and adds task-specific insights through prompts for clinical de-ID. This differs from one-shot augmentation, which additionally uses synthetic examples from the dataset. Increasing the augmentation level in zero-shot might not necessarily generate meaningfully diverse data because of the constraints of the LLM's pre-existing knowledge. Thus, the model is more prone to overfit and experiences a drop in test performance in zero-shot augmentation. Although one-shot augmentation does not show significant degradation, this should not lead to an overly optimistic conclusion that performance will continue to increase with repeated augmentation. It is necessary to explore an appropriate range for the augmentation level to understand its impact comprehensively and to identify the optimal point.

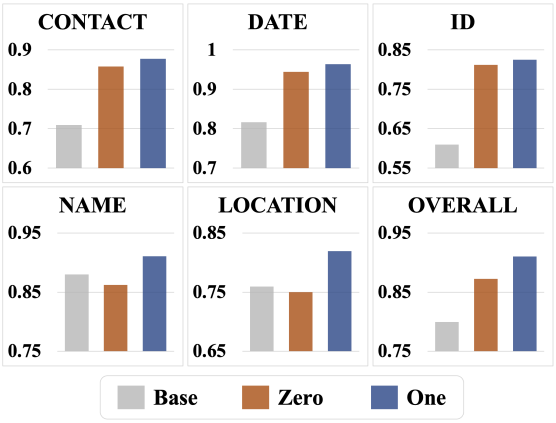


Figure 3: Performance of Bio+ClinicalBERT on i2b2 2014 with different datasets. *Base*, *Zero*, and *One* refer to the model trained on *i2b2 2006* dataset, *zero-shot augmented* dataset, and *i2b2 2006 + one-shot augmented* dataset, respectively.

### 5.4 Performance Boost Across PHI Classes

Figure 3 reports token-level F1 scores across PHI classes. Specifically, in micro-averaged F1 scores ("Overall"), performance on one-shot augmentation is the highest, followed by zero-shot and baseline. The model trained with the one-shot augmented dataset performs best in all classes. Additionally, in "Contact," "Date," and "ID," the zero-shot augmentation is significantly better than the baseline. However, in the "Location" and "Name" classes, the zero-shot performance is slightly worse than the baseline. This indicates that the model trained on a zero-shot augmented dataset may struggle with domain-specific unseen datasets. In synthetic patient reports, the distribution of names and locations may differ with general domains. For instance, names in patient records contain initialized names (i.e., usernames), and locations mainly con-

| Test on | Model | F1 | Trained on | |
|---|---|---|---|---|
| | | | Within-dataset | Cross-dataset (aug) |
| 2014 | Bio+Clinical BERT | B | 98.71 | 94.37 |
| | | T | 98.31 | 91.05 |
| | | E | 96.46 | 82.36 |
| | BioBERT | B | 98.62 | 94.27 |
| | | T | 98.20 | 91.13 |
| | | E | 95.99 | 82.29 |
| 2006 | Bio+Clinical BERT | B | 99.32 | 93.26 |
| | | T | 98.94 | 88.46 |
| | | E | 97.72 | 75.77 |
| | BioBERT | B | 99.16 | 93.77 |
| | | T | 98.71 | 88.46 |
| | | E | 97.65 | 76.40 |

Table 4: Comparison of within-dataset and cross-dataset performance. "B," "T," and "E" in the F1 column refer to binary token-level F1, token-level micro F1, and entity-level micro F1, respectively.

tain addresses related to hospitals.

### 5.5 Within-Dataset Performance

While our proposed methods significantly improve generalization performance, a notable gap still remains when comparing the performance to the within-dataset performance, where models are trained and tested on the same dataset. As shown in Table 4, although the within-dataset results show consistently higher F1 scores for all metrics, the cross-dataset performance is significantly improved through one-shot augmentation, highlighting the challenge of generalizing across different institutions where data formats and annotation standards vary. This gap emphasizes the importance of addressing the generalization problem, as models must perform robustly on unseen datasets from various sources in real-world applications.

## 6 Conclusion

This paper explores and analyzes the potential of one-shot and zero-shot data augmentation for de-identification in cross-dataset testing and extends the result to models pretrained with medical data. The proposed methods leverage GPT-4 to significantly enhance data scalability and diversity, which is particularly vital in healthcare domains where publicly available data is often limited. Moreover, the proposed methods of employing synthetic data generation with PHI-scrubbed datasets offer a strategic solution to addressing privacy concerns for GPT-4 APIs, aligning with HIPAA regulations. Overall, our methodology not only enriches the dataset quality for more effective model training but also aligns with ethical standards and privacy regulations, highlighting its potential to transform the data preparation process in sensitive domains.

## Acknowledgements

We thank the anonymous reviewers and meta-reviewer for their valuable feedback on this paper. We also thank Sangsoo Im, dispatched from the Republic of Korea Army, for his contribution to the experimental process.

Woojin Kim was seconded by the Korean National Police Agency to the Department of Data Science at Seoul National University as part of his master's program, which concluded in February 2024. This work was conducted during his secondment using the facilities provided by Seoul National University.

This work was supported in part by the National Research Foundation of Korea (NRF) grant (No. RS-2023-00222663, Center for Optimizing Hyperscale AI Models and Platforms), by the Institute for Information and Communications Technology Promotion (IITP) grant (No. 2018-0-00581, CUDA Programming Environment for FPGA Clusters), by the BK21 Plus programs for BK21 FOUR Intelligence Computing (Dept. of Computer Science and Engineering, SNU, No. 4199990214639) and Innovative Data Science Talent Education Program (Dept. of Data Science, SNU, No. 5199990914569) through NRF, all funded by the Ministry of Science and ICT (MSIT) of Korea. ICT at Seoul National University provided research facilities for this study.

## Limitations

There are some limitations to our study. First, domain-specific upgrade is required for the module that fills surrogate PHI in PHI-scrubbed data obtained from GPT-4. Upon reviewing performance across PHI classes, the observation that zero-shot performance in certain classes (name, location) is slightly lower than the baseline underscores the importance of domain-specific surrogate PHI. Constructing a universally applicable surrogate-PHI candidate list is deemed a crucial area for subsequent research, especially for practical applications in the healthcare domain.

Second, our study identified the optimal model using a validation set from the same source as the training data. However, further research is needed

to optimize performance for cross-dataset tests, which involve test data compiled from entirely different sources.

Third, while our model evaluates the validity of the augmented dataset through post-training performance comparisons, it is essential to investigate the augmentation's effectiveness further by comparing it with actual medical data, including word distribution. This step will help to ensure that the augmentation process accurately reflects the complexities of medical datasets.

Finally, our study uses GPT-4 to augment data and improve the generalization performance of state-of-the-art models for de-identification tasks, particularly in scenarios where the available dataset is extremely limited or nonexistent. Our decision to benchmark against PHICON was driven by the lack of previous studies that have employed synthetic data generation for clinical de-ID. However, Comparing against additional baselines, such as previously studied rule-based de-identification methods or fine-tuning offline open-source LLMs directly for the task in a manner that does not breach privacy concerns, could provide richer insights.

## Ethics Statement

Our study advances NLP while adhering to ethical guidelines. We use the i2b2 2006 and i2b2 2014 datasets, which contain de-identified Electronic Health Records (EHRs) accessible upon institutional consent[1]. We comply with data use policies and maintain transparency by disclosing our data sources, preprocessing, augmentation methods, and experimental settings.

Additionally, we prioritize maintaining ethical standards by ensuring that all personal identifiers are removed before further use. Our data augmentation includes scrubbing personal health information (PHI) to ensure privacy when using actual medical data with APIs.

## References


Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, et al. 2023. Chataug: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867.

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. Neuroner: an easy-to-use program for named-entity recognition based on neural networks. *arXiv preprint arXiv:1705.05487*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. Daga: Data augmentation with a generation approach for low-resource tagging tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6045–6057.

David A Dorr, WF Phillips, Shobha Phansalkar, Shannon A Sims, and John Franklin Hurdle. 2006. Assessing the difficulty and time cost of de-identification in clinical narratives. *Methods of information in medicine*, 45(03):246–252.

MM Douglass, GD Cliffford, Andrew Reisner, WJ Long, GB Moody, and RG Mark. 2005. De-identification algorithm for free-text nursing notes. In *Computers in Cardiology, 2005*, pages 331–334. IEEE.

Yaroslav Emelyanov. 2021. Towards task-agnostic privacy-and utility-preserving models. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 394–401.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.


[1] https://portal.dbmi.hms.harvard.edu

Pierre Lison, Ildikó Pilán, David Sánchez, 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.

Jialin Liu, Changyu Wang, and Siru Liu. 2023a. Utility of chatgpt in clinical practice. *Journal of Medical Internet Research*, 25:e48568.

Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. 2017. De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of biomedical informatics*, 75:S34–S42.

Zhengliang Liu, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Wei Liu, Dinggang Shen, Quanzheng Li, et al. 2023b. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*.

Stephane M Meystre, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC medical research methodology*, 10(1):1–16.

John Morris, Justin Chiu, Ramin Zabih, and Alexander M Rush. 2022. Unsupervised text deidentification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4777–4788.

Ishna Neamatullah, Margaret M Douglass, Li-Wei H Lehman, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. 2008. Automated de-identification of free-text medical records. *BMC medical informatics and decision making*, 8(1):1–17.

OpenAI. 2023a. Gpt-3.5-turbo. https://platform.openai.com/docs/models/gpt-3.5-turbo.

OpenAI. 2023b. Gpt-4-turbo. https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo.

Amber Stubbs, Michele Filannino, and Özlem Uzuner. 2017. De-identification of psychiatric intake records: Overview of 2016 cegs n-grid shared tasks track 1. *Journal of biomedical informatics*, 75:S4–S18.

Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. 2015. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1. *Journal of biomedical informatics*, 58:S11–S19.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus. *Journal of biomedical informatics*, 58:S20–S29.

Buzhou Tang, Dehuan Jiang, Qingcai Chen, Xiaolong Wang, Jun Yan, and Ying Shen. 2019. De-identification of clinical text via bi-lstm-crf with neural language models. In *AMIA Annual Symposium Proceedings*, volume 2019, page 857. American Medical Informatics Association.

Alexander Turchin, Stanislav Masharsky, and Marinka Zitnik. 2023. Comparison of bert implementations for natural language processing of narrative medical documents. *Informatics in Medicine Unlocked*, 36:101139.

Solomon Ubani, Suleyman Olcay Polat, and Rodney Nielsen. 2023. Zeroshotdataaug: Generating and augmenting training data with chatgpt. *arXiv preprint arXiv:2304.14334*.

Jay Urbain, George Kowalski, Kristen Osinski, Robert Spaniol, Mei Liu, Bradley Taylor, Lemuel R Waitman, et al. 2022. Natural language processing for enterprise-scale de-identification of protected health information in clinical notes. In *AMIA Annual Symposium Proceedings*, volume 2022, page 92. American Medical Informatics Association.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563.

Ozlem Uzuner, Peter Szolovits, and Isaac Kohane. 2006. i2b2 workshop on natural language processing challenges for clinical records. In *Proceedings of the Fall Symposium of the American Medical Informatics Association*. Citeseer.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Xi Yang, Tianchen Lyu, Qian Li, Chih-Yin Lee, Jiang Bian, William R Hogan, and Yonghui Wu. 2019. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC medical informatics and decision making*, 19(5):1–9.

Xiang Yue, Bernal Jimenez Gutierrez, and Huan Sun. 2020. Clinical reading comprehension: a thorough analysis of the emrqa dataset. *arXiv preprint arXiv:2005.00574*.

Xiang Yue and Shuang Zhou. 2020. Phicon: improving generalization of clinical text de-identification models via data augmentation. *arXiv preprint arXiv:2010.05143*.

## A Supplements for Prompts

As shown in Figure 4, prompts for one-shot augmentation include a sample, guidelines, and a task description. In contrast, prompts for zero-shot augmentation do not include samples.

### A.1 Samples

For the one-shot prompts in our study, each prompt incorporates a single sample from the pre-obtained data. This approach ensures that every sample within the dataset is utilized in generating augmented data, maintaining a 1:1 ratio between the original and newly created samples. Specifically, when augmenting the i2b2 2006 dataset once ($\alpha = 1$), each of the 622 samples from the i2b2 2006 training dataset is sequentially used in API calls, resulting in the creation of 622 new samples. This method allows for the comprehensive augmentation of the dataset, leveraging every available sample to enrich the training data with diverse, synthetic instances.

The i2b2 2006 and i2b2 2014 datasets used in this study are only authorized for research purposes, and the distribution of these datasets is strictly prohibited[2]. Therefore, it is not permissible to present the data samples used in the prompts directly within this paper. Instead, only a generalized and brief example can be provided as illustrated below. To prevent data leakage, we provided GPT-4 with PHI-scrubbed patient notes as samples when prompting. A sample of patient notes included in the one-shot prompts is as follows:

[2]https://portal.dbmi.hms.harvard.edu

Hospital: [HOSPITAL]
ID: [ID]
DATE: [DATE]
Discharge Summary: . . . [PATIENT] was transferred to [HOSPITAL] due to acute kidney injury. . .
Return Appointment: [DATE]
Electronically signed by: [DOCTOR] [DATE]
[report end]

### A.2 Guidelines

Specific guidelines included in the prompts are as follows:

1. PHI (Personal Health Information) should be removed and annotated with PHI labels.
2. Types of PHI labels and explanation:
{
"[AGE]": "Placeholder for annotate all ages, not just those over 90, including those for patient's families if they are mentioned",
"[DATE]": "Placeholder for any calendar date, including years, seasons, months, and holidays except time of day",
"[DOCTOR]": "placeholder for specific doctor names. Titles (Dr., Mr., Ms., etc.) do not have to be annotated. Information such as 'M.D.', 'R.N.' do not have to be annotated. If a name is possessive (e.g., Sam's) do not annotate the 's",
"[HOSPITAL]": "placeholder for the names of medical organizations and of nursing homes where patients are treated and may also reside. It cludes room numbers of patients, and buildings and floors

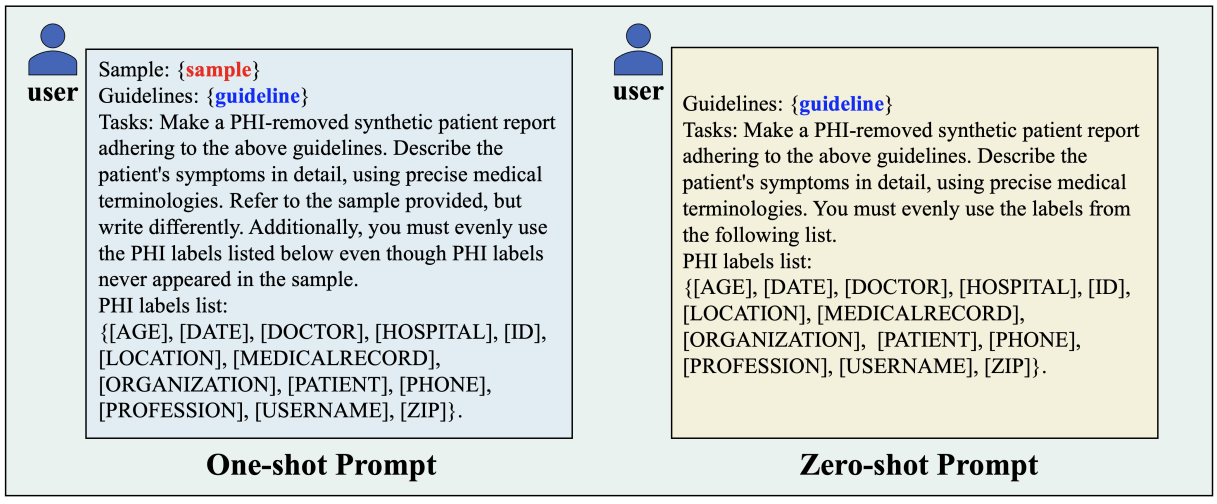


Figure 4: One-shot and zero-shot prompts.

related to doctors' affiliations. e.g, The patient was transferred to [Gates 4]",
"[ID]": "Placeholder for any identity (e.g., social security number, health plan number, account number, license number, vehicle ID, device ID, biometric ID, ID number) of an individual",
"[LOCATION]": "Placeholder for geographic locations such as cities, states, street names, building names,"
"[MEDICALRECORD]": "Placeholder for a medical record number",
"[ORGANIZATION]": "placeholder for specific named organizations",
"[PATIENT]": "placeholder for specific names of a patient. Titles (Dr., Mr., Ms., etc.) do not have to be annotated. Information such as 'M.D.', 'R.N.' do not have to be annotated. If a name is possessive (e.g., Sam's) do not annotate the 's",
"[PHONE]": "Placeholder for a specific phone number",
"[PROFESSION]": "Placeholder for any job that is mentioned that is not held by someone on the medical staff should be tagged",
"[USERNAME]": "Placeholder for a specific username that are initials followed by numbers (i.e., as4)",
"[ZIP]": "Placeholder for a specific zip code",
}
3. Only use PHI labels as described in section 2 and do not use any labels that are not included in this list.
4. Use all of PHI labels in section 2.

**Details of Guidelines.** Guideline 1 and 2 are developed concerning the annotation guidelines of i2b2 2014 in Stubbs and Uzuner (2015). Guideline 3 and 4 are empirically added. We found that without guideline 3, the model often generates data using placeholders that are not on the list. Without guideline 4, the model consistently uses only specific placeholders. This suggests that GPT-4, based on its pre-trained knowledge, tends to focus on commonly included personal information types such as patient names, locations, and phone numbers. As a result, less common types of personal information in the medical domain, such as medical record numbers and usernames, are not used.

### A.3 Task Descriptions

Specific task descriptions included in the prompts are as follows:

Make a PHI-removed synthetic patient report adhering to the above guidelines. Describe the patient's symptoms in detail, using precise medical terminologies. You must evenly use the labels from the following list.
PHI labels list:
{[AGE], [DATE], [DOCTOR], [HOSPITAL], [ID], [LOCATION], [MEDICALRECORD], [ORGANIZATION], [PATIENT], [PHONE], [PROFESSION], [USERNAME], [ZIP]}.

**Details of Task Descrptions.** The directive to 'evenly use' PHI labels is implemented because without this instruction, we empirically found that some placeholders are not used at all. Moreover, to improve generalization to data from different sources, it was necessary to balance the distribution of unbalanced classes. As shown in Figure 5, the distribution of PHI classes between the 2006 and 2014 datasets differs significantly. While the generated data has a class distribution similar to that of the real data, it helps to alleviate the imbalance.

### A.4 PHI Augmentation

Since the generated data using the prompts are PHI-scrubbed, we replace placeholders (e.g., [PATIENT], [DATE]) with appropriate surrogate PHI. We implemented PHI-augmentation as outlined by PHICON (Yue and Zhou, 2020). The augmentation process was divided into two main strategies based on the nature of the PHI labels.

- Systematically generated labels: For PHI labels such as ID, ZIP, and DATE, which follow specific patterns, we utilized regular expressions to generate the data.

- Non-systematically generated labels: For labels such as Organization, Hospital, Location, Patient, and Doctor, which do not follow predictable patterns and thus cannot be systematically generated, lists of candidate PHI for each category were gathered from Wikipedia. Detailed information on sources can be found at the paper (Yue and Zhou, 2020).

The PHI-scrubbed patient reports generated by GPT-4 were then populated by replacing each PHI label placeholder with the corresponding data listed above (e.g., the placeholder [HOSPITAL] was replaced with 'Cleveland Clinic'). This process is illustrated on the left-hand side of Figure 2.

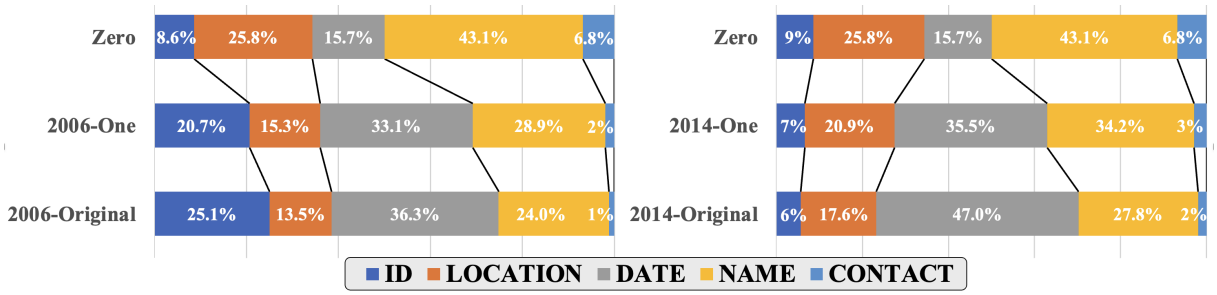


Figure 5: Ratio distribution of PHI classes across datasets based on entity counts.

## B Supplements for Dataset Statistics

We report statistics of train datasets after conducting preprocessing steps as outlined in subsection 4.1. Each dataset include sentences from patient notes that include PHI entities. PHI classes include **ID**, **LOCATION**, **DATE**, **NAME**, **CONTACT**. After tokenizing sentences using *spaCy*, tokens are labelled in BIO format. For instance, *B-NAME* indicates beginning token of a NAME entity whereas *I-NAME* indicates inside token. Tokens labelled with *O* signifies non-PHI tokens.

In the following tables, statistics illuminate the distribution and frequency of PHI classes within the original and augmented datasets. "Tokens" are counted as individual instances of PHI tokens, where each part of PHI like 'B-NAME' and 'I-NAME' is counted separately. "Entities" are counted as unique instances of PHI, where multiple tokens forming a single information unit are counted as one entity. Figure 5 shows ratio distribution of PHI classes across datasets.

### B.1 i2b2 2006 and One-shot Datasets

| **i2b2 2006** | **Original** | | **One-shot** | |
|---|---|---|---|---|
| #Notes | 622 | | 622 | |
| #Sentences | 5,502 | | 4,603 | |
| **PHI Class** | **Tokens** | **Entities** | **Tokens** | **Entities** |
| ID | 4,872 | 3,372 | 4,461 | 2,428 |
| LOCATION | 4,040 | 1,813 | 6,629 | 1,798 |
| DATE | 5,930 | 4,884 | 6,009 | 3,885 |
| NAME | 7,580 | 3,230 | 5,833 | 3,383 |
| CONTACT | 756 | 153 | 611 | 227 |
| Total ($T$) | 23,178 | 13,452 | 23,543 | 11,721 |
| $T$ / sentence | 4.21 | 2.44 | 5.11 | 2.55 |

Table 5: Statistics of the i2b2 2006 training set and the one-shot augmented training set derived from it. "One-shot" statistics represent the average values across five datasets, augmented along the iteration $\alpha = 5$.

### B.2 i2b2 2014 and One-shot Datasets

| **i2b2 2014** | **Original** | | **One-shot** | |
|---|---|---|---|---|
| #Notes | 912 | | 912 | |
| #Sentences | 8,419 | | 6,522 | |
| **PHI Class** | **Token** | **Entity** | **Token** | **Entity** |
| ID | 2,439 | 1,038 | 2,310 | 1,028 |
| LOCATION | 5,468 | 3,252 | 9,660 | 3,255 |
| DATE | 14,539 | 8,680 | 8,301 | 5,535 |
| NAME | 9,555 | 5,140 | 9,138 | 5,337 |
| CONTACT | 998 | 366 | 1,193 | 452 |
| Total ($T$) | 32,999 | 18,476 | 30,602 | 15,607 |
| $T$ / sentence | 3.92 | 2.19 | 4.69 | 2.39 |

Table 6: Statistics of the i2b2 2014 training set and the one-shot augmented training set derived from it. "One-shot" statistics represent the average values across five datasets, augmented along the iteration $\alpha = 5$.

### B.3 Zero-shot Augmented Datasets

| | **Zero-shot** | |
|---|---|---|
| #Notes | 912 | |
| #Sentences | 9,432 | |
| **PHI Class** | **Tokens** | **Entities** |
| ID | 3,848 | 1,761 |
| LOCATION | 16,101 | 5,262 |
| DATE | 4,940 | 3,201 |
| NAME | 15,895 | 8,815 |
| CONTACT | 3,815 | 1,391 |
| Total ($T$) | 44,599 | 20,430 |
| $T$ / sentence | 4.73 | 2.17 |

Table 7: Statistics of the zero-shot augmented training set. Statistics represent the average values across five datasets, augmented along the iteration $\alpha = 5$.

## C Hyperparameters for GPT-4

In our study, we use GPT-4 via its API, strictly adhering to the default settings provided to generate augmented data, without making specific adjustments to the hyperparameters. Details are as follows.

- Temperature: The default setting is used, typically aimed at fostering a balance between creativity and relevance, thus facilitating the generation of realistic yet diverse text outputs.
- Max Tokens: We adhere to the default limit to ensure that the length of outputs was reasonable and comparable to typical patient reports.
- Prompt: Although the content varied depending on the specific sample from the dataset being augmented, the structure of these prompts remained consistent, following the default guidelines for input.
- Stop Sequences: We employ the default settings, allowing the model to naturally conclude its text generation based on its internal algorithms and the content of the prompt.