Controlled Generation for Private Synthetic Text

Zihao Zhao
Johns Hopkins University
zzhao71@jhu.edu

Anjalie Field
Johns Hopkins University
anjalief@jhu.edu

Abstract

Text anonymization is essential for responsibly developing and deploying AI in high-stakes domains such as healthcare, social services, and law. In this work, we propose a novel methodology for privacy-preserving synthetic text generation that leverages the principles of de-identification and the Hiding In Plain Sight (HIPS) theory. Our approach introduces entity-aware control codes to guide controllable generation using either in-context learning (ICL) or prefix tuning. The ICL variant ensures privacy levels consistent with the underlying de-identification system, while the prefix tuning variant incorporates a custom masking strategy and loss function to support scalable, high-quality generation. Experiments on legal and clinical datasets demonstrate that our method achieves a strong balance between privacy protection and utility, offering a practical and effective solution for synthetic text generation in sensitive domains.[1]

1 Introduction

Text anonymization is an essential precursor to responsibly developing, deploying, and auditing AI in high stakes domains, like healthcare (Panchbhai and Pathak, 2022), social services (Gandhi et al., 2023), or law (Zhong et al., 2020), where sharing sensitive data with AI developers or using it to train models risks harmful leakage. Despite its importance, achieving effective anonymization in text is notoriously challenging. Tools for redacting directly identifying content, like names and addresses, have grown increasingly accurate, but they are unlikely to guarantee 100% recall, leaving "residual" identifiers easy to spot (Carrell et al., 2013). They also fail to remove vaguer quasi-identifying information that can still lead to re-identification, like a description of someone's appearance (Lison et al., 2021; Pilán et al., 2022).

Recently, synthetic text has become a potentially more robust alternative to long-standing redaction approaches. The idea involves leveraging Large Language Models' (LLMs) ability to generate highly fluent outputs to create text that is realistic enough to be useful but scrambles and abstracts away identifying or sensitive information. Models are typically fine-tuned over the real data and then prompted to generate new text. As LLMs are prone to memorizing and outputting training data (Carlini et al., 2021), prior work has conducted fine-tuning using differential privacy to prevent leakage of sensitive information (Yue et al., 2023; Kurakin et al., 2023; Mattern et al., 2022a; Putta et al., 2023). However, although language-model-generated data quality is generally high in the above methods, differentially private fine-tuning greatly degrades synthetic text quality (Ramesh et al., 2024). Furthermore, differential privacy guarantees are difficult to maintain in text settings where the unit of privacy is often unclear, and in practice, they do result in leakage of directly identifying information (Ramesh et al., 2024).

In this work, we propose a methodology for privacy-preserving synthetic data generation that aims to balance data utility with protection against the leakage of sensitive identifiers. Rather than relying on differential privacy, our approach draws on the long-established practice of de-identification and the theory of Hiding In Plain Sight (HIPS), which suggests that replacing detected identifiers with realistic surrogates can obscure the presence of any leaked real identifiers, making them more difficult to detect or exploit (Hirschman and Aberdeen, 2010; Carrell et al., 2013).

Our method begins by identifying private entities within the input text and representing them as control codes. These control codes are then used to guide controllable text generation (Keskar

[1]The code is provided in https://github.com/zzhao71/Controlled-Generation-for-Private-Synthetic-Text.git

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 32709–32724

et al., 2019), where falsified sensitive information is injected into the generation process to reduce the likelihood of the model reproducing real identifiers. We propose two variants of this approach: an in-context learning (ICL) method and a prefix tuning-based fine-tuning method.

In the ICL setting, the model is prompted with several example contexts, each consisting of a control code and its corresponding private text, followed by a fictional control code to guide the generation of the synthesized text. We further enhance privacy by directly blocking the output of sensitive tokens from the private examples, thus providing at least the same level of privacy protection as the underlying de-identification system (expecting small easy-to-filter errors due to case sensitivity). This characteristic makes the system suitable for deployment in regulated domains where legal compliance, such as with HIPAA, is essential.

In the prefix tuning variant, the model is fine-tuned using input-output pairs composed of control codes and their associated private texts. A fictional control code is again provided for the generation of synthetic text. This approach is further enhanced by our proposed privacy masking and custom loss function, which improve privacy protection and enable the generation of high-quality synthetic data. The overall method is illustrated in Figure 1.

We conduct experiments over data from two domains with sensitive information: law (Pilán et al., 2022) and healthcare (Johnson et al., 2016), demonstrating that our approach successfully produces synthesized text, prevents leakage of identifiers, and maintains text utility for downstream tasks. Unlike current paradigms that rely on DP-SGD for model fine-tuning (Yue et al., 2023), our novel method achieves a better balance between privacy protection and utility while being significantly more efficient and practical for real-world applications. This approach offers strong potential for advancing the responsible development and deployment of AI in contexts where both data utility and privacy preservation are critical considerations.

2 Methodology

Given a corpus of documents D_{real} containing sensitive information—such as court cases or clinical notes—our objective is to construct a synthetic dataset $D_{\text{synthetic}}$ that preserves the core content and stylistic characteristics of D_{real} while preventing the leakage of private information.

To achieve this, we employ *control codes* to guide the generation process toward producing fabricated sensitive information rather than reproducing real identifiers. In general, control codes have been used to guide models to produce outputs with desired properties (Keskar et al., 2019). Under this framework, both training and inference rely on the conditional probability of the output sequence $x = x_1, x_2, \ldots, x_n$ given the control code c, defined as:

$$P(x \mid c) = \prod_{i=1}^{n} P(x_i \mid x_{<i}, c) \tag{1}$$

We first define our notion of control codes and then how we use them to facilitate generation through either in-context learning or fine-tuning.

2.1 Control Codes

We define control codes as categories of private information (e.g., names, dates, or locations) and their associated values.

For a real document $d \in D_{real}$, which serves as an ICL example or a training data point, we extract all associated identifiers, such as names, ID numbers, or addresses—using either manual annotation or a de-identification model (Mendels et al., 2018). We group all identifiers by type to construct a control code of `<IDENTIFIER TYPE>: <REAL_VALUE>, <REAL_VALUE>, ...` pairs that contains all sensitive information in d. For example, $c =$

```
PERSON: Alice Jones, John Smith
LOC: New York City
...
```

Accordingly, every sensitive span in d is associated with one or more identifier types. The primary role of control codes is to help the model identify sensitive content and learn its relationship to the surrounding context.

At inference time, we construct fictional control codes—synthetic representations that follow the same format but contain falsified values randomly sampled from public lists. These fictional control codes direct the model to generate fake sensitive information over real information in accordance with the HIPS theory (Hirschman and Aberdeen, 2010; Carrell et al., 2013).

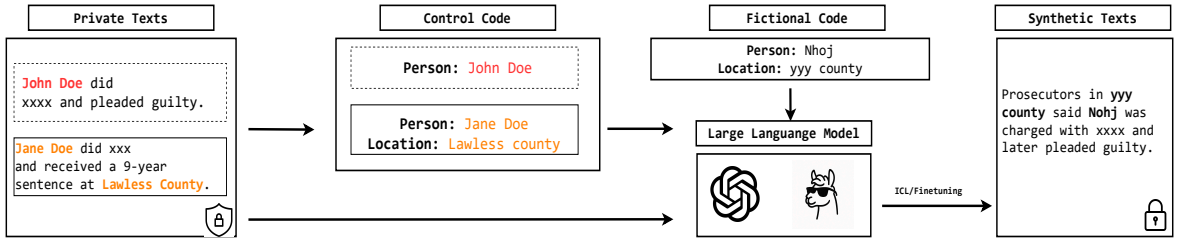

Figure 1: Overview of synthetic data generation

2.2 In-context learning (ICL)

In the ICL approach, we sample three distinct real documents $d_{1:3}$ from D_{real}. We construct associated control codes $c_{1:3}$, and a fictional code c_f. We then prompt an LLM with the three real pairs followed by the fictional code: $< c_1, d_1 >$, $< c_2, d_2 >$, $< c_3, d_3 >$, c_f. Thus, the LLM is prompted to generate a synthetic document consistent with the preceding examples but grounded only in the fictional control code, thereby avoiding the use of any real sensitive data. Examples of contexts in in-context learning generation are shown in Appendix C.

ICL w/ privacy enhancement In the ICL setup, the set of private information that the generation model is exposed to and could possible reproduce is limited to the small finite set of values in $c_{1:3}$. We explicitly designate these values as a set of "bad" tokens that the language model is prohibited from generating.[2] During decoding, these tokens are assigned low or zero probability, effectively preventing the model from selecting them in the output. This approach enforces that sensitive content is excluded from the model's output, ensuring that privacy preservation is at least as good as the hand annotations or deidentification model used to construct $c_{1:3}$.

2.3 Fine-tuning

In the fine-tuning setup, we train the LLM to generate a document from a provided control code. More specifically, from D_{real}, we construct training pairs where c_i is the input and d_i is the output and fine-tune the LLM over these pairs. We employ prefix-tuning (Li and Liang, 2021), which is much more computationally efficient than full fine-tuning and has been shown to successfully adapt models to similar tasks (He and Vechev, 2023). At inference time, fictional codes c_f are used as inputs to direct the model to output documents with fictional sensitive information.

Fine-tuning w/ Masking Inspired by He and Vechev (2023), we introduce a masked loss framework to enhance privacy protection during prefix tuning. We define a binary mask M to indicate the privacy status of each token in the training text. Specifically, private tokens are masked with $M = 0$, while non-private tokens are masked with $M = 1$.

Let P_θ denote the fine-tuned model's probability distribution. The standard language modeling loss $\mathcal{L}_{\text{LM}}$ is defined as:

$$\mathcal{L}_{\text{LM}} = -\sum_t \log P_\theta(y_t \mid y_{<t}) \tag{2}$$

where y_t is the ground truth token at position t, and $y_{<t}$ denotes all preceding tokens before position t.

To further enforce a divergence in behavior on private spans, we define a contrastive loss $\mathcal{L}_{\text{contrastive}}$, which encourages the fine-tuned model P_θ to deviate from the base model P_{base} specifically on private tokens:

$$-\sum_t (1 - M_t) \log \left(\frac{P_\theta(y_t|y_{<t})}{P_\theta(y_t|y_{<t}) + P_{\text{base}}(y_t|y_{<t})} \right) \tag{3}$$

This loss focuses only on private tokens ($M_t = 0$), promoting prediction divergence between P_θ and P_{base}.

In contrast, to preserve utility on non-private content, we introduce a KL divergence loss $\mathcal{L}_{\text{KL}}$, which encourages the fine-tuned model to remain close to the base model on non-private tokens:

$$\mathcal{L}_{\text{KL}} = \sum_t M_t \sum_{v \in V} P_{\text{base}}(v \mid y_{<t}) \log \frac{P_{\text{base}}(v|y_{<t})}{P_\theta(v|y_{<t})} \tag{4}$$

where V denotes the vocabulary and v represents a token in V.

[2] We use `AutoModelForCausalLM.generate(..., bad_words_ids=...)` to hard-block specific tokens at decoding time.

Finally, the overall training objective is a weighted combination of the three losses:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{LM}}\mathcal{L}_{\text{LM}} + \lambda_{\text{contrastive}}\mathcal{L}_{\text{contrastive}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} \quad (5)$$

The hyperparameters λ_{LM}, $\lambda_{\text{contrastive}}$, and λ_{KL} control the relative contributions of each component.

3 Experimental setup

3.1 Dataset

We conduct our experiments using the Text Anonymization Benchmark (TAB) introduced by Pilán et al. (2022), which comprises English-language court cases from the European Court of Human Rights (ECHR) and is manually annotated with personal identifiers. We use both the training (1,014 entries) and test (127 entries) splits. The training set is primarily used for synthetic data generation, while the test set is employed to evaluate the utility of models fine-tuned on the synthetic data.

Following the annotation scheme proposed by Pilán et al. (2022), we categorize information into three types: direct identifiers, quasi-identifiers, and non-sensitive information. Direct identifiers refer to information uniquely associated with an individual, whereas quasi-identifiers are publicly known attributes that may not independently lead to re-identification but could do so when combined with other such attributes. In this work, we define sensitive information as direct identifiers. Due to the model's input token limitations, we split each document into two segments: the first 6 paragraphs and paragraphs 6 to 12. The reported results represent the average scores across these two segments.

For experiments in the healthcare domain, we utilize the widely used dataset: MIMIC-III (Johnson et al., 2016; Goldberger et al., 2000). The MIMIC-III dataset contains over 2 million de-identified clinical notes associated with more than 40,000 patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts. To generate synthetic data from MIMIC-III, we use the string used for deidentification (e.g., [**Hospital 18**]) as control codes. The dataset provides diverse medical text for evaluating the effectiveness of our privacy-preserving synthetic generation methods in high-stakes domains.

3.2 Model

For both synthetic text generation and fine-tuning with synthetic data to evaluate utility metrics, we use the Sheared-LLaMA 1.3B model (Xia et al.). In the in-context learning setting, The generation parameters are set with a maximum of 400 new tokens, a temperature of 0.7, and a top-p value of 0.9. For prefix tuning, we use a default of 20 virtual tokens in the prefix. The model is trained for 3 epochs with a learning rate of 5×10^{-5}. The loss weighting coefficients are set as follows: $\lambda_{\text{LM}} = 1$, $\lambda_{\text{contrastive}} = 1$, and $\lambda_{\text{KL}} = 1$, assuming equal importance across all components.[3]

3.3 Control Code Construction

We evaluate models under two settings for constructing control codes. In the first, **private entity known**, we use the expert annotations of private entities provided in TAB, which are categorized into several types, including *person*, *code*, *location*, *organization*, *demographic*, *datetime*, *quantity*, and *miscellaneous*. A detailed description of these categories is provided in Appendix A.1. For the MIMIC dataset, which has already been anonymized, the anonymized terms will be recognized as control codes.

In the second, **private entity unknown**, we consider how our method might perform for anonymizing a dataset that has not already been annotated by experts for sensitive information, which is a more realistic scenario. In this case, we employ an existing entity recognition tool, Presidio (Mendels et al., 2018), to automatically identify potential private entities. We construct control codes using these inferred sensitive spans rather than the hand-annotated ones. In both settings, we use the hand-annotated sensitive spans in evaluation privacy leakage.

We also construct fictional control codes the same way in both settings. Fictional control codes are randomly generated, and the full details of the generation procedure are provided in Appendix A.2.

3.4 Privacy Metrics

We introduce several metrics to quantify privacy leakage.

Private Information Presence Percentage (PIPP) We define *Private Information Presence Percent-*

[3]We did not perform hyperparameter tuning; it is possible that further tuning could improve the method's performance.

age (PIPP) as the proportion of generated passages that contain at least one private entity. In the in-context learning setting, a privacy leak is recorded if any private term from the provided context appears in the generated output. In the prefix-tuning setting, a generation is considered a leak if any private entity from the entire fine-tuning dataset appears in the output. This metric quantifies the frequency of such leaks relative to the total number of generated samples.

Let $s_1, \ldots, s_n$ denote the n generated outputs. The PIPP is computed as:

$$\text{PIPP} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(s_i \text{ contains private information}) \tag{6}$$

where $\mathbb{I}(\cdot)$ is the indicator function, returning 1 if the condition is true and 0 otherwise.

Entity Leakage Percentage (ELP) The *Entity Leakage Percentage (ELP)* measures the proportion of private entities that appear in the model's generated output. In the ICL setting, we count how many private entities from the input context are present in the synthetic output, then divide this count by the total number of private entities in the context. We report the average value of this ratio across all examples.

In the fine-tuning setting, we first identify the total number of deduplicated private entities in the training set, as the model is exposed to all of these entities and could potentially leak them. We then collect a deduplicated list of leaked private entities found in the generated text. The ELP is computed as the ratio of the number of leaked (non-repetitive) entities to the total number of private entities.

Rouge Score While PIPP and ELP both measure leakage of identifiers, the goal of generating synthetic text instead of just redacting identifiers is for the synthetic text actually to be synthetic: e.g., it should contain new combinations of words and information that mask any residual identifiers, rather than directly copying content from model inputs. In order to evaluate if the generated text is sufficiently synthetic, we also report ROUGE-2 and ROUGE-L scores (Lin, 2004) to evaluate the similarity between the generated text and the reference text. The specific formulas for both ROUGE-2 and ROUGE-L scores are presented in Appendix B.

In the in-context learning setting, the reference text corresponds to the context provided to the model during generation. For the fine-tuning method, we generate each fictional control code for each fine-tuning sample sharing the same identifier type. To check if the output is copying from its fine-tuning data, the ROUGE score is then computed between the model's output and the corresponding text in the fine-tuning dataset, and takes the maximum value. Higher ROUGE scores in this context indicate a greater resemblance between the generated text and the original private content, and thus a higher risk of privacy leakage.

3.5 Utility Metrics

In addition to privacy protection, utility is also an important metric for measuring whether the content of sentences or paragraphs is preserved.

Perplexity After generating the synthetic text, we fine-tune a new instance of the Sheared LLaMA 1.3B model using the synthetic data as training input. To evaluate the utility of the synthetic data, we measure the model's perplexity on the test set (Jelinek, 1998). Perplexity is a commonly used metric in language modeling that quantifies how well a model predicts a sequence of tokens. It is computed as the exponential of the average negative log-likelihood of the true tokens under the model's predicted distribution. Lower perplexity indicates better predictive performance, suggesting that the synthetic data supports effective learning and retains high utility for downstream language modeling tasks.

MAUVE MAUVE (short for Model and Human Outputs Via Empirical divergence) is a metric designed to evaluate the distributional similarity between model-generated text and human-written reference text (Pillutla et al., 2021). Unlike token-level overlap metrics such as ROUGE or BLEU, MAUVE captures high-level properties of natural language, such as fluency, coherence, and diversity, by comparing the distributions of embeddings derived from pretrained language models.

To evaluate MAUVE in our setting, we use the fine-tuned model or the ICL model to generate synthetic outputs for each sample in the test set. Specifically, we input the corresponding control code from each test sample and collect the resulting generated texts. MAUVE is then computed between the set of model-generated outputs and the original texts from the test set. A higher MAUVE score indicates that the generated distribution more

Method	PIPP (%)↓	ELP (%) ↓	ROUGE-2 ↓	ROUGE-L ↓
Baseline	50.0 ± 2.20	30.89 ± 3.15	0.3898 ± 0.02	0.4303 ± 0.01
DP-SGD	1.56 ± 0.02	0.73 ± 0.05	0.0635 ± 0.002	0.1328 ± 0.01
ICL	3.25 ± 0.23	0.36 ± 0.012	0.2978 ± 0.02	0.3073 ± 0.008
ICL w/ privacy enhancement	**0.00 ± 0.00**	**0.00 ± 0.00**	0.3361 ± 0.04	0.3645 ± 0.005
Fine-tuning	3.94 ± 0.2	0.35 ± 0.07	0.017 ± 0.001	0.133 ± 0.004
Fine-tuning w/ masking	2.56 ± 0.89	0.39 ± 0.03	**0.0111 ± 0.003**	**0.0975 ± 0.002**

Table 1: Privacy protection performance under the private entity known setting on the TAB dataset

closely matches the distribution of human-written text, reflecting better language quality and realism.

4 Results

For all reported results with confidence intervals, we repeat each setting three times and compute the 95% confidence interval based on the observed variation across runs.

We include as a baseline performing generation with ICL, where we provide the model with three examples from the training set without using any control codes. We also include DP-SGD, which clips the gradients to limit the contribution of individual samples from the training data and subsequently adds noise from a predefined type of distribution to the sum of the clipped gradients across all samples, as a reference for comparison with our method (Ramesh et al., 2025).

4.1 Private Entity Known

The privacy metrics for all models in the private entity known setting are presented in Table 1. In the experiment, we set epsilon = 8.

All methods achieve much higher privacy protection than the baseline. As expected, `ICL w/ privacy enhancement` achieves better (lower) PIPP and ELP than the other models. While our measured leakage is 0% on average, we find that some leakage is possible in practice. LLaMA tokenizes words in a case-sensitive manner—e.g., "apple" and "Apple" are assigned different token representations. Although we construct a bad token list that accounts for multiple case variants (e.g., lowercase, uppercase), there remains a small chance of private entity leakage due to unaccounted casing variations or tokenization artifacts. Forcing leakage to zero is still possible by repeating each generation until no private entity is detected, and since the rate of leakage is effectively 0%, the expected number of regenerations needed is very small. However, in our experiments, for fair comparison with other methods, we report results from a single generation pass per sample.

While the ICL method supports direct prevention of leakage, leading to better PIPP and ELP metrics, the higher ROUGE scores as compared to the fine-tuning methods indicate that the model copies substantially more content from the input examples. For example, ROUGE-L for ICL is 0.3073 as compared to 0.133 for fine-tuning. This result suggests that the ICL setting is preferred when there is high confidence in the deidentification model and protecting information not specifically marked as identifiable is unimportant. In contrast, the fine-tuning approach offers a greater level of synthesis and thus is more capable of protecting information not explicitly marked as identifying but that may still be considered private.

We also perform evaluations on the MIMIC-III dataset, which follow a similar trend to those observed on the TAB dataset and are presented in Table 3.

4.2 Private entity unknown

In the entity unknown setting (Table 2), as expected privacy is generally worse than in the private entity known setting. For the ICL setting, the PIPP and ELP scores are lower than those in the private entity known setting, primarily because in some generations, the model either failed to produce any meaningful output or generated only a few terms, leading to lower leakage rates relative to the normal outputs. The ICL w/ privacy enhancement still achieves the best PIPP, but ELP is better for the prefix fine-tuning models. This result empirically demonstrates that the greater level of synthesis in the fine-tuning models can help correct for an imperfect deidentification model. Other trends are similar, with the prefix fine-tuning models having better (lower) ROUGE scores than the ICL methods.

Method	PIPP (%) ↓	ELP (%) ↓	ROUGE-2 ↓	ROUGE-L ↓
Baseline	50.0 ± 2.20	30.89 ± 3.15	0.3898 ± 0.02	0.4303 ± 0.01
ICL	0.89 ± 0.80	0.74 ± 0.13	0.4133 ±0.02	0.4877 ± 0.01
ICL w/ privacy enhancement	**0.77 ± 0.02**	0.67± 0.01	0.3968 ± 0.04	0.4478 ± 0.01
Fine-tuning	4.24 ± 0.18	0.28 ± 0.03	0.0235 ± 0.02	**0.0762± 0.01**
Fine-tuning w/ masking	2.84 ± 0.02	**0.11± 0.03**	**0.0098 ± 0.0003**	0.083 ± 0.01

Table 2: Privacy protection performance under the private entity unknown setting on the TAB dataset

Method	PIPP (%) ↓	ELP (%) ↓	ROUGE-2 ↓	ROUGE-L ↓
Baseline	22.3	4.7	0.4557	0.5321
ICL	5.6	1.8	0.3714	0.4147
ICL w/ privacy enhancement	**1.2**	**0.5**	0.3921	0.4263
Fine-tuning	9.7	2.3	0.1478	0.1526
Fine-tuning w/ masking	4.8	1.2	**0.0215**	**0.0874**

Table 3: Privacy protection performance under the private entity known setting on the MIMIC-III dataset. Lower scores indicate stronger privacy protection across entity-level and lexical similarity metrics.

4.3 Downstream Performance

To evaluate the utility of our method, we primarily report perplexity and MAUVE. As shown in Table 4, fine-tuning achieves better utility performance compared to the ICL method, with the fine-tuning w/ masking variant performing the best overall.

In terms of privacy protection, the ICL w/ privacy enhancement achieves the strongest results on certain metrics. However, the use of bad tokens in this setting may impair utility, particularly under the private entity unknown case, where the entity matcher may incorrectly identify common terms as private. Although the baseline appears to perform well in utility metrics, it offers poor privacy protection and is therefore not a viable privacy-preserving solution.

Overall, fine-tuning with masking strikes the best balance between utility and privacy, demonstrating strong performance across both evaluation criteria.

4.4 Ablation Study

4.4.1 Performance on larger model

All experiments described above were conducted using Sheared-LLaMA 1.3B. We extended our evaluation to Llama 3.1-8B Instruct, testing both ICL and ICL with privacy enhancement, shown in Table 5. The overall trends remain consistent: both settings show low risk of leakage, with the privacy enhancement setting providing greater protection. This demonstrates our method's robust performance across different model scales.

While PIPP and ELP metrics indicate slightly higher leakage risk in the larger model, we attribute this to the fact that TAB is not truly a private dataset. Larger LLMs exposed to TAB during pre-training are more likely to have memorized its content. Notably, Rouge-L scores improved compared to the smaller model, suggesting greater degrees of synthesizing at larger scales.

4.4.2 Performance on both direct and quasi identifiers

In the previous experiments, we considered only direct identifiers as private entities. We further evaluate the scenario where both direct and quasi-identifiers are treated as private. As shown in Table 6, performance generally declines due to the significant increase in the number of privacy terms. Nonetheless, the ICL w/ privacy enhancement setting continues to achieve the best results in the PIPP and ELP metrics, demonstrating the effectiveness of the bad token strategy in preventing privacy leakage.

For ROUGE-2 and ROUGE-L, the prefix tuning with masking variant remains the strongest performer. Compared to the setting where only direct identifiers are protected, its privacy protection performance shows only a modest decline, especially relative to other methods. This result highlights the method's robustness and its potential to scale effectively to scenarios involving a larger set of

Method	Private Entity Known		Private Entity Unknown	
	Perplexity ↓	MAUVE ↑	Perplexity ↓	MAUVE ↑
Baseline	11.7	0.78	**11.7**	**0.78**
DP-SGD	12.8	0.70	12.8	0.70
ICL	11.8	0.78	13.5	0.71
ICL w/ privacy enhancement	12.0	0.66	14.2	0.62
Fine-tuning	10.5	0.82	12.1	0.76
Fine-tuning w/ masking	**10.2**	**0.83**	**11.7**	**0.78**

Table 4: Utility performance under the private entity known and unknown settings on the TAB dataset

Method	PIPP (%) ↓	ELP (%) ↓	ROUGE-2 ↓	ROUGE-L ↓
ICL	6.51	2.50	0.2663	0.3201
ICL w/ privacy enhancement	3.55	0.82	0.2793	0.3341

Table 5: Privacy protection performance under the private entity known setting on the TAB dataset for a larger LLM (Llama 3.1-8B Instruct). Lower scores indicate stronger privacy protection across entity-level and lexical similarity metrics.

sensitive entities. The performance on pure quasi identifiers is shown in Appendix G.

5 Related Work

Synthetic data Synthetic data refers to the data that is generated artificially instead of collecting real-world events or annotations. In the context of AI, it specifically refers to the data that is generated by generative models to mimic the characteristics of real data (Liu et al., 2024; Saxton et al., 2019).

In NLP, LLMs have replaced more traditional methods for synthetic data generation, like synonym replacement, random insertion, and back-translation, and are emerging as a potentially viable alternative to human-generated data (Hartvigsen et al., 2022; Ye et al., 2022). Extensive pretraining allows LLMs to generated seeming fluent outputs (Ding et al., 2023), with a high level of controllability and adaptability, allowing researchers to create flexible datasets tailored to specific requirements (Eldan and Li, 2023).

To ensure privacy in synthetic data generation, differential privacy (DP) has emerged as a foundational framework (Dwork et al., 2006). Early efforts applied DP to generative models such as generative adversarial networks (GANs) and variational autoencoders (VAEs) by modifying training procedures using differentially private stochastic gradient descent (DP-SGD) (Xie et al., 2018; Chen et al., 2020; Abadi et al., 2016). More recently, DP has been extended to language models and text generation. For instance, Li et al., Yu et al. (2021), and Ramesh et al. (2024) fine-tune large language models under DP constraints by incorporating gradient clipping and noise injection to mitigate the memorization of sensitive content.

Our approach is complementary to fully differentially private training. Rather than enforcing global privacy guarantees, we focus on entity-aware generation through in-context learning and prefix tuning. This strategy reduces the risk of private information leakage by guiding the model to operate on synthetic or fictional identifiers, without the utility degradation imposed by DP.

Text Sanitization Text sanitization refers to the process of replacing sensitive tokens to protect privacy (Tong et al., 2025). A common approach involves first identifying text spans that contain personally identifiable information (PII) and then replacing them with a default placeholder, such as '***' (Olstad et al., 2023) or a black box (Lison et al., 2021; Pilán et al., 2022). However, this method often reduces the utility of the text. To mitigate this issue, alternative strategies have been developed that replace sensitive content with less risky alternatives, such as synonyms (Dalianis, 2019; Volodina et al., 2020) or synthetic surrogates (Carrell et al., 2013). In the medical domain, for example, patient names may be substituted with randomly selected names from a predefined list (Dalianis, 2019).

In addition, differential privacy (DP) has been

Method	PIPP (%) ↓	ELP (%) ↓	ROUGE-2↓	ROUGE-L ↓
Baseline	66.70 ± 5.22	48.2 ± 4.17	0.3898 ± 0.02	0.4303 ± 0.01
ICL	27.31 ± 6.18	3.20 ± 1.13	0.2669 ± 0.009	0.3034 ± 0.01
ICL w/ privacy enhancement	**0.89 ± 0.332**	**0.06 ± 0.027**	0.2624 ± 0.03	0.2988 ± 0.004
Fine-tuning	2.74 ± 1.91	0.64 ± 0.16	0.012 ± 0.0018	0.11 ± 0.08
Fine-tuning w/ masking	3.17 ± 0.99	0.33 ± 0.09	**0.009 ± 0.001**	**0.0977 ± 0.01**

Table 6: Privacy protection performance under the private entity known setting, where both direct and quasi-identifiers are treated as private entities, on the TAB dataset

applied to text sanitization at both the word level and sentence level. Word-level DP methods perturb individual words to achieve privacy guarantees (Feyisetan et al., 2020, 2019), while sentence-level approaches ensure privacy across entire documents (Igamberdiev and Habernal, 2023; Krishna et al., 2021; Mattern et al., 2022b; Utpala et al., 2023). These DP-based techniques offer formal privacy guarantees for each record in a dataset. However, such methods often come at the cost of degraded text quality and reduced utility. In contrast, our approach strikes a more favorable balance between privacy protection and text utility, generating high-quality synthetic text while mitigating privacy risks.

6 Conclusion

In our work, we proposed two approaches for generating privacy-preserving synthetic text that aims to balance privacy protection with downstream utility: an ICL method and a prefix tuning-based fine-tuning method. While the ICL method offers stronger privacy protection—particularly when combined with the privacy enhancement mechanism—it still exhibits notable lexical overlap with the original private text and has lower utility. In contrast, the prefix tuning with masking method achieves a better trade-off, demonstrating strong performance in both privacy and utility. Although it shows slightly higher entity-level leakage, it produces more diverse and less memorized outputs, which have high applicability in real-world scenarios.

Overall, both approaches demonstrate strong potential for use in synthetic text generation under privacy constraints. In future work, we plan to evaluate these methods on a broader range of datasets and explore additional hyperparameter settings to further improve performance.

7 Acknowledgments

We thank reviewers for their helpful feedback. This work was supported in part by the AI2050 Fellowship program by Schmidt Sciences.

8 Limitations

Our current study has several limitations. First, we evaluate our methods on only two datasets, which limits the generalizability of our findings. Future work should include a broader range of datasets across different domains and scales to assess performance more comprehensively. Second, we use fixed hyperparameters in our experiments. A more extensive hyperparameter search—including variations in the number of in-context examples and the loss weighting coefficients λ_{LM}, $\lambda_{\text{contrastive}}$, and λ_{KL}—could further improve performance. We also fix decoding parameters such as top-p and temperature, which may not be optimal across all settings. Finally, all experiments are conducted using the Sheared LLaMA 1.3B model. Evaluating our methods on larger range of language models would help determine how performance varies and whether models exhibit different privacy-utility trade-offs.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650.

David Carrell, Bradley Malin, John Aberdeen, Samuel Bayer, Cheryl Clark, Ben Wellner, and Lynette

Hirschman. 2013. Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. *Journal of the American Medical Informatics Association*, 20(2):342–348.

Dingfan Chen, Tribhuvanesh Orekondy, and Mario Fritz. 2020. Gs-wgan: A gradient-sanitized approach for learning differentially private generators. *Advances in Neural Information Processing Systems*, 33:12673–12684.

Hercules Dalianis. 2019. Pseudonymisation of swedish electronic patient records using a rule-based approach. In *Proceedings of the Workshop on NLP and Pseudonymisation*, volume 166, pages 16–23.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is gpt-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*.

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020. Privacy-and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the 13th international conference on web search and data mining*, pages 178–186.

Oluwaseyi Feyisetan, Tom Diethe, and Thomas Drake. 2019. Leveraging hierarchical representations for preserving privacy and utility in text. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 210–219. IEEE.

Nupoor Gandhi, Anjalie Field, and Emma Strubell. 2023. Annotating mentions alone enables efficient domain adaptation for coreference resolution. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10543–10558.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

Jingxuan He and Martin Vechev. 2023. Large language models for code: Security hardening and adversarial testing. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 1865–1879.

Lynette Hirschman and John Aberdeen. 2010. Measuring risk and information preservation: toward new metrics for de-identification of clinical texts. In *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, pages 72–75.

Timour Igamberdiev and Ivan Habernal. 2023. Dp-bart for privatized text rewriting under local differential privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13914–13934.

Frederick Jelinek. 1998. *Statistical methods for speech recognition*. MIT press.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Satyapriya Krishna, Rahul Gupta, and Christophe Dupuy. 2021. Adept: Auto-encoder based differentially private text transformation. *arXiv preprint arXiv:2102.01502*.

Alexey Kurakin, Natalia Ponomareva, Umar Syed, Liam MacDermed, and Andreas Terzis. 2023. Harnessing large-language models to generate private synthetic text. *arXiv preprint arXiv:2306.01684*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*.

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 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.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. 2024. Best practices and lessons learned on synthetic data. *arXiv preprint arXiv:2404.07503*.

Justus Mattern, Zhijing Jin, Benjamin Weggenmann, Bernhard Schoelkopf, and Mrinmaya Sachan. 2022a. Differentially private language models for secure data sharing. *arXiv preprint arXiv:2210.13918*.

Justus Mattern, Benjamin Weggenmann, and Florian Kerschbaum. 2022b. The limits of word level differential privacy. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 867–881.

Omri Mendels, Coby Peled, Nava Vaisman Levy, Sharon Hart, Tomer Rosenthal, Limor Lahiani, et al. 2018. Microsoft Presidio: Context aware, pluggable and customizable pii anonymization service for text and images.

Annika Willoch Olstad, Anthi Papadopoulou, and Pierre Lison. 2023. Generation of replacement options in text sanitization. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 292–300.

Bhanudas Suresh Panchbhai and Varsha Makarand Pathak. 2022. A systematic review of natural language processing in healthcare. *Journal of Algebraic Statistics*, 13(1):682–707.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101.

Pranav Putta, Ander Steele, and Joseph W Ferrara. 2023. Differentially private conditional text generation for synthetic data production.

Krithika Ramesh, Nupoor Gandhi, Pulkit Madaan, Lisa Bauer, Charith Peris, and Anjalie Field. 2024. Evaluating differentially private synthetic data generation in high-stakes domains. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15254–15269, Miami, Florida, USA. Association for Computational Linguistics.

Krithika Ramesh, Daniel Smolyak, Zihao Zhao, Nupoor Gandhi, Ritu Agarwal, Margrét Bjarnadóttir, and Anjalie Field. 2025. Synthtexteval: Synthetic text data generation and evaluation for high-stakes domains. *arXiv preprint arXiv:2507.07229*.

David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. 2019. Analysing mathematical reasoning abilities of neural models. *arXiv preprint arXiv:1904.01557*.

Meng Tong, Kejiang Chen, Xiaojian Yuan, Jiayang Liu, Weiming Zhang, Nenghai Yu, and Jie Zhang. 2025. On the vulnerability of text sanitization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5150–5164.

Saiteja Utpala, Sara Hooker, and Pin Yu Chen. 2023. Locally differentially private document generation using zero shot prompting. *arXiv preprint arXiv:2310.16111*.

Elena Volodina, Yousuf Ali Mohammed, Sandra Derbring, Arild Matsson, and Beata Megyesi. 2020. Towards privacy by design in learner corpora research: A case of on-the-fly pseudonymization of swedish learner essays. In *Proceedings of the 28th international conference on computational linguistics*, pages 357–369.

Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. Sheared llama: Accelerating language model pre-training via structured pruning. In *The Twelfth International Conference on Learning Representations*.

Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. 2018. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. Zerogen: Efficient zero-shot learning via dataset generation. *arXiv preprint arXiv:2202.07922*.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. 2021. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*.

Xiang Yue, Huseyin Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. 2023. Synthetic text generation with differential privacy: A simple and practical recipe. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1321–1342.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does nlp benefit legal system: A summary of legal artificial intelligence. *arXiv preprint arXiv:2004.12158*.

A Control Code

A.1 Control Code of TAB Dataset

The control codes in the TAB dataset are categorized into eight types, as shown in Table 7.

A.2 Fictional Control Code

During generation, we create fictional control codes that follow the format and structure of the original context. The types of fictional control codes are determined based on the categories present in the input context. Specifically, we randomly generate values for seven predefined categories—excluding `MISC`—to simulate plausible yet non-identifiable entities. The generation rules and sampling patterns for each category are detailed in Figure 2.

B Rouge Score

ROUGE-2 measures the overlap of bigrams between the generated and reference texts, capturing local phrase-level similarity. Let $|B_{\text{match}}|$ be the number of overlapping bigrams. Then, the precision P and recall R are given by:

$$P = \frac{|B_{\text{match}}|}{|B_{\text{gen}}|}, \quad R = \frac{|B_{\text{match}}|}{|B_{\text{ref}}|}$$

The ROUGE-2 F1 score is computed as the harmonic mean:

$$\text{ROUGE-2} = \frac{2 \cdot P \cdot R}{P + R}$$

ROUGE-L, on the other hand, computes the length of the longest common subsequence (LCS), which reflects the preservation of word order and sentence structure. Let $\text{LCS}(X, Y)$ denote the length of the longest common subsequence between a reference X and a generation Y, and let $|X|$ and $|Y|$ be the lengths of the reference and generation, respectively. Then:

$$P = \frac{\text{LCS}(X, Y)}{|Y|}, \quad R = \frac{\text{LCS}(X, Y)}{|X|}$$

The ROUGE-L F1 score is given by:

$$\text{ROUGE-L} = \frac{(1 + \beta^2) \cdot P \cdot R}{R + \beta^2 \cdot P}, \quad \text{where } \beta = 1$$

C Baseline

The format of baseline ICL in shown in Table 8.

C.1 Generation

An example of a generated output using the baseline ICL method based on the above prompt is shown in Table 9.

D In-context learning

The format of input of in-context learning is shown in Table 10.

D.1 Generation

An example of a generated output using the in-context learning (ICL) method based on the above prompt is shown in Table 11.

E ICL w/ privacy enhancement

The input format follows the same structure used for in-context learning, as illustrated in Table 10.

E.1 Generation

An example of a generated output using the ICL w/ privacy enhancement method based on the above prompt is shown in Table 12.

F Prefix-tuning

The prompt of prefix-tuning will directly be the fictional code. Generally, the final paragraph will be cut off for the downstream performance.

F.1 Generation

An example of a generated output using the prefix tuning method is shown in Table 13.

G Performance on quasi identifiers

Beyond experiments on direct identifiers and combined direct and quasi-identifier settings, we evaluated performance on quasi-identifiers alone.

Figure 14 shows results for quasi-identifiers under the private known setting. As expected, privacy-related tasks show some performance degradation due to their inherent difficulty. The ICL method achieves approximately 31% on PIPP metrics. In contrast, both finetuning and finetuning with masking demonstrate only modest performance drops, confirming their effectiveness for privacy protection. Notably, ICL with privacy enhancement provides strong protection, achieving near-zero scores on both PIPP and ELP metrics. However, this method maintains higher Rouge-2 and Rouge-L scores compared to finetuning approaches, indicating excessive similarity to original articles that could pose privacy risks.

Category	Description
PERSON	Names of individuals, including full names, nicknames, aliases, usernames, and initials.
CODE	Identifying numbers and codes such as Social Security Numbers (SSNs), phone numbers, passport numbers, license plates, and other personal identifiers.
LOC	Locations and addresses, including cities, regions, countries, streets, and named infrastructures (e.g., bus stops, bridges).
ORG	Names of organizations such as companies, educational institutions, government bodies, healthcare facilities, non-governmental organizations, and religious institutions.
DEM	Demographic attributes including ethnicity, language, heritage, job titles, ranks, education levels, physical descriptions, medical diagnoses, birthmarks, and age-related information.
DATETIME	Mentions of specific dates (e.g., "October 3, 2018"), times (e.g., "9:48 AM"), or durations (e.g., "18 years").
QUANTITY	Quantitative values such as percentages, measurements, or monetary amounts.
MISC	Any other information that could describe an individual but does not fall under the above categories.

Table 7: Private entity categories and their descriptions used for privacy annotation.

Control Code Category	Random Generation Pattern and Sample Values
CODE	Random alphanumeric format: ABCDE/XY Example: X5T9L/QZ
PERSON	Title + First Name + Last Name Sample pool: Titles = Mr, Ms, Dr, Prof, First Names = Alex, Blake, Casey, Dana, Elliot, Finley, Harper, Jordan, Kai, Logan, Morgan, Quinn, Riley, Skyler, Last Names = Adams, Baker, Carson, Dawson, Ellis, Foster, Griffin, Hayes, Irwin, Johnson, Kennedy, Lewis Example: Dr Logan Ellis
DATETIME	Random date between 1990–2024 in the format: DD Month YYYY Example: 12 October 2011
LOC	Random choice from: - Cities: Baltimore, Seattle, Tokyo, Munich, Cairo - Countries: USA, Germany, Japan, Kenya, Brazil - Addresses: 221B Baker St, 1600 Amphitheatre Pkwy, 350 Fifth Ave - Infrastructure: London Bridge, Central Station, Pier 39 Example: 1600 Amphitheatre Pkwy
ORG	Sampled from organization names: OpenAI, World Health Organization, Harvard University, UNICEF, St. Mary's Hospital, SpaceX, NASA, MIT, Stanford University, Google Example: Stanford University
DEM	Pattern-based combinations of: - Heritage: Irish-American, Nigerian, Chinese, Latinx, Punjabi - Jobs: software engineer, nurse, professor, mechanic, pilot - Ages: randomly generated as N-year-old Example: 42-year-old pilot or Latinx descent
QUANTITY	Randomly selected format: - Percentage: 45% - Currency: $87,500 Example: $215,000

Figure 2: Pattern-based generation of fictional control code values for each privacy entity category. Each value is randomly sampled using predefined lists and templates to ensure diversity while preserving format consistency.

Figure 15 presents results under the private unknown setting. ICL with privacy enhancement shows increased leakage (10% in PIPP, 3% in ELP) because Presidio's entity recognition algorithm struggles to identify quasi-identifiers. Consequently, many of the identifiers are not set to zero likelihood, resulting in increased privacy leakage. The masking approach is ineffective due to poor quasi-identifier detection. However, the finetuning stage introduces sufficient generation diversity to

In-Context Learning Prompt to LLM
First Context
The case originated in an application (no. 36244/06) against the Kingdom of Denmark lodged with the Court under Article 34 of the Convention for the Protection of Human Rights and Fundamental Freedoms (“the Convention”) by a Danish national, Mr Henrik Hasslund (“the applicant”), on 31 August 2006. The applicant was represented by Mr Tyge Trier, a lawyer practising in Copenhagen. The Danish Government (“the Government”) were represented by their Agent, Ms Nina Holst-Christensen of the Ministry of Justice.
Second Context
The case originated in an application (no. 29366/03) against the Republic of Poland lodged with the Court under Article 34 of the Convention for the Protection of Human Rights and Fundamental Freedoms (“the Convention”) by Mr D. Stępniak on 25 July 2003. The Polish Government (“the Government”) were represented by their Agent, Mr J. Wołąsiewicz of the Ministry of Foreign Affairs.
Third Context
The case originated in an application (no. 5138/04) against the Republic of Turkey lodged with the Court under Article 34 of the Convention for the Protection of Human Rights and Fundamental Freedoms (“the Convention”) by a Turkish national, Mr Nusret Amutgan (“the applicant”), on 29 December 2003. The applicant was represented by Ms B Özpolat, a lawyer practising in Gaziantep. The Turkish Government (“the Government”) were represented by their Agent.

Table 8: Example of a 3-shot baseline in-context learning prompt used to generate a single synthetic passage.

Example generation from Baseline
PROCEDURE The case originated in an application (no. 5138/04) against the Republic of Turkey lodged with the Court under Article 34 of the Convention for the Protection of Human Rights and Fundamental Freedoms (“the Convention”) by Mr Nusret Amutgan (“the applicant”), on 29 December 2003. The applicant was represented by Ms B Özpolat, a lawyer practising in Gaziantep. The Turkish Government (“the Government”) were represented by their Agent. PROCEDURE The case originated in an application (no. 5138/04) against the Republic of Turkey lodged with the Court under Article 34 of the Convention for the Protection of Human Rights and Fundamental Freedoms (“the Convention”) by Mr Nusret Amutgan (“the applicant”), on 29 December 2003. The applicant was represented by Ms B Özpolat, a lawyer practising in Gaziantep. The Turkish Government (“the Government”) were represented by their Agent.

Table 9: Example generation of baseline in-context learning

limit privacy leakage to approximately 7%.

In-Context Learning Prompt to LLM
First Context
CODE: 36244/06 PERSON: Mr Henrik Hasslund, Mr Tyge Trier, Ms Nina Holst-Christensen DATETIME: 31 August 2006 PROCEDURE The case originated in an application (no. 36244/06) against the Kingdom of Denmark lodged with the Court under Article 34 of the Convention for the Protection of Human Rights and Fundamental Freedoms (“the Convention”) by a Danish national, Mr Henrik Hasslund (“the applicant”), on 31 August 2006. The applicant was represented by Mr Tyge Trier, a lawyer practising in Copenhagen. The Danish Government (“the Government”) were represented by their Agent, Ms Nina Holst-Christensen of the Ministry of Justice.
Second Context
CODE: 29366/03 PERSON: Mr D. Stępniak, Mr J. Wołąsiewicz DATETIME: 25 July 2003 PROCEDURE The case originated in an application (no. 29366/03) against the Republic of Poland lodged with the Court under Article 34 of the Convention for the Protection of Human Rights and Fundamental Freedoms (“the Convention”) by Mr D. Stępniak on 25 July 2003. The Polish Government (“the Government”) were represented by their Agent, Mr J. Wołąsiewicz of the Ministry of Foreign Affairs.
Third Context
CODE: 5138/04 PERSON: Mr Nusret Amutgan, Ms B Özpolat DATETIME: 29 December 2003 PROCEDURE The case originated in an application (no. 5138/04) against the Republic of Turkey lodged with the Court under Article 34 of the Convention for the Protection of Human Rights and Fundamental Freedoms (“the Convention”) by a Turkish national, Mr Nusret Amutgan (“the applicant”), on 29 December 2003. The applicant was represented by Ms B Özpolat, a lawyer practising in Gaziantep. The Turkish Government (“the Government”) were represented by their Agent.
Randomly Generated Control Code
CODE: JMIRM/QT PERSON: Ms Elliot Carson DATETIME: 16 June 2007

Table 10: Example of a 3-shot in-context learning prompt used to generate a single synthetic passage.

Example generation from ICL
\nPROCEDURE\n The case originated in an application (no. 29762/99) against the Republic of Yugoslavia lodged with the Court under Article 34 of the Convention for the Protection of Human Rights and Fundamental Freedoms ("the Convention") by a Yugoslav national, Prof Kai Kennedy ("the applicant"), on 20 January 2000. The applicant was represented by Mr J. B. S. Fleming, a lawyer practising in Zagreb. The Republic of Yugoslavia ("the Republic") were represented by their Agent, Ms I. K. Jankovic of the Ministry of Foreign Affairs.

Table 11: Example paragraph generated from the above prompt using the in-context learning (ICL) method

Generated Paragraph from ICL w/ privacy enhancement
PROCEDURE The case originated in an application (no. 2AGVC/7B) against the Republic of Turkey lodged with the Court under Article 34 of the Convention for the Protection of Human Rights and Fundamental Freedoms ("the Convention") by Professor Skyler Baker ("the applicant"), on 26 August 2006. The applicant was represented by Mr M. S. Yildirim, a lawyer practising in Istanbul. The Turkish Government ("the Government") were represented by their Agent, Ms Z. Özkan of the Ministry of Foreign Affairs.

Table 12: Example paragraph generated using icl w/ privacy enhancement method.

Example Paragraph of prefix-tuning
LOCATION: DUYNK Skyler Johnson was 19 years old when he died in a car crash on the night of 9 September 2018. He was in the passenger seat of a grey Holden Commodore which collided with a white Toyota Corolla on the night of 9 September 2018. The Corolla driver, 38 year old, was not injured.

Table 13: Full academic biography paragraph and publication entries used as example input text.

Method	**PIPP (%)** ↓	**ELP (%)** ↓	**ROUGE-2**↓	**ROUGE-L** ↓
ICL	31.36	3.94	0.2726	0.3057
ICL w/ privacy enhancement	**0.88**	**0.06**	0.2624	0.2988
Fine-tuning	2.76	1.02	0.0087	0.0794
Fine-tuning w/ masking	2.88	1.07	**0.0084**	**0.0793**

Table 14: Privacy protection performance under the private entity known setting, where only quasi identifiers are treated as private entities, on the TAB dataset

Method	**PIPP (%)** ↓	**ELP (%)** ↓	**ROUGE-2**↓	**ROUGE-L** ↓
ICL	40.21	6.79	0.3233	0.0.3487
ICL w/ privacy enhancement	13.21	3.06	0.2624	0.2988
Fine-tuning	**7.21**	3.02	0.053	0.0981
Fine-tuning w/ masking	7.34	**2.38**	**0.019**	**0.0823**

Table 15: Privacy protection performance under the private entity unknown setting, where only quasi identifiers are treated as private entities, on the TAB dataset