Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Membership Disclosure Evaluation for Synthetic Clinical Text Generated by LLM via Dynamic Few-Shot In-Context Learning

Sen Li

Sen.Li.Sprout@gmail.com

University of Ottawa

Fida K. Dankar

CHEO Research Institute

Khaled El Emam

University of Ottawa

Research Article

Keywords: MIA, Synthetic Note, LLM, In-Context Learning, Few-Shot

Posted Date: March 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8997592/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Membership Disclosure Evaluation for Synthetic Clinical Text Generated by LLM via Dynamic Few-Shot In-Context Learning

Sen Li[1,2*], Fida K. Dankar[2] and Khaled El Emam[1,2†]

[1*]School of Epidemiology and Public Health, University of Ottawa, Ottawa, ON, Canada.
[2]CHEO Research Institute, Ottawa, Ontario, Canada.

*Corresponding author(s). E-mail(s): Sen.Li.Sprout@gmail.com;
Contributing authors: fdankar@cheo.on.ca; kelemam@uottawa.com;
†These authors contributed equally to this work.

Abstract

Purpose: LLMs are increasingly used to generate synthetic clinical notes for data sharing, but membership inference attacks (MIAs) can reveal whether specific real notes appeared in prompts or training data. Prior studies often examine static few-shot prompts, i.e., fixed exemplar notes that remain the same across queries. We investigate MIAs on synthetic clinical text datasets as an evaluation metric for membership disclosure risk from the perspective of data custodians, focusing on synthetic outputs generated by dynamic few-shot ICL with LLMs. Our goal is to examine whether it is possible to estimate the membership disclosure risk of the generative pipeline using only the real dataset, by simulating MIAs on synthetic outputs and a limited adversarial dataset. **Methods:** We develop a custodial auditing framework based on sentence-embedding similarity between real notes and the synthetic corpus, using three features: Max (nearest-neighbor similarity), Top-3 (mean of three strongest similarities), and Mean (centroid alignment). We formalize membership under both static and dynamic ICL and introduce a population-aware partitioning protocol that preserves realistic priors for evaluation. We report F1 and an adjusted metric M=(F1-Fmax)/(1-Fmax), where Fmax denotes the F1 score achieved by the naive adversarial strategy of classifying all records as members. Experiments span static and dynamic ICL with $\boldsymbol{k} \in \{\mathbf{1}, \mathbf{2}, \mathbf{3}\}$ and include extensive resampling to assess stability. The entire pipeline is text-only and requires no access to prompts, gradients, or model weights. **Results:** Static 1-shot ICL exhibits high

membership disclosure risk, with Max and Top-3 achieving strong attack performance (high F1 and large positive $\boldsymbol{M}$), indicating substantial leakage. Increasing the number of shots to 2 or 3 in static ICL markedly reduces risk, yielding $\boldsymbol{M}$ below conservative thresholds and bringing performance close to membership-preserving regimes. Dynamic ICL consistently suppresses disclosure across all $\boldsymbol{k}$, with F1 near chance and $\boldsymbol{M} \approx \mathbf{0}$ or slightly negative, signifying negligible exploitable signal. The largest contrast between regimes occurs at 1-shot, underscoring prompt reuse as the primary driver of risk. **Conclusion:** Dynamic few-shot ICL offers a practical mitigation for membership disclosure in synthetic clinical text. Static 1-shot prompting should be avoided for public releases; when static prompting is unavoidable, using 2–3 examples substantially mitigates leakage. Our text-only, model-agnostic auditing provides custodians with deployable tools and an interpretable metric to decide release readiness. Ongoing experiments will expand datasets, LLMs, and auxiliary data sizes, and assess robustness to domain drift and prompt formatting changes.

Keywords: MIA, Synthetic Note, LLM, In-Context Learning, Few-Shot

1 Introduction

Large language models (LLMs) have transformed synthetic text generation, enabling applications from automated content creation to customized decision support systems. Trained on vast text corpora, these models produce fluent, contextually relevant outputs that can be repurposed for downstream tasks, including data augmentation, event prediction, knowledge extraction, etc. However, this capability raises significant privacy concerns. In particular, membership disclosure or membership inference - which aims to determine whether a specific record contributed to a model's training data or appeared in its prompt - poses concrete risks of sensitive information disclosure, intellectual property leakage, and regulatory non-compliance in high-stakes domains such as healthcare and finance [1].

In the context of synthetic data, a common (and often implicit) assumption is that synthesizing new samples from a trained generative system removes or substantially reduces privacy risk. Yet, prior work shows that synthetic data can still leak information about individuals in the training set [2]. This leakage can occur when a generator memorizes examples or encodes subtle signals that enable an attacker to distinguish members from non-members. Recent evidence in tabular data, including diffusion-based synthesis in healthcare and finance, corroborates this risk, demonstrating non-trivial membership inference attack (MIA) performance against state-of-the-art generators [3]. In text, the risk is amplified by the rich lexical and semantic structure of natural language and by the increasing use of LLMs to produce synthetic medical contexts in privacy-sensitive settings (e.g., clinical note synthesis) [4].

Beyond training-time privacy, prompting-time privacy is increasingly important because many LLM applications employ in-context learning (ICL) rather than fine-tuning. Given a small set of input–output demonstrations in a prompt, an LLM can perform a new task without parameter updates [5, 6]. This paradigm is widely used

in healthcare for triage assistants, report drafting, and data harmonization [7–9]. If an adversary can infer that a specific patient's data appeared in an ICL prompt, then sensitive attributes (e.g., diagnoses) may be revealed [10, 11]. Recent studies indicate that prompted (non-fine-tuned) LLMs can be more vulnerable to MIAs than fine-tuned models: text-only attacks against static ICL achieve high success rates, revealing membership of prompt examples under static few-shot prompting [7, 10].

These developments motivate a careful reevaluation of membership risks when LLMs with ICL are used to generate synthetic clinical text. In practice, data custodians increasingly consider LLM-generated synthetic notes as substitutes for real notes in method development and sharing. However, two key gaps remain. First, existing work on MIAs for ICL largely focuses on fixed few-shot prompts, whereas many real-world systems use dynamic few-shot selection: for each input, the prompt is constructed on-the-fly by retrieving task-relevant exemplars from a reference pool. This dynamic construction can potentially reduce memorization-like behaviors and blur membership signals, but its implications for membership disclosure through synthetic outputs are not well evaluated. Second, most ICL-focused MIAs assume access to the prompt or model scores. In contrast, custodians releasing synthetic corpora often only expose the syntheses. There is a need for text-only auditing methods that estimate membership disclosure risk from synthetic text alone, possibly supplemented with a small auxiliary corpus drawn from the same distribution.

MIA methodologies have received substantial attention in the literature as a fundamental privacy metric. Generally, MIA methods can be organized into three broad categories: threshold-based, similarity-based, and model-based attacks [12–14]. **Threshold-based MIAs** rely on fixed or adaptive thresholds applied to output metrics—such as confidence scores, loss values, or prediction entropy—to determine whether an input sample belongs to the training set [1, 7, 15–17]. **Similarity-based MIAs** are behavior-driven pairwise comparison techniques that exploit inherent data relationships observed in the target model's output behaviors (e.g., embeddings or gradients), rather than relying on raw input records [18–24]. If one point within a cluster is known to be a training member, it is often assumed that all points in the same cluster are likely members as well. **Attack-model-based MIAs** involve training shadow (surrogate) models that replicate the behavior of the target model on datasets with known membership labels. These surrogate models are then used to construct a specialized attack classifier based on their behavior, which is subsequently applied to predict membership in the target model [25–27]. The overall taxonomy is useful because it distinguishes simple statistical tests (threshold-based MIAs) from richer representation-based attacks (similarity-based MIAs) and from learned-model attacks (attack-model-based MIAs) [28]. By categorizing attacks in this way, researchers can systematically compare methods along dimensions such as required assumptions, data availability, and computational cost, thereby selecting the approach that best aligns with their downstream tasks.

Problem setting and scope

We investigate MIAs on synthetic clinical text datasets as an evaluation metric for membership disclosure risk from the perspective of data custodians, focusing on synthetic outputs generated by dynamic few-shot ICL with LLMs. In this setting, for each synthetic note generation, k real clinical notes are randomly sampled from a source pool to construct the prompt. The LLM conditions on these k-shot examples and subsequently produces a synthetic clinical note.

We assume an adversary who has access only to the released synthetic corpus, without visibility into the prompts, token-level confidence scores, or model gradients. The adversary may, however, possess a small auxiliary dataset of clinical notes drawn from the same distribution as the real corpus (e.g., originating from the same institution and time period).

In practice, data custodians have access solely to the real note dataset, but not to the adversary's auxiliary data or to the underlying population from which the dataset was sampled. Consequently, custodians must treat the available real dataset as a surrogate for the population.

Our analysis centers on text-only auditing: specifically, whether it is possible to estimate the membership disclosure risk of the generative pipeline using only the real dataset, by simulating MIAs on the basis of synthetic outputs and a limited adversarial dataset.

Our approach

We propose a custodial, text-only auditing framework to evaluate membership disclosure risk in synthetic clinical note generation pipelines. First, we introduce a population-aware data partitioning strategy for the real-note corpus that preserves realistic class priors: the original dataset is split into a Member source pool and a Non-Member source pool, from which we sample both the simulated note set under attack and the adversarial note set. We formalize membership under both static and dynamic k-shot ICL, ensuring that evaluation reflects the operational conditions faced by data custodians.

To support this evaluation, we design similarity-based MIAs tailored to the clinical auditing scenario. The central idea is to compute sentence-level embeddings for the synthetic corpus and measure semantic proximity to each candidate real note drawn from the same distribution. We instantiate three model-agnostic features that do not require access to prompts, confidence scores, gradients, or model weights: (i) Max, the nearest-neighbor similarity between a real note and the synthetic corpus; (ii) Mean, the centroid alignment between a real note and the synthetic corpus; and (iii) Top-3, the mean of the three strongest similarities.

For risk quantification, we report both F1 and an adjusted explainable metric $M = (\mathrm{F1} - \mathrm{F}_{\max})/(1 - \mathrm{F}_{\max})$, where $\mathrm{F}_{\max}$ denotes the F1 achieved by a naive adversary that classifies all records as members. Our experiments span static and dynamic ICL with $k \in \{1, 2, 3\}$ and include extensive resampling to assess stability and variance. The entire pipeline operates in a text-only setting and is readily implementable by custodians without access to model internals.

Summary of findings

Across clinical datasets and LLM backbones, we observe:

1. Static 1-shot ICL exhibits high membership disclosure risk: Max and Top-3 features yield very strong attack performance (F1 near 0.89) with large positive M, indicating substantial leakage.
2. Increasing k to 2 or 3 in static ICL sharply reduces attack performance, bringing M below 0.2 across features and entering a membership-preserving regime.
3. Dynamic ICL drastically suppresses disclosure across all k, with F1 near chance and M approximately zero or slightly negative, pointing to negligible exploitable signal under text-only attacks.

Overall, prompt reuse is the dominant driver of risk; per-input re-sampling in dynamic ICL obfuscates exemplar-specific signatures more effectively than increasing k alone.

Contributions

This work makes the following contributions:

1. We formulate and define the problem of evaluating membership disclosure risk in the generative pipeline of synthetic clinical text produced via both static and dynamic few-shot ICL, from the perspective of data custodians, with a focus on a text-only adversarial setting that reflects practical data release scenarios.
2. We introduce three semantic similarity features based on sentence embeddings for auditing membership disclosure risk using only the synthetic corpus and a small auxiliary real-text set from the same distribution.
3. We provide an empirical study on clinical datasets showing that single learning sample and prompt reuse are the dominant driver of membership disclosure risk, and dynamic few-shot ICL lowers membership disclosure risk compared to static few-shot ICL.

2 Related Work

2.1 In-Context Learning and Dynamic Few-Shot Prompting

LLMs can solve new tasks at inference time by conditioning on prompts that include instructions or a few demonstrations, a paradigm known as in-context learning (ICL) [5, 29, 30]. Few-shot ICL leverages pre-trained knowledge and adapts to tasks without parameter updates, in contrast to traditional few-shot learning that fine-tunes on small datasets [7]. Beyond fixed few-shot prompts, a growing body of work studies dynamic few-shot selection, where the examples in the prompt are retrieved on-the-fly based on the input. This approach is often implemented via retrieval-augmented generation (RAG) and has been shown to yield performance benefits by tailoring context to each query [31]. Notably, the optimal number of in-context examples appears to depend on input complexity: more complex samples benefit from additional demonstrations, whereas overly many demonstrations can degrade accuracy for simple inputs [30–32]. These observations motivate our focus on dynamic few-shot prompting in clinical text

generation, where the output diversity is substantial and static prompting may be sub-optimal.

2.2 Membership Inference Attacks: General and for LLMs

MIAs aim to determine whether a data record contributed to a model's training data (or to the inference context in the case of ICL). MIAs are a central tool for privacy auditing across modalities and learning paradigms [12–14]. For LLMs, recent evaluations report that standard MIAs often underperform, especially when models are trained on large, deduplicated corpora with limited epochs, which blur the distinction between members and non-members due to high lexical overlap in natural language [1]. Apparent MIA success is frequently attributable to distribution shifts (e.g., temporal drift) rather than true memorization, and attacks can be brittle to minor text perturbations [1]. These findings underscore the need for auditing protocols that reflect realistic deployment conditions and account for the inherently fuzzy membership boundary in text.

2.3 MIAs for Static In-Context Learning and RAG Systems

A complementary line of work studies MIAs in prompted (non-fine-tuned) LLMs. In static few-shot prompting, several studies demonstrate that membership of prompt exemplars can be inferred with text-only signals or with limited model feedback [7, 10]. These results highlight a distinct threat surface at inference time: sensitive examples embedded in static prompts may leave detectable traces in generated outputs, enabling membership disclosure without access to model parameters or training pipelines.

RAG-based systems introduce additional considerations due to retrieval dynamics. Because the context is retrieved per query, the exposure of any specific exemplar is stochastic and task-dependent. Recent work begins to probe membership risks in this setting. Li et al. [33] propose S^2MIA, a semantic-similarity attack that infers whether a query appears in a RAG knowledge base by combining cosine similarity and perplexity-based features, achieving strong performance on medical QA benchmarks. Anderson et al. [34] show that membership of documents in retrieval indices can be detected efficiently in black-box and gray-box regimes by crafting appropriate prompts. Beyond text-only RAG, multimodal systems also exhibit membership leakage: Yang et al. [35] introduce MrM, a black-box MIA for vision–language RAG that uses object-aware perturbations and counterfactual masking to induce retrieval and expose membership signals, demonstrating robustness across commercial models. Together, these studies suggest that RAG-based systems do not eliminate membership risk; instead, they reshape the attack surface toward semantic similarity and retrieval dynamics.

2.4 Membership Disclosure in Synthetic Data and Clinical Text

Synthetic data generation is often adopted to mitigate privacy risks, yet multiple studies show that synthetic datasets may leak membership information, including in healthcare contexts [2, 3]. In text, rich lexical and semantic structures can make leakage more subtle but still consequential, particularly in clinical narratives where

rare conditions, unique phrasing, or institution-specific templates can act as quasi-identifiers [4]. Prior work in clinical NLP has largely focused on de-identification or release policies, with fewer studies quantifying membership disclosure from synthetic clinical notes. Moreover, most existing evaluations either assume access to prompts or model confidences, or they study fixed few-shot ICL, leaving open questions about dynamic few-shot generation and text-only auditing.

2.5 Positioning and Distinctions

Our work differs from the above in three respects aligned with the deployment realities highlighted in Section 1: (i) we study dynamic few-shot ICL for synthetic clinical note generation, where exemplars are randomly sampled per input rather than fixed in a static prompt; (ii) we evaluate membership disclosure under a text-only adversarial model, assuming no access to prompts, confidences, gradients, or training pipelines; and (iii) we operationalize similarity-based MIAs tailored to this auditing scenario, using sentence-level semantic features that capture proximity between released synthetic notes and candidate real notes drawn from the same distribution. This complements prior MIAs for LLMs and RAG by focusing on corpus-level auditing of synthetic outputs, which matches how custodians typically release data, and by providing evidence about how the number of dynamic shots modulates risk in clinical settings.

3 Methods

3.1 Problem Formulation and Definitions

In this section, We formulate and define the problem of evaluating membership disclosure risk in the generative pipeline of synthetic clinical text produced via few-shot ICL, both dynamically and statically, from the perspective of data custodians, with a focus on a text-only adversarial setting that reflects practical data release scenarios.

3.1.1 Membership in Dynamic Few-Shot In-Context Learning

In the dynamic few-shot ICL setting, each synthetic text is generated by conditioning the LLM on a uniquely sampled combination of few-shot examples. Specifically, for each generation request, we randomly sample a small number of examples (i.e., k-shot) from a real text set—such as a private collection of real clinical notes—and use them as the context prompt. Each generated synthetic text s is therefore associated with a unique index tuple that identifies the specific examples used in its prompt. To avoid reuse, we re-sample whenever a previously used index tuple is encountered. This dynamic sampling strategy ensures that each synthetic instance is generated with a distinct prompt configuration, effectively simulating the diversity observed across real patients.

In contrast, the static few-shot ICL setting fixes a single set of k examples that are reused across multiple generation requests. Under this paradigm, a group of synthetic texts share the same context prompt, enable more efficient generation when prompt construction is costly, and possibly exhibit reduced variance due to the lack of prompt

changes. While static ICL facilitates consistency and repeatability, it may capture less variability than the dynamic approach.

We establish a clear notation hierarchy by first defining abstract concepts, then concrete instances. Let $\mathcal{R}$ represent an **abstract set of real clinical text records**, and $\mathcal{S}$ an **abstract set of synthetic clinical text records** generated from some $\mathcal{R}$. Concrete datasets are then specified using subscripts:

- $\mathcal{R}_0$: Private real note set (target dataset under attack).
- $\mathcal{S}_0$: Synthetic note set intended for release.
- $\mathcal{R}_\mathrm{A}$: Real note set accessible to the adversary.

Individual records are denoted as $r_i \in \mathcal{R}$ where $i \in [1, |\mathcal{R}|]$ indexes records within any real set. For synthetic text generation under k-shot ICL, we define an *index tuple* $t^k = (i_1, \ldots, i_k)$ identifying the specific k real examples sampled from $\mathcal{R}_0$ to generate a synthetic record.

This allows unified definitions of **Member** (superscript M) and **Non-Member** (superscript NM) applicable to:

- Any abstract real set $\mathcal{R}$.
- The concrete target real set $\mathcal{R}_0$.
- The concrete adversarial real set $\mathcal{R}_\mathrm{A}$.
- The concrete real set $\mathcal{D}$ that is accessible to data custodians, who evaluate the membership disclosure risks of $\mathcal{S}_0$. This notation is adopted from [18].

within the few-shot ICL paradigm:

- A **Member** is defined as a data record r_i from an abstract real set $\mathcal{R}$ that is used at least once as a few-shot example in the synthetic text generation process. For each generation request indexed by $g \in [1, |\mathcal{S}|]$, a k-shot index tuple $t^k_g = (i_{g,1}, \ldots, i_{g,k})$ is sampled from the real dataset and used as the prompt for producing a synthetic text s_g. The number of times a real record r_i is exposed to the target LLM across all generation requests is denoted as its *exposure count*, defined by

$$j_i = \left|\{\, g : i \in t^k_g \,\}\right|.$$

  A record r_i is considered a Member if and only if $j_i \geq 1$; that is, it appears in at least one sampled prompt and is therefore exposed to the target model at least once. The set of all such records constitutes the Member real dataset, denoted by

$$\mathcal{R}^\mathrm{M} = \{\, r_i \in \mathcal{R} : j_i \geq 1 \,\}.$$

  In the dynamic one-shot ICL setting ($k = 1$), each Member record has exactly one exposure ($j_i = 1$) if sampled, or zero exposures ($j_i = 0$) otherwise. In contrast, under dynamic k-shot ICL with $k > 1$, a record may appear in multiple distinct k-shot prompts across generations, yielding $j_i \geq 1$. In the static k-shot ICL setting, where a fixed tuple t^k is reused across multiple generation requests, the exposure counts j_i for records included in the fixed tuple may be substantially larger, corresponding to the number of times the static prompt is applied.

- A **Non-Member** is defined as a data record r_i if and only if $j_i = 0$; that is, it does not appear in any sampled prompt and is never exposed to the target LLM during the generation process. The set of all such records constitutes the Non-Member real dataset, denoted by
$$\mathcal{R}^{\mathrm{NM}} = \{\, r_i \in \mathcal{R} : j_i = 0 \,\}.$$

3.1.2 Membership Evaluation by Data Custodians

The objective of membership disclosure risk evaluation in dynamic few-shot ICL for synthetic text generation is to quantify the vulnerability of the generative system to MIAs. This evaluation is typically conducted by the data custodian prior to the public release of a synthetic text dataset. The central question is whether an adversary can reliably distinguish whether a target data record x, drawn from $\mathcal{R}_{\mathrm{A}}$, was included in the few-shot context during synthetic text generation (i.e., $x \in \mathcal{R}_0^{\mathrm{M}}$) or excluded (i.e., $x \in \mathcal{R}_0^{\mathrm{NM}}$).

We adopt the notation of population $\mathcal{P}$ from [18]. In clinical practice, data custodians only have access to the real dataset $\mathcal{D}$, however, neither to the adversarial dataset $\mathcal{R}_{\mathrm{A}}$ nor to the underlying population $\mathcal{P}$ from which $\mathcal{D}$ is sampled. Therefore, custodians must rely exclusively on the available real dataset $\mathcal{D}$, treating it as a surrogate for the population $\mathcal{P}$. To simulate realistic adversarial conditions, $\mathcal{D}$ is partitioned into a Member sub-pool $\mathcal{D}^{\mathrm{M}}$ and a Non-Member sub-pool $\mathcal{D}^{\mathrm{NM}}$, such that the relative size of the Member sub-pool reflects the sampling fraction of $\mathcal{D}$ with respect to the population:
$$\frac{|\mathcal{D}^{\mathrm{M}}|}{|\mathcal{D}|} = \frac{|\mathcal{D}|}{|\mathcal{P}|}.$$
Under this construction, the membership disclosure risk for the full synthetic dataset $\mathcal{S}^{\mathcal{D}}$, generated from $\mathcal{D}$, can be estimated by evaluating the sub-synthetic dataset $\mathcal{S}^{\mathcal{D}^{\mathrm{M}}}$, generated from $\mathcal{D}^{\mathrm{M}}$. In other words, in the practical evaluation of the membership disclosure risk given solely $\mathcal{D}$, our strategy is to re-define the real text set and synthetic text that are under attack, i.e., $\mathcal{R}_0 = \mathcal{D}^{\mathrm{M}}$ and $\mathcal{S}_0 = \mathcal{S}^{\mathcal{D}^{\mathrm{M}}}$. The corresponding simulated adversarial dataset $\mathcal{R}_{\mathrm{A}}$ is then constructed by combining two subsets: (i) a Member dataset $\mathcal{R}_{\mathrm{A}}^{\mathcal{D}^{\mathrm{M}}}$, sampled from $\mathcal{D}^{\mathrm{M}}$, and (ii) a Non-Member dataset $\mathcal{R}_{\mathrm{A}}^{\mathcal{D}^{\mathrm{NM}}}$, sampled from $\mathcal{D}^{\mathrm{NM}}$. The member ratio within the simulated adversarial dataset is thus preserved as

$$\frac{|\mathcal{R}_{\mathrm{A}}^{\mathcal{D}^{\mathrm{M}}}|}{|\mathcal{R}_{\mathrm{A}}|} = \frac{|\mathcal{D}^{\mathrm{M}}|}{|\mathcal{D}|} = \frac{|\mathcal{D}|}{|\mathcal{P}|}. \tag{1}$$

This partitioning strategy ensures that the simulated adversarial attack remains realistic, since an adversary randomly sampling from the population $\mathcal{P}$ would, in expectation, overlap with the real dataset at approximately the rate $\frac{|\mathcal{D}|}{|\mathcal{P}|}$. If condition Equation (1) is not satisfied, the resulting membership disclosure evaluation, measured for instance by the F1 score, may be misleadingly high or low, thereby yielding an inaccurate estimate of membership disclosure risk [18]. An illustrative example of the data partitioning procedure is provided in **Fig. 1**.

In this text-only auditing from the aspect of data custodians, the simulated adversary is assumed to have access to:

1. The complete synthetic text dataset $\mathcal{S}_0$;
2. A small pool of unlabeled real records $\mathcal{R}_\text{A}$, of size $|\mathcal{R}_\text{A}|$.

A general formulation of the adversarial membership attack can be expressed as the execution of the function

$$A(x \mid \mathcal{R}_\text{A}, \mathcal{S}_0) \;\rightarrow\; \{\text{M-}\mathcal{R}_0,\;\; \text{NM-}\mathcal{R}_0\}$$

where the adversary A takes as input a target data record x and outputs a membership prediction indicating whether x is a Member (M) or Non-Member (NM) of the ground-truth dataset $\mathcal{R}_0$. To address this auditing problem, we introduce three semantic similarity features derived from sentence embeddings, as described in Section 3.2.

3.2 Similarity-based MIAs

To quantify membership disclosure risk in the text-only auditing, where only the synthetic dataset $\mathcal{S}_0$ and a set of unlabeled real records $\mathcal{R}_\text{A}$ are accessible to the adversary, the key intuition is that real examples used in few-shot ICL prompts may exert semantic influence over the generated synthetic texts, resulting in higher textual or representational similarity between member records and the synthetic corpus. In contrast, non-member records are expected to exhibit weaker similarity. Therefore, by quantifying this similarity, one can infer potential membership status.

We propose unsupervised similarity-based MIA methods that exploit semantic relationships between a real record x and the synthetic dataset $\mathcal{S}_0$. Let $\text{nemb}(\cdot)$ denote a pre-trained sentence embedding model (e.g., Sentence-BERT), such that for any text y, $\text{nemb}(y) \in \mathbb{R}^d$ returns a normalized sentence embedding vector satisfying

$$\|\text{nemb}(y)\|_2 = 1.$$

The central idea is to represent each adversarial real record by a scalar feature θ that quantifies its semantic similarity or distance to the synthetic dataset $\mathcal{S}_0$ based on normalized sentence embedding vectors, where each vector encodes the holistic meaning of an entire note. To ensure that θ reflects the relationship between a real record and the overall distributional characteristics of $\mathcal{S}_0$—rather than its proximity to any single synthetic instance—we consider dataset-level similarity features. Once all real records are mapped to their corresponding θ values, an unsupervised clustering algorithm (e.g., Gaussian Mixture Model or k-means) is applied to partition the records into two groups. The cluster with the higher average similarity score $\bar{\theta}$ to $\mathcal{S}_0$ is inferred to represent Member records, while the other is classified as Non-Members.

In summary, for each adversarial record $x \in \mathcal{R}_\text{A}$, we compute a scalar feature θ that quantifies its distributional connection to the synthetic dataset $\mathcal{S}_0$. The proposed unsupervised, similarity-based MIA method comprises three principal steps:

1. Derive a dataset-level scalar feature θ that measures the similarity (or, conversely, the distance) between a target record x and the synthetic text dataset $\mathcal{S}_0$ based on normalized Sentence-Embeddings.
2. Compute this feature θ for all records in the unlabeled adversarial set $\mathcal{R}_\text{A}$.

3. Partition the resulting θ values (and correspondingly, the adversarial records) into *Member* and *Non-Member* clusters using an unsupervised learning algorithm, e.g., Gaussian Mixture Model or k-means.

In the following sections, we introduce and evaluate three distinct dataset-level similarity features designed for this purpose.

3.2.1 Maximum Cosine Similarity

The maximum cosine similarity measures the "closest" relationship between a real record $x \in \mathcal{R}_{\mathrm{A}}$ and any synthetic record in $\mathcal{S}_0$. The underlying intuition is that member records, having influenced at least one synthetic generation, are more likely to exhibit a higher maximum similarity score with at least one record in $\mathcal{S}_0$ compared to non-member records. This feature is a scalar quantity that extends to the entire dataset $\mathcal{S}_0$ by focusing on the closest match.

For each real record $x \in \mathcal{R}_{\mathrm{A}}$, we calculate the cosine similarity between the normalized embeddings of x and each synthetic text $s \in \mathcal{S}_0$. This similarity, denoted as $\mathrm{sim}(x, s)$, is computed as follows:

$$\mathrm{sim}(x, s) = \mathrm{nemb}(x) \cdot \mathrm{nemb}(s) \tag{2}$$

where $\mathrm{nemb}(x)$ and $\mathrm{nemb}(s)$ are the normalized embeddings of x and s, respectively, and $\cdot$ denotes the dot product. The cosine similarity metric ranges from -1 to 1, where higher values indicate greater semantic similarity between the embeddings.

We then define the feature θ for x as the maximum similarity across all synthetic texts:

$$\theta_{\max}(x) = \text{max-sim}(x) = \max_{s \in \mathcal{S}_0} \mathrm{sim}(x, s) \tag{3}$$

which is a scalar value representing the strongest semantic connection between x and any synthetic text. This formulation aligns with the hypothesis that member records will exhibit at least one highly similar synthetic output.

3.2.2 Mean Cosine Similarity

The maximum cosine similarity $\theta_{\max}(x)$ can be susceptible to outliers and may fail to accurately capture the overall influence of a target record x on the synthetic text dataset $\mathcal{S}_0$ in the dynamic few-shot ICL setting. A more robust and distribution-aware alternative is the mean cosine similarity $\theta_{\mathrm{mean}}(x)$, which aggregates the similarity of x to every synthetic text $s \in \mathcal{S}_0$:

$$\begin{aligned} \theta_{\mathrm{mean}}(x) &= \frac{1}{|\mathcal{S}_0|} \sum_{s \in \mathcal{S}_0} \mathrm{sim}(x, s) \\ &= \frac{1}{|\mathcal{S}_0|} \sum_{s \in \mathcal{S}_0} \mathrm{nemb}(x) \cdot \mathrm{nemb}(s) \end{aligned} \tag{4}$$

The mean cosine similarity measures the central tendency of the similarity between a target record x and the entire distribution of the synthetic text dataset $\mathcal{S}_0$ in

the embedding space. It effectively assesses how "aligned" a target record is with the semantic manifold of the synthetic text dataset. This feature mitigates the dilution effect of k-shot ($k > 1$) ICL generation while reducing sensitivity to individual anomalous synthetic texts.

To gain further geometric insight into this feature, it is instructive to note that the mean cosine similarity $\theta_{\text{mean}}(x)$ is directly related to the cosine similarity between target record x and the **centroid** $c_{\mathcal{S}_0}$ of the semantic manifold defined by $\mathcal{S}_0$ in the embedding space. The unnormalized centroid embedding is computed as:

$$\text{emb}(c_{\mathcal{S}_0}) = \frac{1}{|\mathcal{S}_0|} \sum_{s \in \mathcal{S}_0} \text{nemb}(s). \tag{5}$$

Since $\mathcal{S}_0$ is generated by conditioning on members of $\mathcal{R}_0$, its centroid is heuristically aligned with the semantic distribution of the member records. For a target real record $x \in \mathcal{R}_{\text{A}}$, the centroid cosine similarity feature $\theta_{\text{centroid}}(x)$ is defined as:

$$\theta_{\text{centroid}}(x) = \text{nemb}(x) \cdot \frac{\text{emb}(c_{\mathcal{S}_0})}{||\text{emb}(c_{\mathcal{S}_0})||_2} \tag{6}$$

Therefore, the mean cosine similarity θ_{mean} can be expressed by the centroid cosine similarity θ_{centroid} following:

$$\begin{aligned}
\theta_{\text{mean}}(x) &= \frac{1}{|\mathcal{S}_0|} \sum_{s \in \mathcal{S}_0} \text{nemb}(x) \cdot \text{nemb}(s) && \text{(Equation 4)} \\
&= \text{nemb}(x) \cdot \frac{1}{|\mathcal{S}_0|} \sum_{s \in \mathcal{S}_0} \text{nemb}(s) \\
&= \text{nemb}(x) \cdot \text{emb}(c_{\mathcal{S}_0}) \cdot \frac{||\text{emb}(c_{\mathcal{S}_0})||_2}{||\text{emb}(c_{\mathcal{S}_0})||_2} && \text{(Equation 5)} \\
&= \theta_{\text{centroid}}(x) \cdot ||\text{emb}(c_{\mathcal{S}_0})||_2 && \text{(Equation 6)}
\end{aligned}$$

The mean cosine similarity θ_{mean} equals the centroid cosine similarity θ_{centroid} scaled by the constant coherence factor $||\text{emb}(c_{\mathcal{S}_0})||_2$, which measures the directional concentration of $\mathcal{S}_0$ in embedding space. This relationship establishes mean cosine similarity as an interpretable measure of the semantic alignment between x and the central tendency of $\mathcal{S}_0$, making it a suitable feature for membership inference.

Since $\|\text{emb}(c_{\mathcal{S}_0})\|_2$ is constant for all x, θ_{centroid} and θ_{mean} induce identical relative orderings. Thus, θ_{centroid} preserves rank-based relationships while offering computational efficiency: it requires one embedding aggregation for $\mathcal{S}_0$ and one similarity computation per x, whereas θ_{mean} requires $|\mathcal{S}_0|$ similarity calculations per x. For large $\mathcal{S}_0$, θ_{centroid} is preferred.

3.2.3 Top-3 Cosine Similarity

While the maximum cosine similarity $\theta_{\max}(x)$ is sensitive to outliers or spurious alignments arising from noise, the mean cosine similarity $\theta_{\text{mean}}(x)$ may become overly diluted when the synthetic dataset $\mathcal{S}_0$ is large, since the majority of synthetic records are not semantically related to the target record x. To achieve a balance between robustness and sensitivity, we introduce the "Top-3 Cosine Similarity" feature, which computes the average of the three highest cosine similarity scores between a real record x and all synthetic texts in $\mathcal{S}_0$.

For a target record $x \in \mathcal{R}_{\text{A}}$, we define the Top-3 Cosine Similarity feature:

$$\theta_{\text{top3}}(x) = \frac{1}{3} \sum_{i=1}^{3} \max_{s \in \mathcal{S}_0}(i)\ \text{sim}(x, s) \tag{7}$$

where $\max_{s \in \mathcal{S}_0}(i)\ \text{sim}(x, s)$ denotes the i-th largest cosine similarity between x and all $s \in \mathcal{S}_0$. This feature captures the average semantic proximity between a real record and its three most similar synthetic counterparts. It mitigates the over-sensitivity of $\theta_{\max}(x)$ to anomalously high but unrepresentative similarities, while also reducing the dilution effect inherent in $\theta_{\text{mean}}(x)$ for large synthetic corpora. It is expected to provide a more balanced and reliable indicator of membership influence, particularly suitable for dynamic few-shot ICL settings, where the association between a Member record and its generated synthetic texts can be inherently weak or dispersed across multiple prompts.

3.3 Evaluation Metric

The F1 score, defined as the harmonic mean of precision and recall, provides a balanced measure of a model's predictive performance. Precision represents the proportion of true positive predictions among all positive predictions, while recall denotes the proportion of true positives among all actual positives. By combining these two metrics, the F1 score captures the trade-off between precision and recall, making it particularly valuable in scenarios characterized by class imbalance, where minimizing both false positives and false negatives is equally important. As such, the F1 score is widely employed in domains such as information retrieval [36–38], medical diagnosis [39–42], and fraud detection [43–45], where the costs of misclassification are high. Formally, the F1 score is expressed as

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.$$

In the context of clinical synthetic dataset MIAs, where the adversary aims to determine whether a target record was used during synthetic data generation, the F1 score provides a particularly suitable evaluation metric. Clinical synthetic dataset MIAs often involve inherent class imbalance, as the proportion of member to non-member samples is typically skewed, making accuracy an unreliable measure. By

taking the harmonic mean of precision and recall, the F1 score offers a balanced evaluation of two critical risks: false positives, where non-members are incorrectly classified as members and the attack's success is overstated, and false negatives, where true members are overlooked and the attack's threat is understated. This balanced treatment of errors enables a realistic assessment of an adversary's effectiveness, capturing their ability to infer membership without disproportionately emphasizing one type of error. A high F1 score therefore indicates that the attack achieves an effective balance between precision and recall, which is essential for objective comparisons of attack methodologies and for evaluating the membership disclosure risks associated with releasing synthetic datasets. Accordingly, the F1 score has become a standard metric in assessing membership disclosure risks [14, 46–48].

However, the F1 metric is inherently dependent on data distribution, as both precision and recall are directly influenced by the relative proportions of positive and negative samples; consequently, the common assumption of a balanced Member/Non-Member ratio (i.e., 1:1) in much of the literature can lead to inaccurate estimates of membership disclosure risk [18]. In disclosure risk estimation, particularly in clinical practice, evaluation datasets are expected to reflect the natural distribution of the population from which they are drawn, often requiring the specification of population size values to ensure meaningful contextualization [18, 49–51]. Accordingly, to obtain valid assessments of membership disclosure in synthetic datasets using the F1 score, the simulated adversarial dataset must maintain the expected proportion of Member records as specified in Equation (1), thereby mimicking the sampling behavior of an adversary operating on the same population as the real dataset.

While a correct F1 score serves as a balanced indicator of an adversary's success in membership inference, it is insufficient as a universal judge of whether a synthetic dataset can be considered membership-preserving, primarily due to its intrinsic dependence on dataset distribution. In particular, an F1 score threshold that indicates acceptable privacy in one population may not be transferable to another, limiting its interpretability across heterogeneous datasets. Metric M an interpretable adjustment of the F1 score [18] that reduces distributional sensitivity while retaining the intuitive interpretability of F1:

$$M = \frac{F1 - F_{\max}}{1 - F_{\max}},$$

where $F_{\max}$ denotes the F1 score achieved by the naive adversarial strategy of classifying all records as members, which yields the maximum possible baseline performance without any background knowledge of the real dataset. By subtracting this baseline, M measures the improvement of an adversary over the strongest naive strategy, thereby mitigating the influence of data distribution on the evaluation. Consequently, M provides a more consistent and interpretable basis for comparing membership disclosure risks across different synthetic datasets. Following established practice [18], we adopt a threshold of $M \leq 0.2$, whereby a synthetic dataset is deemed membership-preserving if the adversary's disclosure performance reflects no more than an approximate 20% improvement over the naive baseline.

4 Experiments

In this section, we present the experimental setup used to evaluate the membership disclosure risk within the generative pipeline of synthetic clinical text produced via k-shot ICL, where $k \in \{1, 2, 3\}$. The evaluation is conducted from the perspective of data custodians, employing similarity-based MIAs. Section 4.1 details the construction of six synthetic text datasets $\mathcal{S}_0^k$ under both dynamic and static ICL settings, as well as the sampling strategy for the adversarial real-text sets. UltraMedical-70B [52] is employed for synthetic note generation, with a temperature of 0.2 applied. Section 4.2 describes the text-only MIA experiments conducted on these six synthetic datasets.

4.1 Dataset Preparation

MIMIC-IV-Note [53, 54] is a collection of deidentified free-text clinical notes linked to patients in the MIMIC-IV clinical database [54–56], a contemporary electronic health record (EHR) dataset encompassing hospital admissions from 2008 to 2019. It contains 331,794 deidentified discharge summaries from 145,915 patients admitted to the hospital and emergency department at the Beth Israel Deaconess Medical Center in Boston, MA, USA. All notes have been stripped of protected health information (PHI) in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor provision. The notes are linkable to MIMIC-IV, providing valuable clinical context.

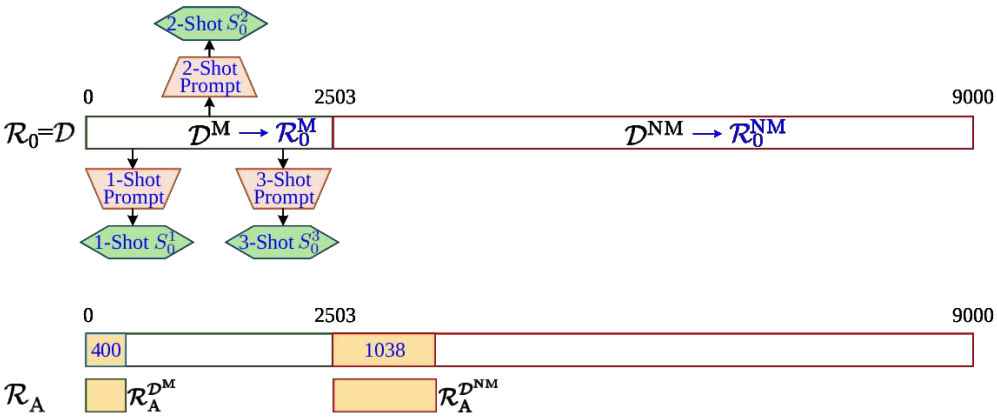


Fig. 1 **Dynamic** ICL Data Preparation of $\mathcal{R}_0$, $\mathcal{S}_0^k$ and $\mathcal{R}_{\mathrm{A}}$. The source dataset $\mathcal{D}$ consists of 9,000 oncology-related discharge summaries. It is partitioned into a Member set $\mathcal{D}^{\mathrm{M}}$ and a Non-Member set $\mathcal{D}^{\mathrm{NM}}$. The Member set $\mathcal{D}^{\mathrm{M}}$ serves as the source pool of real notes, i.e., $\mathcal{R}_0^{\mathrm{M}}$, for generating $\mathcal{S}_0^k$, while $\mathcal{D}^{\mathrm{NM}}$ serves as $\mathcal{R}_0^{\mathrm{NM}}$. The unlabeled adversarial subsets $\mathcal{R}_{\mathrm{A}}^{\mathcal{D}^{\mathrm{M}}}$ and $\mathcal{R}_{\mathrm{A}}^{\mathcal{D}^{\mathrm{NM}}}$ are randomly sampled (1,000 times to obtain average MIA performance) from $\mathcal{D}^{\mathrm{M}}$ and $\mathcal{D}^{\mathrm{NM}}$, respectively, as indicated by their superscripts. We did not use M and NM as their superscripts, as these subsets are unlabeled and their true membership info is used solely for MIA evaluations rather than for the MIAs themselves. This partitioning strategy follows Equation (1), where $|\mathcal{P}| = 32{,}362$, $|\mathcal{R}_0| = |\mathcal{D}| = 9{,}000$, $|\mathcal{D}^{\mathrm{M}}| = 2{,}503$, $|\mathcal{R}_{\mathrm{A}}| = 1{,}438$, $|\mathcal{R}_{\mathrm{A}}^{\mathcal{D}^{\mathrm{M}}}| = 400$. $|\mathcal{S}_0^1| = |\mathcal{S}_0^2| = |\mathcal{S}_0^3| = 2{,}503$.

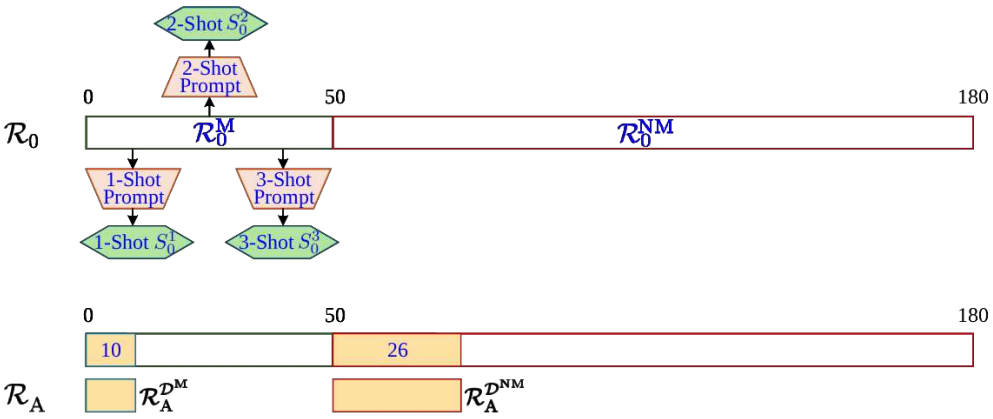


Fig. 2 **Static** ICL Data Preparation of $\mathcal{R}_0$, $\mathcal{S}_0^k$, **and** $\mathcal{R}_{\mathrm{A}}$. The subsets $\mathcal{R}_0^{\mathrm{M}}$ and $\mathcal{R}_0^{\mathrm{NM}}$ are randomly sampled from $\mathcal{D}^{\mathrm{M}}$ and $\mathcal{D}^{\mathrm{NM}}$, respectively. Then, $\mathcal{R}_{\mathrm{A}}^{\mathcal{D}^{\mathrm{M}}}$ and $\mathcal{R}_{\mathrm{A}}^{\mathcal{D}^{\mathrm{NM}}}$ are randomly sampled from $\mathcal{R}_0^{\mathrm{M}}$ and $\mathcal{R}_0^{\mathrm{NM}}$, respectively. Note that the random sampling for $\mathcal{R}_0$ is performed only once and then fixed, whereas $\mathcal{R}_{\mathrm{A}}$ is resampled 1,000 times to obtain the average MIA performance. $|\mathcal{R}_0| = 180$, $|\mathcal{R}_0^{\mathrm{M}}| = 50$, $|\mathcal{R}_{\mathrm{A}}| = 36$, and $|\mathcal{R}_{\mathrm{A}}^{\mathcal{D}^{\mathrm{M}}}| = 10$. $|\mathcal{S}_0^1| = |\mathcal{S}_0^2| = |\mathcal{S}_0^3| = 2{,}500$, where each unique k-shot prompt is used 50 times to generate 50 corresponding synthetic notes.

MIMIC-IV-Ext-BHC [54, 57, 58] is a curated collection of preprocessed and labeled clinical notes derived from the MIMIC-IV-Note database, with a specific focus on discharge summaries. The dataset comprises 270,033 meticulously cleaned and standardized clinical notes, with an average token length of 2,267, making it suitable for machine learning (ML) applications. The preprocessing pipeline employs regular expressions to address common issues in raw clinical text—such as special characters, extraneous whitespace, inconsistent formatting, and irrelevant content—resulting in a high-quality structured dataset. Clinical note sections are clearly delineated using appropriate headings, preserving the semantic structure of each document. These notes provide detailed narratives of patient hospitalizations, including sections such as medical history, treatments, and discharge instructions.

In this study, we direct our attention to the specialized and clinically significant field of oncology. To construct a deidentified real oncology-specific free-text dataset, we filtered the preprocessed discharge summaries by keyword search for "oncology", yielding 9,000 oncology-related note records from Boston. This dataset, denoted as $\mathcal{D}$, served as the source pool for sampling the sub-datasets $\mathcal{R}_0$ and $\mathcal{R}_{\mathrm{A}}$, as illustrated in **Fig. 1**.

In 2018, Boston had a population of 695,925 residents, consisting of 333,815 males and 362,111 females [59]. According to the Health of Boston 2023: The Cancer Report [60], the age-adjusted incidence rate of all invasive cancers was 461.2 cases per 100,000 male residents and 387.3 cases per 100,000 female residents. Applying these rates to the respective male and female populations yields an estimated 1,540 cases among males and 1,402 cases among females, resulting in a total of approximately 2,942 cancer cases in Boston during 2018. Based on the 2018 cancer incidence estimate, we project the total number of cancer patients in Boston between 2008 and

2019. Assuming a consistent population size and stable cancer incidence rate throughout this period, the approximated number of cases over the 11 years covered by the MIMIC-IV database is $2{,}942 \times 11$. In other words, we approximate the population of cancer patients in Boston across the 11-year period as $|\mathcal{P}| = 32{,}362$.

Fig. 1 and **Fig. 2** illustrate the dataset preparation for MIAs under dynamic ICL and static ICL, respectively. To ensure accurate F1 scores as well as their corresponding interpretable metric M for estimating membership disclosure risk in both MIAs, our data sampling strictly adheres to Equation (1):

$$\frac{|\mathcal{R}_{\mathrm{A}}^{\mathcal{D}^{\mathrm{M}}}|}{|\mathcal{R}_{\mathrm{A}}|} = \frac{|\mathcal{R}_0^{\mathrm{M}}|}{|\mathcal{R}_0|} = \frac{\mathcal{D}^{\mathrm{M}}}{\mathcal{D}} = \frac{\mathcal{D}}{\mathcal{P}} = 27.81\%.$$

This proportion reflects both the operational reality of synthetic text generation and the statistical requirements of membership inference evaluation.

We adhere to the following rules when sampling k-shot real notes during synthetic note generation:

1. For all k-shot ICL synthetic set generations, where $k \in \{1, 2, 3\}$, and for both the dynamic and static ICL paradigms, the number of unique prompts—that is, the size of the index tuple set $\mathcal{T}^k$ (where each $t^k \in \mathcal{T}^k$)—is always equal to the size of the Member real note set:
$$|\mathcal{T}^k| = |\mathcal{R}_0^{\mathrm{M}}|.$$
2. For each k-shot sampling, consider the jth unique index tuple $t_j^k = (i_{j,1}, \ldots, i_{j,k})$, where $1 \leq j \leq |\mathcal{T}^k|$. The following constraints apply to the construction of each t_j^k:
   (a) Set $i_{j,1} = j$. In other words, the first real note used as the in-context learning example for the jth prompt is deterministically chosen to be the jth real note from $\mathcal{R}_0^{\mathrm{M}}$.
   (b) For $k \in \{2, 3\}$, remove the index j from the integer set $\{1, \ldots, |\mathcal{R}_0^{\mathrm{M}}|\}$ to obtain a subset of size $|\mathcal{R}_0^{\mathrm{M}}| - 1$. Then, randomly sample $k - 1$ distinct indices *without replacement* from this subset to serve as the remaining indices $(i_{j,2}, \ldots, i_{j,k})$.

These real-note sampling rules ensure that:

- Each index tuple $t^k = (i_1, \ldots, i_k)$ for $k \in \{1, 2, 3\}$ is unique.
- The total number of possible index tuple combinations of k-shot is given by
$$\binom{|\mathcal{R}_0^{\mathrm{M}}|}{k} = \frac{|\mathcal{R}_0^{\mathrm{M}}|!}{k!\,(|\mathcal{R}_0^{\mathrm{M}}| - k)!}.$$
  For $k > 1$, this number is far greater than $|\mathcal{R}_0^{\mathrm{M}}|$ and may lead to potential bias in the evaluation results. To mitigate this bias, each index tuple t^k is sampled in a randomized manner.
- Every real note in $\mathcal{R}_0^{\mathrm{M}}$ is exposed to the target LLM at least once, thereby guaranteeing its inclusion as a member instance.

4.2 Experimental Setups

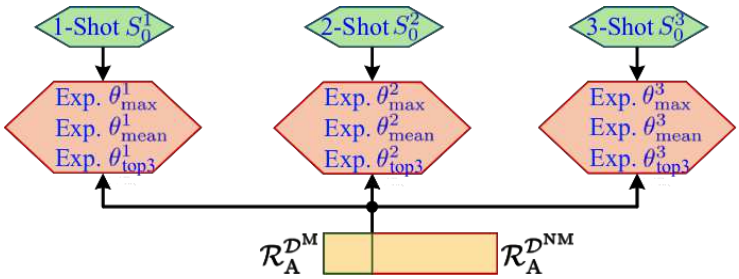


Fig. 3 Similarity-based MIAs, as introduced in Section 3.2, against oncology synthetic notes generated via k-shot ICL, where $k \in \{1, 2, 3\}$. This figure applies to both the dynamic and static ICL paradigms. The datasets $\mathcal{S}_0^k$, $\mathcal{R}_{\mathrm{A}}^{\mathcal{D}^{\mathrm{M}}}$ and $\mathcal{R}_{\mathrm{A}}^{\mathcal{D}^{\mathrm{NM}}}$ in the dynamic and static ICL experiments correspond to **Fig. 1** and **Fig. 2**, respectively. In each paradigm, for every synthetic set $\mathcal{S}_0^k$, three MIA experiments are conducted using $\theta^k_{\max}$, θ^k_{mean}, and θ^k_{top3}, respectively.

As illustrated in **Fig. 3**, the simulated adversary is assumed to possess access to the complete set of synthetic texts $\mathcal{S}_0^k$ generated under k-shot ICL, where $k \in \{1, 2, 3\}$, and to a small pool of unlabeled real records $\mathcal{R}_{\mathrm{A}}$. The evaluation procedure follows the similarity-based MIA method described in Section 3.2. For each real record $x \in \mathcal{R}_{\mathrm{A}}$, a dataset-level scalar similarity feature θ is computed, quantifying the semantic proximity between x and the synthetic dataset $\mathcal{S}_0^k$ in the sentence embedding space. Clinical-ModernBERT [61] is employed for generating sentence embeddings.

Three feature formulations are considered: (1) the maximum cosine similarity $\theta^k_{\max}$, which captures the strongest semantic connection between a real record and any synthetic instance, (2) the mean cosine similarity θ^k_{mean}, which aggregates similarities across the entire synthetic corpus $\mathcal{S}_0^k$ to characterize the record's overall alignment with the distribution of $\mathcal{S}_0^k$, and (3) the mean of top-3 maximum cosine similarity θ^k_{top3}, which is expected to retain the advantage of $\theta^k_{\max}$ and θ^k_{mean} and be more reliable.

Once the feature values θ are computed for all records in $\mathcal{R}_{\mathrm{A}}$, membership prediction is performed using an unsupervised Gaussian Mixture Model (GMM) clustering algorithm. The GMM partitions the feature distribution into two clusters, with the cluster exhibiting the higher average similarity score inferred as corresponding to the Member class and the other to the Non-Member class.

5 Results and Discussion

Table 1 reports the membership disclosure evaluation results across six synthetic corpora generated by static and dynamic k-shot ICL ($k \in \{1, 2, 3\}$), using three similarity-based features (Max, Mean, Top-3) and unsupervised clustering for text-only MIAs. We discuss trends along three axes: (i) increasing k within static ICL; (ii) increasing k within dynamic ICL; and (iii) static versus dynamic ICL at fixed k. We interpret performance using both F1 and the interpretable improvement-over-baseline

Experiments	Similarity	F1 (average +- std)	M (average +- std)
Static 1-Shot	Max	0.8991 +- 0.1754	0.8215 +- 0.3103
	Mean	0.7448 +- 0.1706	0.5485 +- 0.3018
	Top-3	0.8946 +- 0.1759	0.8136 +- 0.3112
Static 2-Shot	Max	0.5040 +- 0.1348	0.1225 +- 0.2384
	Mean	0.5187 +- 0.1465	0.1484 +- 0.2592
	Top-3	0.5012 +- 0.1310	0.1176 +- 0.2317
Static 3-Shot	Max	0.4720 +- 0.0685	**0.0659 +- 0.1212**
	Mean	0.5358 +- 0.1679	0.1787 +- 0.2970
	Top-3	0.4812 +- 0.0958	0.0822 +- 0.1695
Dynamic 1-Shot	Max	0.4339 +- 0.0024	-0.0024 +- 0.0042
	Mean	0.4283 +- 0.0058	-0.0123 +- 0.0103
	Top-3	0.4335 +- 0.0025	-0.0031 +- 0.0045
Dynamic 2-Shot	Max	0.4337 +- 0.0025	-0.0027 +- 0.0044
	Mean	0.4277 +- 0.0057	-0.0134 +- 0.0100
	Top-3	0.4333 +- 0.0029	-0.0034 +- 0.0051
Dynamic 3-Shot	Max	0.4338 +- 0.0025	-0.0026 +- 0.0044
	Mean	0.4279 +- 0.0057	-0.0130 +- 0.0101
	Top-3	0.4335 +- 0.0026	-0.0031 +- 0.0047

Table 1 Membership Disclosure Risk Evaluation Results. In each experiment, $\mathcal{R}_{\text{A}}$ is randomly sampled 1,000 times to get the average MIA results.

metric M (Section 3.3), where lower M indicates weaker adversarial advantage beyond the naive maximum membership strategy.

Static k-shot ICL: k is important

Under static 1-shot prompting, the Max and Top-3 features exhibit very high attack success (average value F1 ≈ 0.89, $M \approx 0.82$), while Mean is moderately strong (F1 ≈ 0.74, $M \approx 0.55$). The small gap between Max and Top-3 suggests that the dominant signal is largely captured by the highest sentence-embedding similarity between a target real note and the synthetic corpus, and averaging over the few largest embedding similarities does not substantially attenuate it. The elevated M values (well above the $M \leq 0.2$ threshold) indicate substantial membership leakage: fixed reuse of a single exemplar per prompt imprints strong semantic proximity that is detectable by similarity-based MIAs.

Moving to static 2-shot, the Max and Top-3 features fall sharply toward average value $M \approx 0.12$ and Mean falls to $M \approx 0.15$, suggesting that adding a second exemplar disrupts exemplar-specific traces sufficiently to impede text-only MIAs. For static 3-shot, while the Mean rises slightly ($M \approx 0.18$), both Max and Top-3 decline further ($M \approx 0.07$–0.08), entering the membership-preserving regime under the $M \leq 0.2$ criterion, even when accounting for the standard deviation. The Mean feature's uptick likely reflects its sensitivity to the overall corpus centroid: as k increases, prompts may induce a more coherent global style in the synthetic corpus, improving centroid alignment for both member and some non-member records, which can inflate Mean-based separability in certain resamples. Overall, within static ICL, increasing k effectively attenuates membership signals, with Max and Top-3 behaving as more robust indicators thank Mean.

Dynamic k-shot ICL: membership-preserving

Across dynamic 1-, 2-, and 3-shot settings, the F1 scores concentrate around 0.43 and the corresponding M values are approximately 0 or slightly negative, with very small variance across the 1,000 resamples. These M values are far below the adoption threshold of $M \leq 0.2$ commonly used to judge membership preservation. Given the calibrated member proportion of 27.81% (Section 4.1), these values imply that an adversary with text-only access gains negligible advantage from semantic similarity. This pattern holds for Max, Mean, and Top-3 features alike, suggesting that, under dynamic prompting, semantic influence from any particular in-context exemplar is weakly expressed in the released synthetic corpus. These findings support the intuition that dynamic exemplar selection disperses and attenuates prompt-specific signals across outputs, blurring membership cues.

Static versus Dynamic

Comparing static and dynamic regimes at matched k highlights a clear pattern: dynamic ICL dramatically suppresses membership disclosure relative to static ICL. The contrast is most striking at 1-shot: static 1-shot produces $M \approx 0.82$ (Max/Top-3) versus $M \approx -0.003$ for dynamic 1-shot. This gap persists, albeit smaller, at higher k. The core mechanism is prompt diversity: dynamic selection reduces the consistency and salience of any single exemplar's semantic signature across the released corpus, thereby diffusing the membership cues that enable text-only attacks. In static prompts, repeated exposure of the same exemplar(s) imprints a stronger, more learnable similarity pattern, especially when k is small.

Comparison across similarity features

- Max vs. Top-3: Both Max and Top-3 perform similarly and strongly, with Top-3 slightly smoothing out extreme outliers but preserving sensitivity to repeated alignment.
- Mean: Mean is more sensitive to corpus-level centroid effects. In static 3-shot, Mean yields a higher M than Max/Top-3, reflecting that distribution-level alignment can sometimes reintroduce weak separability. For practitioners, centroid-based scores can complement nearest-neighbor scores to detect style-level versus instance-level leakage.

Implications for custodial policy and evaluation

The results support two operational recommendations for custodians considering the release of synthetic clinical corpora generated by ICL:

1. Prefer dynamic over static few-shot prompting, even with small k, to substantially reduce membership leakage; and
2. If static prompting is unavoidable, use at least 2–3 examples and ensure sufficient exemplar diversity to dilute membership signals. In both cases, custodians should perform corpus-level auditing using distribution-aware evaluation (Section 3.1) and report both F1 and M to contextualize risk relative to a naive baseline.

Limitations and scope

Our attacks are text-only and rely on sentence-level embeddings and unsupervised clustering; stronger adversaries with prompt access, confidence scores, or black-box probing could increase attack power. We evaluate only oncology discharge summaries within a single institution. Although this focuses the study on a clinically relevant subset with structured narratives, future work should assess generalizability across note types, institutions, and languages, and examine additional backbone LLMs and embedding models. Finally, while M improves interpretability across populations, deployment decisions should consider institutional risk thresholds and regulatory constraints.

6 Conclusion

We formalized membership in dynamic few-shot ICL, and proposed a practical, population-aware auditing framework for membership disclosure risk in synthetic clinical text generated via few-shot in-context learning. Our approach reflects the data custodian's constraints by assuming text-only adversaries with access to the released synthetic corpus and a small auxiliary real-text pool. We introduced three semantic similarity features based on sentence embeddings and evaluated them against synthetic corpora produced by static and dynamic k-shot ICL.

Our experiments on oncology discharge summaries demonstrate that dynamic few-shot ICL yields membership disclosure performance near the naive baseline across features and $k \in \{1, 2, 3\}$, with $M \approx 0$, indicating that released synthetic notes are membership-preserving under our evaluation protocol. In contrast, static 1-shot ICL exhibits a high disclosure risk, suggesting that repeated reuse of fixed exemplars leaves detectable membership traces even under text-only membership inference attacks (MIAs). This risk is substantially mitigated as k increases to $2, 3$. Overall, dynamic few-shot ICL markedly reduces membership disclosure risk compared to static prompting.

These findings have immediate implications for safe synthetic data release practices in clinical NLP: when using LLMs to synthesize notes, dynamic exemplar selection should be preferred over static prompting, and if static prompting is required, adopt multi-shot prompts. Beyond these results, our population-aware partitioning and interpretable metric M offer a replicable template for custodial auditing that avoids common pitfalls stemming from miscalibrated member ratios.

Future work will broaden the evaluation to additional LLMs, clinical note types, institutions, and languages; explore stronger black-box probes and hybrid attacks combining similarity with generative likelihood signals; and assess complementary mitigations such as controllable paraphrasing, differentially private prompt retrieval, and content filtering. Together, these directions aim to provide actionable, quantitative guarantees that synthetic clinical corpora produced by ICL-enabled LLMs can be shared responsibly while preserving patient privacy.

7 Declarations

Funding

Khaled El Emam received funding from Selfiie (www.selfiie.com). The specific funding details will be provided upon request.

Competing interest

The authors declare that they have no competing interests as defined by BMC, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

Author contribution

Sen Li and Fida K. Dankar wrote the manuscript text. Sen Li prepared Figures 1–3. All authors reviewed the manuscript.

Data availability

This study used MIMIC-IV-Note, a collection of deidentified free-text clinical notes linked to patients in the MIMIC-IV clinical database (covering hospital admissions from 2008 to 2019). Access to MIMIC-IV/MIMIC-IV-Note is available via PhysioNet under its data use requirements. Additional details are provided in Section 4.1 (Data Preparation).

Ethics declaration

This study analyzed deidentified clinical notes from the MIMIC-IV-Note resource. Ethics approval and consent procedures are governed by the MIMIC-IV/PhysioNet data resource; no additional ethics approval was required for the authors' analysis of deidentified data.

References

[1] Duan, M., Suri, A., Mireshghallah, N., Min, S., Shi, W., Zettlemoyer, L.S., Tsvetkov, Y., Choi, Y., Evans, D., Hajishirzi, H.: Do membership inference attacks work on large language models? ArXiv **abs/2402.07841** (2024)

[2] Zhang, Z., Yan, C., Malin, B.A.: Membership inference attacks against synthetic health data. Journal of biomedical informatics, 103977 (2021)

[3] German, E., Samira, D., Elovici, Y., Shabtai, A.: Mia-ept: Membership inference attack via error prediction for tabular data. (2025). https://api.semanticscholar.org/CorpusID:281325097

[4] Miranda, L., Dunstan, J., Toro, M., Olmedo, F., Melo, F.: Evaluating privacy risks in synthetic clinical text generation in spanish. https://api.semanticscholar.org/CorpusID:279242041

[5] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners (2020) arXiv:2005.14165

[6] Liu, X., McDuff, D.J., Kovács, G., Galatzer-Levy, I.R., Sunshine, J., Zhan, J., Poh, M.-Z., Liao, S., Achille, P.D., Patel, S.N.: Large language models are few-shot health learners. ArXiv **abs/2305.15525** (2023)

[7] Duan, H., Dziedzic, A., Yaghini, M., Papernot, N., Boenisch, F.: On the privacy risk of in-context learning. ArXiv **abs/2411.10512** (2024)

[8] Giorgi, J., Toma, A., Xie, R., Chen, S.S., An, K.R., Zheng, G.X., Wang, B.: Wanglab at mediqa-chat 2023: Clinical note generation from doctor-patient conversations using large language models. In: Clinical Natural Language Processing Workshop (2023). https://api.semanticscholar.org/CorpusID:259076043

[9] Yashwanth, Y.S., Shettar, R.: Zero and few short learning using large language models for de-identification of medical records. IEEE Access **12**, 110385–110393 (2024)

[10] Wen, R., Li, Z., Backes, M., Zhang, Y.: Membership inference attacks against in-context learning. Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security (2024)

[11] Hamid, R., Brohi, S.N.: A review of large language models in healthcare: Taxonomy, threats, vulnerabilities, and framework. Big Data Cogn. Comput. **8**, 161 (2024)

[12] Hu, H., Salcic, Z.A., Sun, L., Dobbie, G., Yu, P., Zhang, X.: Membership inference attacks on machine learning: A survey. ACM Computing Surveys (CSUR) **54**, 1–37 (2021)

[13] Smith, V., Shamsabadi, A.S., Ashurst, C., Weller, A.: Identifying and mitigating privacy risks stemming from language models: A survey. ArXiv **abs/2310.01424** (2023)

[14] Niu, J., Liu, P., Zhu, X., Shen, K., Wang, Y., Chi, H., Shen, Y., Jiang, X., Ma, J., Zhang, Y.: A survey on membership inference attacks and defenses in machine learning. Journal of Information and Intelligence (2024)

[15] Sablayrolles, A., Douze, M., Schmid, C., Ollivier, Y., Jégou, H.: White-box vs black-box: Bayes optimal strategies for membership inference. In: International Conference on Machine Learning (2019). https://api.semanticscholar.org/CorpusID:174799799

[16] Watson, L., Guo, C., Cormode, G., Sablayrolles, A.: On the importance of difficulty calibration in membership inference attacks. ArXiv **abs/2111.08440** (2021)

[17] Puerto, H., Gubri, M., Yun, S., Oh, S.J.: Scaling up membership inference: When and how attacks succeed on large language models. ArXiv **abs/2411.00154** (2024)

[18] Emam, K.E., Mosquera, L., Fang, X.: Validating a membership disclosure metric for synthetic health data. JAMIA Open **5** (2022)

[19] Sun, R., Hu, H., Luo, W., Zhang, Z., Zhang, Y., Yuan, H., Zhang, L.Y.: When better features mean greater risks: The performance-privacy trade-off in contrastive learning. ArXiv **abs/2506.05743** (2025)

[20] Mendelevitch, O., Lesh, M.D.: Fidelity and privacy of synthetic medical data. ArXiv **abs/2101.08658** (2021)

[21] Zhang, Z., Yan, C., Mesa, D.A., Sun, J., Malin, B.A.: Ensuring electronic medical record simulation through better training, modeling, and evaluation. Journal of the American Medical Informatics Association : JAMIA (2019)

[22] Yan, C., Zhang, Z., Nyemba, S., Malin, B.A.: Generating electronic health records with multiple data types and constraints. Annual Symposium proceedings. AMIA Symposium **2020**, 1335–1344 (2020)

[23] Gonçalves, A.R., Ray, P., Soper, B.C., Stevens, J.L., Coyle, L., Sales, A.P.: Generation and evaluation of synthetic patient data. BMC Medical Research Methodology **20** (2020)

[24] Chen, D., Yu, N., Zhang, Y., Fritz, M.: Gan-leaks: A taxonomy of membership inference attacks against generative models. Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security (2019)

[25] Shi, H., Tu, O., Wang, A.: Learning-based difficulty calibration for enhanced membership inference attacks. 2024 IEEE 9th European Symposium on Security and Privacy (EuroS&P), 62–77 (2024)

[26] Sarkar, A.R., Chuang, Y.-S., Mohammed, N., Jiang, X.: De-identification is not enough: a comparison between de-identified and synthetic clinical notes. Scientific Reports **14**(1), 29669 (2024) https://doi.org/10.1038/s41598-024-81170-y

[27] Bai, Y., Pradhan, G., Tobaben, M., Honkela, A.: Empirical comparison of membership inference attacks in deep transfer learning. In: ICML 2025 Workshop on Reliable and Responsible Foundation Models (2025). https://openreview.net/forum?id=eR8mAEt1jT

[28] Dubi'nski, J., Kowalczuk, A.D., Pawlak, S., Rokita, P., Trzci'nski, T., Morawiecki, P.: Towards more realistic membership inference attacks on large diffusion models. 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 4848–4857 (2023)

[29] Xie, S.M., Raghunathan, A., Liang, P., Ma, T.: An explanation of in-context learning as implicit bayesian inference. ArXiv **abs/2111.02080** (2021)

[30] Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., Zettlemoyer, L.: Rethinking the role of demonstrations: What makes in-context learning work? ArXiv **abs/2202.12837** (2022)

[31] Zhou, W., Jiang, Y.E., Cotterell, R., Sachan, M.: Efficient prompting via dynamic in-context learning. ArXiv **abs/2305.11170** (2023)

[32] Dai, D., Sun, Y., Dong, L., Hao, Y., Ma, S., Sui, Z., Wei, F.: Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. (2022). https://api.semanticscholar.org/CorpusID:258686544

[33] Li, Y., Liu, G., Wang, C., Yang, Y.: Generating is believing: Membership inference attacks against retrieval-augmented generation. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (2024). https://api.semanticscholar.org/CorpusID:272911477

[34] Anderson, M., Amit, G., Goldsteen, A.: Is my data in your retrieval database? membership inference attacks against retrieval augmented generation. ArXiv **abs/2405.20446** (2024)

[35] Yang, P., Yin, J., Zheng, H., Bai, X., Wang, H., Sun, Y., Li, X., Wang, S., Huang, Y., Qi, T.: Mrm: Black-box membership inference attacks against multimodal rag systems. ArXiv **abs/2506.07399** (2025)

[36] Hripcsak, G., Rothschild, A.S.: Agreement, the f-measure, and reliability in information retrieval. Journal of the American Medical Informatics Association **12**(3), 296–298 (2005) https://doi.org/10.1197/jamia.M1733

[37] Derczynski, L.: Complementarity, f-score, and nlp evaluation. In: International Conference on Language Resources and Evaluation (2016). https://api.semanticscholar.org/CorpusID:1414144

[38] Christen, P., Hand, D.J., Kirielle, N.: A review of the f-measure: Its history, properties, criticism, and alternatives. ACM Computing Surveys **56**, 1–24 (2023)

[39] Li, Y., Konomis, D., Neubig, G., Xie, P., Cheng, C., Xing, E.P.: Convolutional neural networks for medical diagnosis from admission notes. ArXiv **abs/1712.02768** (2017)

[40] Hicks, S.A., Strùke, I., Thambawita, V.L., Hammou, M., Halvorsen, P., Riegler, M.A., Parasa, S.: On evaluation metrics for medical applications of artificial intelligence. Scientific Reports **12** (2021)

[41] Ferhi, L.A., Amar, M.B., Choubani, F., Bouallègue, R.: Enhancing diagnostic accuracy in symptom-based health checkers: a comprehensive machine learning approach with clinical vignettes and benchmarking. Frontiers in Artificial Intelligence **7** (2024)

[42] Ng, C., Sun, L., Tang, S.: X-ray-cot: Interpretable chest x-ray diagnosis with vision-language models via chain-of-thought reasoning. (2025). https://api.semanticscholar.org/CorpusID:280677334

[43] Zhu, M., Zhang, Y., Gong, Y., Xu, C., Xiang, Y.: Enhancing credit card fraud detection a neural network and smote integrated approach. ArXiv **abs/2405.00026** (2024)

[44] Chen, Y., Zhao, C., Xu, Y., Nie, C.: Year-over-year developments in financial fraud detection via deep learning: A systematic literature review. ArXiv **abs/2502.00201** (2025)

[45] Hayat, K., Magnier, B.: Data leakage and deceptive performance: A critical examination of credit card fraud detection methodologies. ArXiv **abs/2506.02703** (2025)

[46] Hui, B., Yang, Y., Yuan, H., Burlina, P., Gong, N.Z., Cao, Y.: Practical blind membership inference attack via differential comparisons. ArXiv **abs/2101.01341** (2021)

[47] Rezaei, S., Liu, X.: On the difficulty of membership inference attacks. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7888–7896 (2021)

[48] Ren, J., Chen, K., Chen, C., Sehwag, V., Xing, Y., Tang, J., Lyu, L.: Self-comparison for dataset-level membership inference in large (vision-)language model. Proceedings of the ACM on Web Conference 2025 (2024)

[49] Dankar, F.K., Emam, K.E., Neisa, A., Roffey, T.: Estimating the re-identification risk of clinical data sets. BMC Medical Informatics and Decision Making **12**, 66–66 (2012)

[50] Rocher, L., Hendrickx, J.M., Montjoye, Y.-A.: Estimating the success of re-identifications in incomplete datasets using generative models. Nature Communications **10** (2019)

[51] Jiang, Y., Mosquera, L., Jiang, B., Kong, L., Emam, K.E.: Measuring re-identification risk using a synthetic estimator to enable data sharing. PLoS ONE

17 (2022)

[52] Zhang, K., Zeng, S., Hua, E., Ding, N., Chen, Z., Ma, Z., Li, H., Cui, G., Qi, B., Zhu, X., Lv, X., Hu, J., Liu, Z., Zhou, B.: Ultramedical: Building specialized generalists in biomedicine. ArXiv **abs/2406.03949** (2024)

[53] Johnson, A., Pollard, T., Horng, S., Celi, L.A., Mark, R.: MIMIC-IV-Note. SCR_007345, RRID (2023). https://doi.org/10.13026/1n74-ne17

[54] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P.C., Mark, R., others, Stanley, H.E.: Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. circulation **101**, 23 (2000). RRID:SCR_007345

[55] Johnson, A., Bulgarelli, L., Pollard, T., Gow, B., Moody, B., Horng, S., Celi, L.A., Mark, R.: MIMIC-IV (version 3.1). PhysioNet. SCR_007345, RRID (2024). https://doi.org/10.13026/kpb9-mt58

[56] Johnson, A.E.W., Bulgarelli, L., Shen, L.e.a.: Mimic-iv, a freely accessible electronic health record dataset. Sci Data **10**, 1 (2023) https://doi.org/10.1038/s41597-022-01899-x

[57] Aali, A., Van Veen, D., Arefeen, Y., Hom, J., Bluethgen, C., Reis, E.P., Gatidis, S., Clifford, N., Daws, J., Tehrani, A., Kim, J., Chaudhari, A.: MIMIC-IV-Ext-BHC. SCR_007345, RRID (2025). https://doi.org/10.13026/5gte-bv70

[58] Aali, A., Van Veen, D., Arefeen, Y., Hom, J., Bluethgen, C., Reis, E.P., Gatidis, S., Clifford, N., Daws, J., Tehrani, A., Kim, J., Chaudhari, A.: A dataset and benchmark for hospital course summarization with adapted large language models. Journal of the American Medical Informatics Association (2024)

[59] Boston planning and development agency, research division (2020)

[60] Boston Public Health Commission, Health of Boston: The cancer report. Boston, Massachusetts **2023** (2023)

[61] Lee, S.A., Wu, A., Chiang, J.N.: Clinical modernbert: An efficient and long context encoder for biomedical text. arXiv preprint arXiv:2504.03964 (2025)