Privacy Leakage and Memorization in Fine-Tuned Clinical Language Models: A Controlled Study of Defenses and Backbone Choice on Clinical Narrative Transcriptions

Yassine Chahid[1], Anas Chahid[2], Ismail Chahid[3], Aissa Kerkour Elmiad[4]
ACSA, Mohammed First University, Oujda, 60000, Morocco[1]
SmartICT, Mohammed First University, Oujda, 60000, Morocco[2]
LARI, Mohammed First University, Oujda, 60000, Morocco[3,4]

***Abstract*—The increasing adoption of large language models (LLMs) and domain-adapted transformers in healthcare has created a new privacy challenge: fine-tuned models may memorize rare clinical strings and later reveal them through generation or scoring behavior. A controlled study of privacy leakage and memorization in clinical language models trained on narrative transcriptions is presented. A canary-based audit pipeline was instantiated on a 4,000-note subset of the Medical Transcriptions (MTSamples) corpus, with 40 synthetic secrets injected only into the training partition and evaluated using three complementary attack families: prompt extraction, exposure-style ranking, and reference-based membership inference. Two experiments are reported. Experiment I compares baseline fine-tuning, early stopping, and a conservative regularized training profile combining lower learning rate, higher weight decay, and partial layer freezing. Experiment II fixes the training protocol and compares DistilGPT2, GPT-2, and BioGPT. A clear privacy-utility tension was observed. In Experiment I, early stopping produced the best held-out language-model utility, whereas the combined regularized profile eliminated observed prompt leakage and reduced membership-inference strength, at the cost of worse perplexity. In Experiment II, stronger and more domain-specialized backbones achieved better clinical language modeling but also exhibited higher leakage and stronger membership-inference signals, with BioGPT yielding the strongest utility and the highest privacy risk under the evaluated attacks. These results indicate that privacy auditing should accompany utility evaluation in clinical LLM adaptation, and that backbone choice can materially affect memorization risk in this controlled setting.**

***Keywords*—Clinical language models; privacy leakage; memorization; membership inference; canary exposure; BioGPT; GPT-2; healthcare NLP; model auditing**

I. Introduction

Large language models are rapidly becoming part of healthcare AI pipelines, including clinical decision support, summarization of electronic health records (EHRs), documentation assistance, patient communication, and biomedical question answering [2], [3], [4], [5], [6]. Their appeal is clear: once adapted to clinical language, such models can absorb domain vocabulary, structure, and reasoning patterns that are costly to encode manually. At the same time, their deployment introduces an acute privacy question: *when a general or biomedical backbone is fine-tuned on clinical text, does it merely generalize, or does it also memorize rare sensitive sequences strongly enough to leak them later?*

This question is especially pressing in healthcare because patient narratives often contain sparse, low-frequency identifiers and clinically meaningful attributes. Even when direct identifiers are removed, combinations of age, disease code, specialty, procedural detail, and treatment narrative can become quasi-identifying. Responsible machine learning in medicine therefore requires more than utility evaluation alone; it requires explicit privacy auditing and leakage measurement [1], [7], [33], [8].

Privacy concerns in language models are not hypothetical. Work on unintended memorization, training-data extraction, and membership inference has shown that generative models can reveal exact or near-exact traces of the data seen during training [9], [10], [11], [12], [13], [14]. Recent findings further suggest that stronger or more domain-adapted models may become more data efficient but simultaneously more privacy sensitive [15], [16], [17], [19].

Clinical adaptation adds another layer of complexity. Biomedical and clinical language models such as BioBERT, PubMedBERT, GatorTron, Med-BERT, BEHRT, and BioGPT outperform general models on many healthcare tasks [26], [27], [29], [30], [31], [28]. However, stronger in-domain fit may also increase the probability of memorizing rare strings. The literature has discussed this possibility conceptually, but practical, reproducible demonstrations on clinical-style corpora are still limited, especially when both defenses and backbone choice are evaluated within the same study.

That gap is addressed here through a controlled dual-experiment study on clinical narrative transcriptions. The Medical Transcriptions (MTSamples) dataset is used as a multi-specialty corpus for clinical free text, and a canary-based audit protocol is designed in which synthetic secrets are inserted *only into the training split*. The resulting fine-tuned models are then tested for leakage through generation and for memorization through lower loss and higher ranking scores (Fig. 1).

Four contributions are made:

- A defense-oriented privacy audit. Baseline fine-tuning, early stopping, and a combined regularized training

profile are compared on the same controlled clinical fine-tuning task.

- A backbone-oriented privacy audit. DistilGPT2, GPT-2, and BioGPT are compared under the same dataset, split, canary set, and attack suite.
- A multi-attack evaluation protocol. Leakage is quantified with prompt-based extraction, candidate ranking with exposure-style scoring, and reference-based membership inference using ROC-AUC, PR-AUC, accuracy, precision, recall, F1, specificity, and balanced accuracy.
- A defensible privacy and utility analysis. The reported results indicate that stronger utility does not guarantee better privacy, and that backbone choice is itself a privacy-relevant design decision in clinical LLM fine-tuning.

II. Related Work

A. Healthcare Language Models and Clinical Adaptation

Clinical and biomedical transformer models have progressed rapidly, from early domain adaptation strategies to large-scale specialized pretraining. BioBERT improved biomedical text mining with domain-adapted contextual representations [26]. PubMedBERT showed that in-domain pretraining from scratch can outperform general-domain initialization in biomedical NLP [27]. Structured-clinical representation learning advanced through Med-BERT and BEHRT, which demonstrated that transformer-based pretraining on large EHR repositories benefits downstream prediction tasks [30], [31]. At larger scale, GatorTron and BioGPT extended this line to generative and clinical-free-text modeling [29], [28]. More recently, Med-PaLM-style systems and health system-scale language models have highlighted the emerging power of foundation models in healthcare [4], [5], [6], [41].

B. Privacy Risks in Machine Learning and Clinical AI

The healthcare literature has repeatedly emphasized that high utility alone is insufficient in medical AI. Responsible deployment requires privacy, governance, reproducibility, and harm-aware design [1], [2], [3]. In medical imaging and distributed healthcare learning, privacy-preserving federated methods and secure training protocols have become major themes [7], [32], [33], [34], [35], [8]. Yet privacy issues in generative clinical text models remain comparatively underexplored, especially under white-box or score-based audit settings.

C. Memorization, Training-Data Extraction, and Membership Inference

Membership inference emerged as a canonical privacy threat in machine learning through black-box attacks that infer whether a sample belonged to the training set [11]. In language models, memorization and extraction risks became especially visible with canary-based audits and direct generation attacks [9], [10]. Later work showed that deduplication, training dynamics, and neighborhood-based attacks strongly affect leakage behavior [15], [16], [12], [13], [17], [18], [19]. Recent work has extended the threat model to code models, context-aware membership inference, unlearning of personally identifiable information, and broad privacy-risk surveys for LLMs [20], [21], [22], [25], [23], [24].

D. Differential Privacy and Related Defenses

Differential privacy remains the most principled formal defense against memorization, beginning with DP-SGD [36]. Its use in language modeling and large-model fine-tuning is active but computationally expensive, and often introduces a non-trivial utility penalty [37], [38], [39]. Outside formal privacy, practical mitigations include data deduplication, regularization, early stopping, dropout, freezing submodules, prompt filtering, and unlearning [15], [16], [22]. However, these defenses are rarely compared side by side on clinical-style generative fine-tuning under the same controlled attack suite.

The study is positioned at this intersection: clinical-language fine-tuning, memorization-aware auditing, defense comparison, and backbone comparison, all grounded on the same clinical narrative corpus.

III. Materials and Methods

A. Dataset and Sampling Protocol

Both experiments were conducted on the Medical Transcriptions (MTSamples) dataset, available at https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions. The corpus contains 4,999 clinical-style transcriptions spanning 40 specialty categories. After preprocessing, 4,000 notes were sampled. The resulting splits followed an 80/10/10 partition: 3,200 training notes, 400 validation notes, and 400 test notes. For both experiments, exact duplicates were removed before splitting.

The selected corpus is appropriate for privacy auditing because it contains long-form narrative prose, specialty-specific vocabulary, structured reporting habits, and heterogeneous note styles. These properties are important because memorization risk is driven not only by raw dataset size, but also by the presence of rare phrasings, repeated templates, and clinically meaningful identifiers embedded in free text.

TABLE I. Dataset and Core Protocol Settings Used in Both Experiments

Setting	Value
Base corpus	Medical Transcriptions (MTSamples)
Prepared note count	4,000 sampled notes
Train/validation/test split	3,200 / 400 / 400
Sequence length	128 tokens
Batch size	2
Gradient accumulation	4
Injected canaries	40
Control canaries	40
Canary repetitions in training	2
Prompt families	4
Candidate pool for ranking attack	64 condition codes

B. Synthetic Canary Design

To measure memorization under controlled conditions, synthetic canaries of the form below were injected:

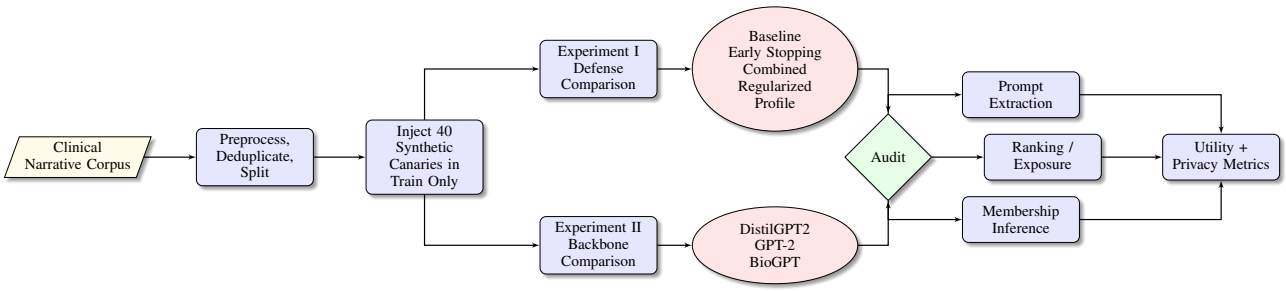


Fig. 1. Overall study design. A clinical narrative corpus is sampled, preprocessed, split, and augmented with train-only synthetic canaries. Experiment I compares defenses on one backbone, while Experiment II compares backbones under a fixed training protocol. All models are audited with prompt extraction, ranking and exposure analysis, and reference-based membership inference.

```
Patient identifier ZXQ-XXXX has condition
OMEGA-YY.
```

Each patient identifier and condition code was unique. The canaries were embedded in realistic clinical-note fragments and inserted *only into the training split*. An additional set of 40 control canaries was generated with the same syntactic structure but never injected into the model. This control set enabled membership inference and ranking-based discrimination between members and non-members.

C. Experiment I: Defense Comparison

Experiment I used DistilGPT2 as the backbone and compared three training settings:

- Baseline: 3 epochs, learning rate 5×10^{-5}, weight decay 0.01.
- Early stopping: up to 6 epochs with patience 2, same optimizer scale as baseline.
- Combined regularized profile: 4 epochs maximum, learning rate 2×10^{-5}, weight decay 0.05, warmup ratio 0.05, and freezing of the lower half of the transformer stack.

Because this third setting combines several interventions, the resulting privacy and utility changes should be interpreted at the profile level rather than attributed to any single component in isolation.

D. Experiment II: Backbone Comparison

Experiment II fixed the training protocol and compared three backbones:

- DistilGPT2 (81.9M parameters),
- GPT-2 (124.4M parameters),
- BioGPT (346.8M parameters).

All three models were trained for up to 4 epochs with learning rate 5×10^{-5}, weight decay 0.01, warmup ratio 0.03, and early stopping patience 2. This made the comparison as controlled as possible: only the backbone changed.

E. Experimental Architecture and Reproducibility Controls

Both notebooks implemented the same end-to-end experimental architecture. First, the corpus was downloaded, normalized, and partitioned into train, validation, and test splits with fixed seeds. Second, synthetic canaries were generated automatically and inserted only into the training split, while matched control canaries were retained as non-members. Third, a chosen causal-language-model backbone was fine-tuned under an explicitly declared profile. Fourth, the trained model was audited by three privacy-evaluation layers: prompt extraction, ranking-based memorization analysis, and reference-based membership inference. Finally, every run exported CSV summaries, plots, and serialized artifacts for downstream inspection and manuscript generation.

Several design choices strengthen scientific defensibility. The experiments used the same prompt templates across all runs, fixed train/validation/test boundaries, CPU-compatible batch sizes, explicit configuration blocks, persistent result files, and no post-hoc retuning of the privacy protocol after observing outcomes. In the backbone-comparison notebook, hyperparameters were held constant across backbones so that privacy differences could be interpreted with minimal optimization asymmetry. This improves comparability, although it does not eliminate configuration sensitivity. Reproducibility is further supported through fixed random seeds, persistent exported artifacts, and complete reporting of the principal training settings in Tables I, III, and IV. Software-version pinning, runtime-variance profiling across repeated executions, and memory-usage tracking were not analyzed in this study and should therefore be considered outside the present reproducibility envelope.

F. Attack-Suite Design

The attack suite was designed to cover multiple black-box interaction patterns rather than a single prompt template. Leakage can surface through chart-style lookups, incomplete clinical notes, routine continuation requests, or repeated model queries. For that reason, the notebooks tested several prompt families rather than one handcrafted query. Greedy decoding alone can underestimate risk, so beam-search and stochastic-sampling variants were also included. This mattered empirically: in Experiment I, some canaries only became exact leaks under sampled decoding, while greedy and beam decoding produced only partial continuations.

The ranking and membership attacks deepen the analysis beyond visible completions. Exposure-style ranking does not require a verbatim leak; it asks whether the true secret is scored unusually well relative to many plausible alternatives. Membership inference asks a related question: can an adversary infer whether a sensitive pattern was used during fine-tuning? In healthcare, even that signal can be consequential because it may reveal that a rare condition template, documentation style, or institution-specific phrase was present in the fine-tuning distribution. Taken together, the three attack families provide a stronger audit than prompt extraction alone.

G. Attack Suite

Three families of privacy attacks were used.

1) Prompt-based extraction: For each canary, four prompt families were evaluated:

- prefix completion,
- direct recall,
- partial note completion,
- chart lookup.

Each prompt was decoded with greedy decoding, beam search, and stochastic sampling. The prompt-based metrics were:

- exact-match recovery rate,
- condition recovery rate,
- patient-identifier recovery rate,
- partial-match rate,
- token recovery rate,
- average string similarity,
- leakage rate (fraction of canaries leaked at least once).

2) Ranking and exposure: Prompt attacks can underestimate memorization because a secret may be internally preferred without always appearing under one decoding strategy. Multiple candidate condition codes were therefore scored given a fixed patient identifier, and the true code was ranked among 64 candidates. Exposure-style memorization was computed as:

$$E = \log_2(N) - \log_2(r), \tag{1}$$

where, N is the candidate-pool size and r is the rank of the true secret. Higher exposure indicates stronger memorization.

3) Reference-based membership inference: For each fine-tuned model, a reference model with the same base weights before fine-tuning was retained. Each member and non-member canary note was scored by both the target model and the reference model. The primary membership score was the loss gap:

$$\Delta\mathcal{L} = \mathcal{L}_{\text{reference}} - \mathcal{L}_{\text{target}}. \tag{2}$$

ROC-AUC, PR-AUC, accuracy, precision, recall, F1, specificity, and balanced accuracy were then evaluated.

H. Utility Metrics

To assess language-model quality independently of privacy, validation loss, test loss, and perplexity were reported:

$$\text{Perplexity} = \exp(\mathcal{L}), \tag{3}$$

where, $\mathcal{L}$ is the mean next-token cross-entropy loss on the corresponding split.

I. Statistical Uncertainty

The model-comparison notebook additionally computed bootstrap confidence intervals for leakage rate, exposure, and ROC-AUC, as well as paired Wilcoxon signed-rank tests on per-canary exposure values. The bootstrap intervals are used descriptively to summarize uncertainty around the reported metrics, whereas the Wilcoxon tests are used only for cautious pairwise interpretation of exposure differences (Table II).

IV. Results: Experiment I: Defense Comparison

A. Training Dynamics and Utility

The three training settings converged cleanly, but their utility profiles diverged markedly. As shown in Fig. 2, early stopping achieved the strongest held-out language modeling with validation loss 1.770 and validation perplexity 5.87, outperforming the baseline (validation perplexity 7.10). The combined regularized profile substantially weakened language-model fit, yielding validation perplexity 10.46 and test perplexity 10.36.

Fig. 2 shows that the training curves do not simply indicate optimization success; they also frame the privacy interpretation. The early-stopping run descends below the baseline on validation loss and preserves stable generalization behavior, confirming that it is the strongest utility configuration in Experiment I. The combined regularized profile, by contrast, converges more conservatively and remains consistently higher on validation loss, which is consistent with stronger capacity control.

The privacy summary in the companion panel is equally important. It shows that the configuration with the best language-model fit is not the one with the best privacy behavior. This visual contrast is central to the study's argument because it demonstrates that utility metrics alone would have selected a more privacy-sensitive model.

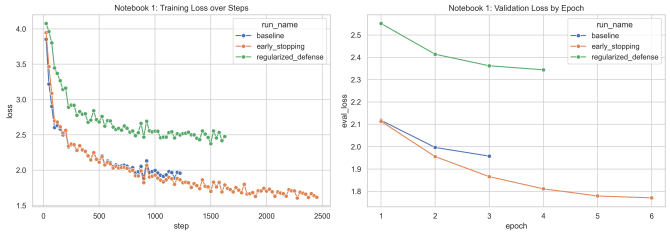
B. Prompt Leakage

The prompt-based results show that utility improvements did not translate into better privacy. Baseline and early stopping both leaked 7 canaries out of 40, corresponding to a leakage rate of 0.175. The combined regularized profile reduced the number of leaked canaries to zero. Exact-match recovery was also lowest for that profile (0.000), compared with 0.0075 for the baseline and 0.01375 for early stopping.

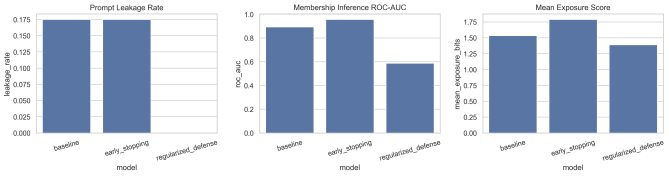
An important nuance is that the patient-identifier recovery rate saturated at 1.0 for all settings because the prompts explicitly included the patient identifier. Accordingly, the most meaningful prompt-level privacy indicators in this study are

TABLE II. ATTACK TAXONOMY AND EVALUATION METRICS USED THROUGHOUT THE STUDY

Attack family	Operational idea	Primary metrics
Prompt extraction	Generate continuations from clinically plausible prompts	Exact match, condition recovery, leakage rate
Ranking / exposure	Rank the true condition against 64 candidate codes	Mean rank, top-1 success, exposure bits
Membership inference	Distinguish seen from unseen canary notes using loss-gap scores	ROC-AUC, PR-AUC, F1, specificity, balanced accuracy



(a) Training loss and validation loss for Experiment I.



(b) Privacy summary for Experiment I.

Fig. 2. Experiment I results. Early stopping gave the best utility, whereas the combined regularized profile most strongly reduced the observed privacy risk metrics.

the condition recovery rate, exact-match recovery rate, and per-canary leakage rate. On those metrics, early stopping slightly worsened privacy relative to the baseline, while the combined regularized profile eliminated observed prompt leakage. Because the profile bundles lower learning rate, higher weight decay, and lower-layer freezing, these gains should be interpreted as the effect of the combined training configuration rather than as evidence for any single defense mechanism.

C. Ranking and Membership Inference

Ranking-based memorization and membership inference reinforce the same conclusion. Early stopping had the highest mean exposure score (1.79 bits) and the strongest membership-inference signal (ROC-AUC 0.956), whereas the combined regularized profile had the lowest mean exposure score (1.39 bits) and a much weaker membership signal (ROC-AUC 0.586). The baseline was intermediate, with ROC-AUC 0.893 and mean exposure 1.53 bits.

Experiment I therefore indicates that the strongest utility configuration was also the most privacy-sensitive, while the combined regularized profile reduced the observed leakage metrics at the cost of substantially worse clinical language modeling. In other words, in this setting, early stopping was not a privacy defense even though it remained a useful optimization strategy (Fig. 3).

The attack-type breakdown adds another layer of interpretation. Direct-recall prompts were the least effective across settings, whereas prefix completion and partial-note completion were more likely to reveal condition codes. This matters because it shows that privacy leakage is tied closely to continuation behavior in clinically shaped contexts. In practical terms, the attack surface is not limited to explicit question prompts; it extends to standard completion workflows that resemble documentation assistance or chart continuation.

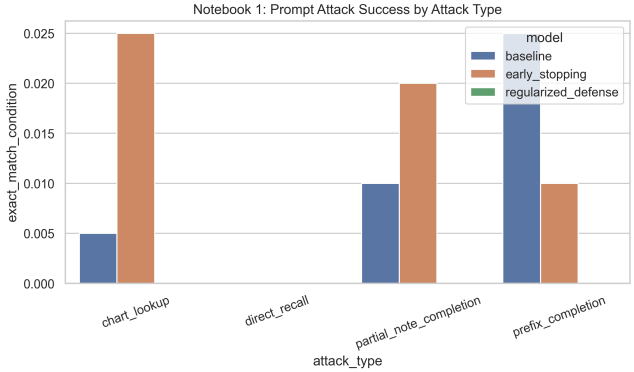


Fig. 3. Prompt attack success by attack type in Experiment I. Prefix-style and partial-note completions were more effective than direct recall prompts.

V. RESULTS: EXPERIMENT II: BACKBONE COMPARISON

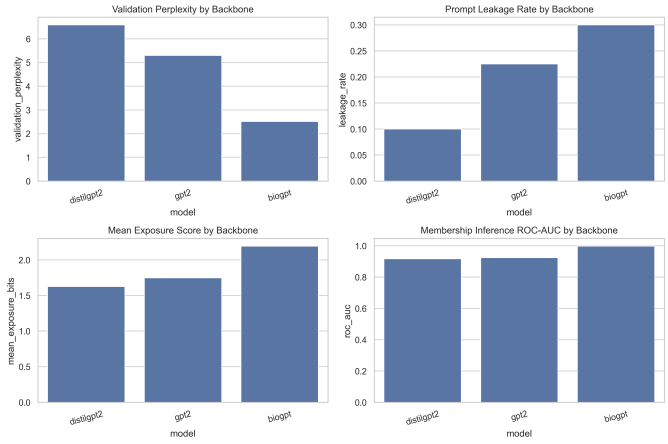
A. Utility Comparison Across Backbones

The model-comparison study yielded a clear performance ordering. DistilGPT2 achieved validation perplexity 6.59 and test perplexity 6.55. GPT-2 improved on this with validation perplexity 5.30 and test perplexity 5.33. BioGPT substantially outperformed both, reaching validation perplexity 2.52 and test perplexity 2.60. The training curves in Fig. 4 show faster and deeper optimization for BioGPT than for GPT-2 or DistilGPT2.

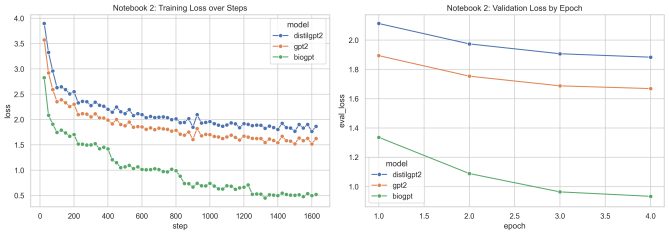
The two panels in Fig. 4 deserve joint interpretation. The utility panel shows a sharp improvement from DistilGPT2 to GPT-2 and then to BioGPT, especially in perplexity. The training-curve panel complements this by showing that BioGPT does not merely finish with a lower loss; it reaches a lower validation regime earlier and maintains that advantage across epochs. This pattern is consistent with a backbone that starts closer to the biomedical target distribution and adapts more efficiently to the corpus.

TABLE III. DEFENSE COMPARISON RESULTS (EXPERIMENT I)

Setting	Val. PPL	Test PPL	Leakage	Exact Match	Exposure	ROC-AUC	PR-AUC	F1
Baseline	7.10	7.05	0.175	0.0075	1.533	0.893	0.895	0.831
Early stopping	5.87	5.88	0.175	0.0138	1.788	0.956	0.953	0.907
Combined regularized profile	10.46	10.36	0.000	0.0000	1.388	0.586	0.537	0.692



(a) Utility and privacy summary across backbones.



(b) Training and validation curves across backbones.

Fig. 4. Experiment II results. BioGPT achieved the best utility, but also the highest leakage, exposure, and membership-inference risk.

B. Prompt Leakage and Memorization

The privacy ordering was the inverse of the "safest utility" interpretation. DistilGPT2 had the lowest leakage rate (0.10, i.e., 4 leaked canaries out of 40), GPT-2 leaked 9 canaries (0.225), and BioGPT leaked 12 canaries (0.30). Exact-match recovery rate rose from 0.00625 for DistilGPT2 to 0.02344 for GPT-2 and 0.03594 for BioGPT. Likewise, condition recovery rate increased monotonically across the same ordering.

Ranking-based exposure confirmed this pattern. Mean exposure rose from 1.63 bits for DistilGPT2 to 1.75 bits for GPT-2 and 2.19 bits for BioGPT. The average rank of the true secret improved (that is, privacy worsened) from 26.1 for DistilGPT2 to 25.1 for GPT-2 and 19.4 for BioGPT. The best-performing clinical backbone therefore also ranked the true secret highest among decoys.

C. Membership Inference

Membership inference produced the strongest privacy signal in Experiment II. DistilGPT2 and GPT-2 already had very high ROC-AUC values (0.918 and 0.925), but BioGPT approached perfect separation with ROC-AUC 0.998 and PR-AUC 0.998. The corresponding 95% bootstrap confidence interval for BioGPT was [0.991, 1.000], making it highly unlikely that the observed effect is a numerical artifact. This finding means that, under the evaluated loss-gap attack, the fine-tuned BioGPT almost perfectly distinguished seen canary notes from unseen controls.

D. Attack-Type Analysis

The backbone comparison also revealed prompt-family differences. Direct recall prompts were the weakest overall, whereas prefix completion, chart lookup, and partial-note completion were more successful. BioGPT dominated these attack families, suggesting that clinically plausible continuation contexts are especially effective at surfacing memorized strings from stronger biomedical backbones.

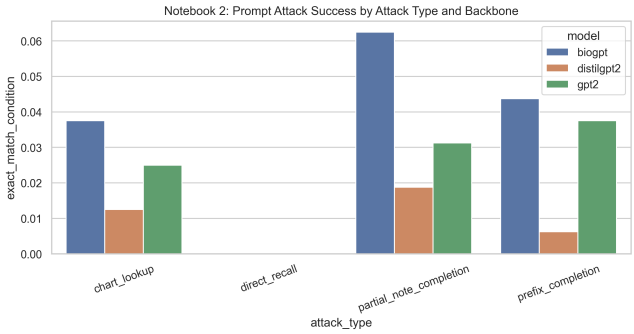


Fig. 5. Prompt attack success by attack type and backbone. Clinically natural completion prompts were more effective than direct recall prompts, especially against BioGPT.

This figure (Fig. 5) is useful not only because BioGPT is highest in aggregate, but also because the ordering is stable across attack families. That stability reduces the likelihood that the result is an artifact of one particular prompt wording. Instead, it suggests that the observed privacy ordering is not confined to a single prompt family.

E. Confidence Intervals and Paired Statistics

Bootstrap intervals preserved the same descriptive ordering. DistilGPT2 had a leakage-rate interval of [0.025, 0.200], GPT-2 had [0.075, 0.375], and BioGPT had [0.175, 0.450].

TABLE IV. BACKBONE COMPARISON RESULTS (EXPERIMENT II)

Backbone	Val. PPL	Test PPL	Leakage	Exact Match	Exposure	ROC-AUC	PR-AUC	F1	Train Time (s)	Params
DistilGPT2	6.59	6.55	0.100	0.0063	1.627	0.918	0.918	0.843	5226	81.9M
GPT-2	5.30	5.33	0.225	0.0234	1.748	0.925	0.922	0.822	7842	124.4M
BioGPT	2.52	2.60	0.300	0.0359	2.194	0.998	0.998	0.974	13588	346.8M

Membership ROC-AUC intervals also separated the models, with BioGPT concentrated near one. The paired Wilcoxon analysis on per-canary exposure values showed directionally higher exposure for BioGPT than for the other backbones, although at this sample size the pairwise p-values (0.086 to 0.100) remained above 0.05. DistilGPT2 and GPT-2 were not significantly different on exposure under the same test ($p = 0.615$).

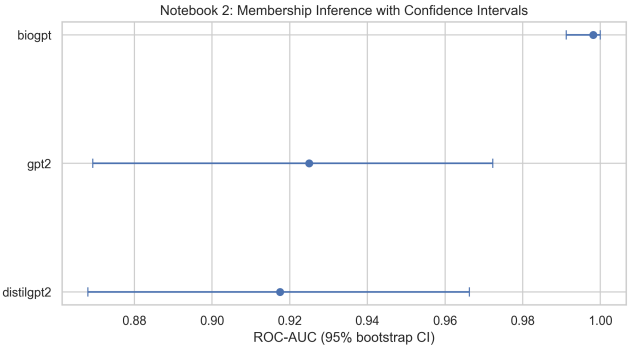


Fig. 6. Membership-inference ROC-AUC with 95% bootstrap confidence intervals in Experiment II. BioGPT exhibits near-perfect separability between member and non-member canary notes.

model_a	model_b	mean_exposure_a	mean_exposure_b	wilcoxon_statistic	p_value
biogpt	distilgpt2	2.1937	1.6275	287.5	0.0996
biogpt	gpt2	2.1937	1.7481	252.0	0.0857
distilgpt2	gpt2	1.6275	1.7481	354.0	0.6154

Fig. 7. Pairwise Wilcoxon exposure comparison among the three backbones. The direction of the effect consistently favors higher exposure for BioGPT, although the current sample size does not yield sub-0.05 p-values and the pairwise differences should therefore be interpreted cautiously.

The confidence-interval figure (Fig. 6) strengthens the interpretation by showing that the membership-inference advantage of BioGPT is not marginal. Its interval is compressed near one, while the smaller backbones remain clearly lower. The pairwise-exposure figure (Fig. 7) adds a useful nuance: although the direction of the effect consistently favors higher exposure for BioGPT, the present sample size limits formal pairwise significance. That nuance is important because it keeps the interpretation statistically disciplined while preserving the descriptive cross-metric ordering.

VI. DISCUSSION

The two experiments jointly support several main conclusions.

A. Better Utility Does Not Imply Better Privacy

Experiment I showed that early stopping improved utility substantially but did not reduce leakage. In fact, it produced the highest exposure and strongest membership-inference signal among the three training settings. This is an important practical lesson: interventions typically associated with better generalization in task performance should not automatically be described as privacy defenses. In a clinical LLM setting, the privacy and utility relationship is more nuanced.

At the same time, the combined regularized profile should be interpreted carefully. Because lower learning rate, higher weight decay, and lower-layer freezing were applied together, the present results support the privacy value of that conservative training profile as a whole, but they do not isolate the contribution of each component.

B. Backbone Choice Is a Privacy-Relevant Design Decision

Experiment II yielded a monotonic descriptive ordering across the evaluated backbones. DistilGPT2 had the weakest utility but the lowest privacy risk. GPT-2 improved utility and leaked more. BioGPT, the strongest and most domain-specialized model, achieved the best clinical language modeling but also the highest leakage, exposure, and membership-inference susceptibility. The exposure-based pairwise tests did not reach conventional significance thresholds at the current sample size, but the cross-metric pattern remained consistent across prompt leakage, exposure, and membership inference. Backbone selection in healthcare should therefore be treated not only as a performance decision but also as a privacy-governance decision.

C. Why Did BioGPT Leak More?

The most plausible explanation is not simply model size, but the combination of size and domain specialization. BioGPT began closer to the target distribution and therefore fit the clinical corpus more effectively. Better fit can improve generalization, but it can also increase the probability that the model assigns high confidence to rare, repeated strings such as synthetic canaries. The exposure and membership results support this interpretation. The model was not only better at generating clinically plausible text; it was also more sensitive to whether a canary note had been seen during training.

D. What the Prompt Results Mean

The prompt results suggest that privacy leakage is more easily elicited in clinically natural completion settings than in explicit question-answer forms. Direct recall prompts were generally weak, whereas partial-note completion and prefix-style prompts were more successful. This is important for real-world security because an attacker is not required to ask bluntly for the secret; instead, leakage may surface through

routine-looking note continuations, chart lookups, or audit-style prompts.

E. Interpreting the Saturated Identifier Metric

One metric deserves careful interpretation: patient-identifier recovery rate was 1.0 in several tables. This does not mean the models invented the identifier. The prompts themselves already contained the patient identifier. Accordingly, that metric is less informative than exact-match recovery, condition recovery, per-canary leakage rate, exposure, and membership ROC-AUC. These latter metrics should be emphasized in any formal presentation of the results.

F. Implications for Privacy-Aware Clinical NLP

Taken together, the findings imply that healthcare organizations adapting language models to internal notes should not rely on utility validation alone. A model with excellent perplexity may still leak or encode member-specific traces strongly enough to support membership inference. Conversely, stronger empirical mitigation profiles can reduce leakage, but may impose a non-trivial utility cost. This underscores the need for explicit privacy auditing, careful dataset governance, privacy-preserving representation design [40], and, where feasible, formal defenses such as differential privacy when the underlying data are truly sensitive [36], [37], [39]. The present study is empirical rather than formally private: no differentially private optimization arm was included, and the reported mitigations should not be interpreted as providing a formal privacy guarantee.

G. Practical Recommendations for Deployment

The results support several concrete recommendations. First, privacy auditing should be integrated into model-selection workflows before deployment approval rather than treated as a retrospective security test. Second, backbone choice should be documented in governance reviews as a privacy-sensitive design decision, especially when moving from a smaller general-purpose backbone to a larger or biomedical-specialized one. Third, when private optimization is not yet deployed, conservative regularization profiles and dataset deduplication should be treated as default baselines rather than optional extras. Fourth, prompt-space stress testing should emphasize clinically natural completions and multiple decoding modes, because these were more effective than direct recall prompts in both notebooks. Fifth, deployment review should weigh computational cost together with utility and privacy: in Experiment II, the highest-risk backbone was also the most computationally expensive, so more computation did not buy greater privacy. Finally, institutions fine-tuning on sensitive records should view the present protocol as a minimum empirical audit layer that complements, but does not replace, stronger controls such as secure infrastructure, access governance, red-team testing, data-minimization practice, documented approval workflows, and broader privacy-preserving machine-learning governance frameworks [42].

VII. Scope and Generalization Considerations

The present study is intentionally focused and controlled, which supports clear interpretation of the observed effects. Several points nevertheless help frame the scope of the findings.

1) Corpus scope: The experiments target a multi-specialty clinical transcription corpus with long-form narrative structure. This is a strong setting for comparative privacy analysis, although MTSamples is not a modern production EHR repository and does not fully reflect current documentation workflows, institution-specific templates, longitudinal note chains, or heterogeneous authoring practices. Exact leakage magnitudes may therefore vary for other note distributions, longer contexts, or operational EHR settings.

2) Configuration sensitivity: The reported rankings were obtained under one controlled configuration per experiment, including fixed sequence length, batch size, warmup, optimizer settings, and training duration. The controlled setup supports cleaner comparison, but it does not establish invariance of the results under all plausible hyperparameter changes.

3) Run-to-run uncertainty: The principal comparisons were executed with fixed seeds to preserve strict comparability, and bootstrap confidence intervals were reported for key privacy metrics. Additional repeated-seed experiments would refine uncertainty estimates and runtime-variance analysis without changing the central comparative design.

4) Threat-model coverage: Canary insertion, exposure scoring, and membership inference provide complementary views of memorization. They do not exhaust all possible privacy risks, such as attribute inference, model inversion, gradient leakage, embedding reconstruction, adaptive adversarial prompting, or linkage-based attacks, but they capture three established and measurable leakage pathways in generative-model auditing.

5) Defense scope: The study emphasizes controlled comparison among practical fine-tuning profiles and backbone choices. Formal privacy mechanisms, including end-to-end differentially private optimization, remain an important extension for settings in which auditable privacy guarantees are required.

VIII. Conclusion

A controlled study of privacy leakage and memorization in fine-tuned clinical language models was presented using clinical narrative transcriptions, train-only synthetic canaries, and three complementary attack families. Two experiments were performed. The first compared training settings and showed that early stopping improved utility but did not mitigate privacy risk, whereas a combined regularized training profile eliminated observed prompt leakage and sharply reduced membership-inference susceptibility at the cost of degraded utility. The second compared backbones and showed a clear privacy and utility trade-off: DistilGPT2 was the safest but weakest model, GPT-2 was intermediate, and BioGPT achieved the best utility but also the highest prompt leakage, exposure, and membership-inference risk under the evaluated attacks.

The results support a practical but important conclusion for healthcare AI: *better clinical language modeling can coincide with worse privacy behavior, and backbone choice itself is a privacy-relevant design decision*. For deployable clinical LLMs, privacy auditing should therefore be treated as a first-class evaluation axis alongside utility, latency, and resource

cost. The exposure-based pairwise statistics in Experiment II remained underpowered for conventional significance claims, but the overall descriptive pattern across multiple privacy metrics remained consistent.

Future work should extend the same protocol to credentialed EHR corpora, repeated seeds, formally private optimization, and more adaptive extraction attacks. Nonetheless, the present study already offers a reproducible, defensible, and methodologically strong baseline for auditing privacy leakage in clinical language-model fine-tuning.

ACKNOWLEDGMENT

This research received no external funding. Appreciation is extended to Mohammed First University for providing the computational environment used to conduct the experiments.

REFERENCES

[1] J. Wiens, S. Saria, M. Sendak, M. Ghassemi, V. X. Liu, F. Doshi-Velez, K. Jung, K. Heller, D. Kale, R. Saeed, P. N. Ossorio, S. Thadaney-Israni, and A. Goldenberg, "Do no harm: a roadmap for responsible machine learning for health care," *Nature Medicine*, vol. 25, no. 9, pp. 1337–1340, 2019, doi: 10.1038/s41591-019-0548-6.

[2] J. He, S. L. Baxter, J. Xu, J. Xu, X. Zhou, and K. Zhang, "The practical implementation of artificial intelligence technologies in medicine," *Nature Medicine*, vol. 25, no. 1, pp. 30–36, 2019, doi: 10.1038/s41591-018-0307-0.

[3] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019, doi: 10.1038/s41591-018-0300-7.

[4] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature Medicine*, vol. 29, no. 8, pp. 1930–1940, 2023, doi: 10.1038/s41591-023-02448-8.

[5] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. T. Tanwani, H. Cole-Lewis, S. Pfohl *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, pp. 172–180, 2023, doi: 10.1038/s41586-023-06291-2.

[6] L. Y. Jiang, X. C. Liu, N. Pour Nejatian, M. Nasir-Moin, D. Wang, A. Abidin, M. A. Eaton, H. W. Riina, M. J. Laufer, P. S. Kim *et al.*, "Health system-scale language models are all-purpose prediction engines," *Nature*, vol. 619, pp. 357–362, 2023, doi: 10.1038/s41586-023-06160-y.

[7] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Machine Intelligence*, vol. 2, no. 6, pp. 305–311, 2020, doi: 10.1038/s42256-020-0186-1.

[8] M. Adnan, S. Kalra, J. C. Cresswell, G. W. Taylor, and H. R. Tizhoosh, "Federated learning and differential privacy for medical image analysis," *Scientific Reports*, vol. 12, art. no. 3034, 2022, doi: 10.1038/s41598-022-05539-7.

[9] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, "The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks," in *28th USENIX Security Symposium*, 2019, pp. 267–284. [Online]. Available: https://www.usenix.org/conference/usenixsecurity19/presentation/carlini

[10] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, Ú. Erlingsson *et al.*, "Extracting Training Data from Large Language Models," in *30th USENIX Security Symposium*, 2021, pp. 2633–2650. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting

[11] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," in *2017 IEEE Symposium on Security and Privacy*, 2017, pp. 3–18, doi: 10.1109/SP.2017.41.

[12] J. Mattern, F. Mireshghallah, Z. Jin, B. Schölkopf, M. Sachan, and T. Berg-Kirkpatrick, "Membership Inference Attacks against Language Models via Neighbourhood Comparison," in *Findings of ACL 2023*, pp. 11330–11343, 2023, doi: 10.18653/v1/2023.findings-acl.719.

[13] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramèr, "Membership Inference Attacks From First Principles," in *2022 IEEE Symposium on Security and Privacy*, 2022, pp. 1897–1914, doi: 10.1109/SP46214.2022.9833649.

[14] S. Ishihara, "Training Data Extraction From Pre-trained Language Models: A Survey," in *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing*, 2023, pp. 171–183, doi: 10.18653/v1/2023.trustnlp-1.23.

[15] K. Lee, D. Ippolito, A. Nystrom, C. Zhang, D. Eck, C. Callison-Burch, and N. Carlini, "Deduplicating Training Data Makes Language Models Better," in *Proceedings of ACL 2022*, pp. 8424–8445, 2022, doi: 10.18653/v1/2022.acl-long.577.

[16] F. Mireshghallah, A. Uniyal, T. Wang, D. Evans, and T. Berg-Kirkpatrick, "An Empirical Analysis of Memorization in Fine-tuned Autoregressive Language Models," in *Proceedings of EMNLP 2022*, pp. 1816–1831, 2022, doi: 10.18653/v1/2022.emnlp-main.119.

[17] Q. Anthony, S. Biderman, U. Prashanth, S. Purohit, E. Raff, H. Schoelkopf, and V. Veliche, "Emergent and Predictable Memorization in Large Language Models," in *Advances in Neural Information Processing Systems 36*, 2023, doi: 10.52202/075280-1219.

[18] A. Aghajanyan, A. Markosyan, K. Tirumala, and L. Zettlemoyer, "Memorization Without Overfitting: Analyzing the Training Dynamics of Large Language Models," in *Advances in Neural Information Processing Systems 35*, 2022, doi: 10.52202/068431-2773.

[19] S. Ishihara and H. Takahashi, "Quantifying Memorization and Detecting Training Data of Pre-trained Language Models using Japanese Newspaper," in *Proceedings of the 17th International Natural Language Generation Conference*, 2024, pp. 177–188, doi: 10.18653/v1/2024.inlg-main.14.

[20] S. Zhang, H. Li, and R. Ji, "Code Membership Inference for Detecting Unauthorized Data Use in Code Pre-trained Language Models," in *Findings of EMNLP 2024*, pp. 10629–10642, 2024, doi: 10.18653/v1/2024.findings-emnlp.621.

[21] H. Chang, A. S. Shamsabadi, K. Katevas, H. Haddadi, and R. Shokri, "Context-Aware Membership Inference Attacks against Pre-trained Large Language Models," in *Proceedings of EMNLP 2025*, 2025, doi: 10.18653/v1/2025.emnlp-main.370.

[22] D. Parii, T. van Osch, and C. Sun, "Machine Unlearning of Personally Identifiable Information in Large Language Models," in *Proceedings of the Natural Legal Language Processing Workshop 2025*, 2025, doi: 10.18653/v1/2025.nllp-1.6.

[23] K. Edemacu and X. Wu, "Privacy Preserving Prompt Engineering: A Survey," *ACM Computing Surveys*, 2025, doi: 10.1145/3729219.

[24] C. Peris, C. Dupuy, J. Majmudar, R. Parikh, S. Smaili, R. Zemel, R. Gupta, and A. Kumar, "Privacy in the Time of Language Models," in *Proceedings of WSDM 2023*, pp. 1291–1299, 2023, doi: 10.1145/3539597.3575792.

[25] K. Chen, X. Zhou, Y. Lin, S. Feng, L. Shen, P. Wu, and Z. Liu, "A survey on privacy risks and protection in large language models," *Journal of King Saud University Computer and Information Sciences*, 2025, doi: 10.1007/s44443-025-00177-1.

[26] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020, doi: 10.1093/bioinformatics/btz682.

[27] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, art. 2, 2022, doi: 10.1145/3458754.

[28] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, "BioGPT: generative pre-trained transformer for biomedical text generation and mining," *Briefings in Bioinformatics*, vol. 23, no. 6, 2022, doi: 10.1093/bib/bbac409.

[29] X. Yang, N. PourNejatian, H. C. Shin, K. E. Smith, C. Parisien, C. Compas, C. Martin, M. B. Flores, Y. Zhang, T. Magoc *et al.*, "GatorTron: A Large Language Model for Clinical Natural Language Processing," *medRxiv*, 2022, doi: 10.1101/2022.02.27.22271257.

[30] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," *npj Digital Medicine*, vol. 4, art. 86, 2021, doi: 10.1038/s41746-021-00455-y.

[31] Y. Li, S. Rao, J. R. Ayala Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, A. Rahimi, K. Salimi-Khorshidi, G. Zottoli *et al.*, "BEHRT: Transformer for Electronic Health Records," *Scientific Reports*, vol. 10, art. 7155, 2020, doi: 10.1038/s41598-020-62922-y.

[32] G. Kaissis, A. Ziller, J. Passerat-Palmbach, T. Ryffel, D. Usynin, A. Trask, I. Lima, J. Mancuso, F. Jungmann, M. Steinborn *et al.*, "End-to-end privacy preserving deep learning on multi-institutional medical imaging," *Nature Machine Intelligence*, vol. 3, no. 6, pp. 473–484, 2021, doi: 10.1038/s42256-021-00337-8.

[33] N. Rieke, J. Hancox, W. Li, F. Milletarì, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. Landman, K. Maier-Hein *et al.*, "The future of digital health with federated learning," *npj Digital Medicine*, vol. 3, art. 119, 2020, doi: 10.1038/s41746-020-00323-1.

[34] X. Li, Y. Gu, N. Dvornek, L. H. Staib, P. Ventola, and J. S. Duncan, "Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results," *Medical Image Analysis*, vol. 65, art. 101765, 2020, doi: 10.1016/j.media.2020.101765.

[35] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, "Federated Learning for Healthcare Informatics," *Journal of Healthcare Informatics Research*, vol. 5, pp. 1–19, 2021, doi: 10.1007/s41666-020-00082-4.

[36] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep Learning with Differential Privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 308–318, doi: 10.1145/2976749.2978318.

[37] D. Yu, S. Naik, A. Backurs, S. Gopi, H. A. Inan, G. Kamath, J. Kulkarni, Y. T. Lee, and L. Wang, "Differentially Private Fine-tuning of Language Models," *Journal of Privacy and Confidentiality*, vol. 14, no. 1, 2024, doi: 10.29012/jpc.880.

[38] A. El Ouadrhiri and A. Abdelhadi, "Differential Privacy for Deep and Federated Learning: A Survey," *IEEE Access*, vol. 10, pp. 22359–22380, 2022, doi: 10.1109/ACCESS.2022.3151670.

[39] K. Pan, Y.-S. Ong, M. Gong, H. Li, A. K. Qin, and Y. Gao, "Differential privacy in deep learning: A literature survey," *Neurocomputing*, vol. 604, art. 127663, 2024, doi: 10.1016/j.neucom.2024.127663.

[40] Y. Li, T. Baldwin, and T. Cohn, "Towards Robust and Privacy-preserving Text Representations," in *Proceedings of ACL 2018, Short Papers*, 2018, pp. 25–30, doi: 10.18653/v1/P18-2005.

[41] Y. Zhang, H. Pei, S. Zhen, Q. Li, and F. Liang, "Chat Generative Pre-Trained Transformer (ChatGPT) usage in healthcare," *Gastroenterology & Endoscopy*, vol. 2, no. 3, pp. 337–343, 2023, doi: 10.1016/j.gande.2023.07.002.

[42] L. Ramachandrappa, K. H. Krishnappa, A. Salim, R. Warren, R. Madhura, and R. Manasa, "Privacy-Preserving Machine Learning in Healthcare," in *BioMed Bots Assurance*, Boca Raton, FL, USA: CRC Press, 2025, ch. 15, doi: 10.1201/9781003561309-15.

AUTHORS' PROFILE

Yassine Chahid received the M.Sc. degree in science and technology from the Faculty of Science and Technology, Settat, Morocco, in 2014, and the Ph.D. degree in mathematics and computer science from the Faculty of Sciences, Mohammed First University, Oujda, Morocco, in 2022. Since 2014, he has worked with several private-sector companies, where he has held technical and research-oriented positions. He is currently a Technical Lead specializing in artificial intelligence and cybersecurity solutions. His research interests include federated learning, information security, cryptography, and distributed systems.

Anas Chahid received the Engineering degree in computer science from the National School of Applied Sciences (ENSA), Oujda, Morocco. He is currently a Software Engineer and a Ph.D. student in medical artificial intelligence. His work focuses on the design of intelligent and customized software systems. His professional interests include artificial intelligence, data analytics, software engineering, and healthcare applications.

Ismail Chahid received the DUT degree in information technology from the École Supérieure de Technologie, Oujda, Morocco, in 2008, the B.Sc. degree in IT management from the Faculty of Polydisciplinary Studies, Tétouan, Morocco, in 2009, and the M.Sc. degree in business intelligence from the Faculty of Science and Technology, Béni Mellal, Morocco, in 2012. He is currently the Head of the IT Department at the Faculty of Medicine and Pharmacy of Oujda, where he has been working since 2019. His professional interests include business intelligence, data warehousing, information systems management, and software engineering.

Aissa Kerkour El Miad received the M.Sc. degree in operational research and informatics and the Ph.D. degree in computer science from Mohammed First University, Oujda, Morocco. He is currently a Professor with the Department of Computer Science, Faculty of Sciences, Mohammed First University, Oujda. His research interests include image processing, artificial intelligence, high-performance computing, and data mining.