How to Train Private Clinical Language Models: A Comparative Study of Privacy-Preserving Pipelines for ICD-9 Coding

Mathieu Dufour
Department of Mathematics
Imperial College London
mathieu.dufour23@imperial.ac.uk

Andrew Duncan
Department of Mathematics
Imperial College London
a.duncan@imperial.ac.uk

Abstract

Large language models trained on clinical text risk exposing sensitive patient information, yet differential privacy (DP) methods often severely degrade the diagnostic accuracy needed for deployment. Despite rapid progress in DP optimisation and text generation, it remains unclear which privacy-preserving strategy actually works best for clinical language tasks. We present the first systematic head-to-head comparison of four training pipelines for automated diagnostic coding from hospital discharge summaries. All pipelines use identical 1B-parameter models and matched privacy budgets to predict ICD-9 codes. At moderate and relaxed privacy budgets ($\varepsilon \in \{4, 6\}$), knowledge distillation from DP-trained teachers outperforms both direct DP-SGD and DP-synthetic data training, recovering up to 63% of the non-private performance whilst maintaining strong empirical privacy (membership-inference AUC ≈ 0.5). These findings expose large differences in the privacy-utility trade-off across architectures and identify knowledge distillation as the most practical route to privacy-preserving clinical NLP.

1 Introduction

Large language models trained on clinical text can inadvertently memorise and reveal sensitive patient details [Carlini et al., 2021b], raising serious concerns about their deployment in healthcare. Differential privacy (DP) offers formal guarantees that individual patient records cannot be reliably inferred from model outputs [Dwork and Roth, 2014], but applying DP training to large models often comes at a steep price: accuracy drops of 40% or more have been reported for clinical NLP tasks [Dadsetan et al., 2024]. This tension between diagnostic utility and patient confidentiality has become a central challenge for trustworthy clinical AI.

A variety of strategies have been proposed to mitigate these trade-offs. Parameter-efficient tuning methods such as Low-Rank Adaptation (LoRA) restrict memorisation through low-rank updates and may offer partial implicit privacy [Malekmohammadi and Farnadi, 2025]. Generative approaches train large DP-protected models to produce synthetic clinical notes that can be used freely for downstream tasks [Yue et al., 2023]. More recently, differentially private knowledge distillation allows a private "teacher" to transfer its knowledge to a smaller "student" via synthetic data and soft labels, preserving the teacher's privacy guarantee through post-processing [Flemings and Annavaram, 2024].

Despite this progress, no prior work has compared these methods under identical conditions. Existing studies vary in model size, dataset, or privacy accounting, making it unclear which approach provides the best balance of accuracy, efficiency, and protection. For practitioners deciding how to deploy language models in hospitals, this lack of systematic evidence leaves a crucial gap.

This study asks a single question: "Given a fixed model capacity, which privacy mechanism best balances diagnostic accuracy and formal privacy guarantees in clinical NLP?"

Preprint.

We answer this through a controlled, head-to-head comparison of four privacy-preserving training pipelines for automated ICD-9 diagnostic coding, i.e. the task of mapping hospital discharge summaries to standardised disease codes, using the publicly available MIMIC-III dataset [Johnson et al., 2016]. All final classifiers share the same architecture and capacity (1B parameters) to isolate the impact of the privacy mechanism itself. In pipelines involving teacher-student architectures (DP-Synthetic, DP-Distil), we train larger 3B "teacher" models under DP-SGD, then use their synthetic outputs and/or soft labels to train the final 1B "student" classifier. Because the student only accesses post-processed outputs of the DP teachers, it inherits the same formal privacy guarantees without direct exposure to real patient records.

The evaluated pipelines are: **(1) DP-Small**—direct DP-SGD on the 1B model; **(2) DP-Synthetic**—training on synthetic notes from a 3B DP-protected generator; **(3) DP-Distil**—knowledge distillation from DP-trained 3B teachers providing synthetic data and soft labels; **(4) LoRA-No-DP**—non-private LoRA fine-tuning as a utility upper bound.

We assess each approach across privacy budgets $\varepsilon \in \{2, 4, 6\}$ with fixed $\delta = 10^{-5}$, measuring utility (micro/macro-F_1, AUPRC) and empirical privacy via membership-inference attacks [Carlini et al., 2021a]. By holding model capacity constant, our study isolates how different training pipelines reshape the privacy-utility frontier in clinical NLP. The results reveal that architectural choices, not merely the privacy budget, substantially shape this trade-off, highlighting knowledge distillation as an effective pathway for developing differentially private language models that preserve clinical usefulness.

2 Methods

2.1 Dataset

We evaluate all methods on the MIMIC-III v1.4 critical-care database [Johnson et al., 2016], a collection of de-identified electronic health records from intensive care unit admissions. We use discharge summaries—clinical narratives summarising each hospital stay—associated with the 50 most frequent ICD-9 diagnostic codes. ICD-9 (International Classification of Diseases, 9th Revision) is a standardised medical coding system used for billing and epidemiology. The final dataset contains 44,778 training notes, 5,624 validation notes, and 5,586 test notes, with splits defined at the admission level to prevent patient overlap. Notes are tokenised using the LLaMA-3.2 tokeniser and truncated to 512 tokens. The prediction task is multi-label ICD-9 classification.

2.2 Model Architecture

All final classifiers share an identical 1B-parameter LLaMA-3.2 backbone [Touvron et al., 2023, Grattafiori et al., 2024], fine-tuned via LoRA [Hu et al., 2021] on the query (Q) and value (V) projections (rank $r = 4$, $\alpha = 16$). The DP "teacher" and "generator" models use the 3B-parameter variant with LoRA rank $r = 8$ and $\alpha = 32$. These settings yield approximately 0.53M trainable parameters for 1B models and 2.3–2.4M for 3B models, keeping adapter capacity roughly proportional to backbone size and ensuring a controlled comparison across training pipelines. This design follows empirical observations that larger models tend to retain higher utility under DP-SGD at fixed privacy budgets [Yu et al., 2021].

We standardise on 1B students rather than deploying 3B teachers directly for two reasons: (1) to ensure fair comparison across pipelines at identical capacity, and (2) to enable resource-efficient deployment. The 1B models offer 2.7× faster inference at 13ms versus 36ms per sample and require only 4.4GB VRAM compared to 8.5GB for the teachers (Table 5), making them practical for clinical settings with hardware constraints.

2.3 Differential privacy mechanisms

Differentially private training follows DP-SGD [Abadi et al., 2016] with per-example gradient clipping and Gaussian noise calibrated via Rényi DP accounting in Opacus [Yousefpour et al., 2021]. Unless stated otherwise, $\delta = 10^{-5}$ and $\varepsilon \in \{2, 4, 6\}$, corresponding to strong, moderate, and relaxed privacy regimes commonly reported in medical NLP [Dadsetan et al., 2024]. Gradient clipping thresholds are $C = 1.0$ for 1B models and $C = 0.7$ for 3B models. In DP-Distil, each teacher model consumes $\varepsilon/2$ to preserve overall (ε, δ) under sequential composition [Dwork and Roth, 2014].

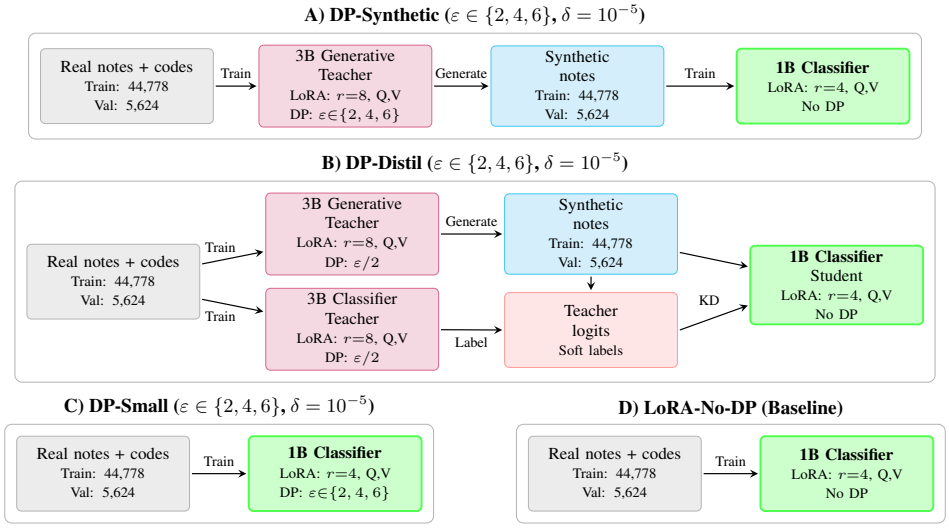

Figure 1: Four training pipelines producing identical 1B classifiers (green). **(A)** DP-trained 3B generator creates synthetic data. **(B)** DP-trained 3B teachers (each $\varepsilon/2$) provide synthetic data and soft labels for knowledge distillation. **(C)** Direct DP-SGD on 1B model. **(D)** Non-private LoRA baseline. All DP pipelines use $\varepsilon \in \{2, 4, 6\}$ with $\delta = 10^{-5}$.

2.4 Privacy-preserving training pipelines

Figure 1 illustrates the four pipelines, all producing identical 1B classifiers for fair comparison.

DP-Synthetic: A 3B generative model is fine-tuned under DP-SGD for conditional generation with control codes [Yue et al., 2023], producing 44,778 synthetic notes (nucleus sampling: $p = 0.9$, $T = 0.8$) to train a 1B classifier inheriting the generator's (ε, δ) via post-processing.

DP-Distil: Two 3B teachers trained under DP-SGD (each $\varepsilon/2, \delta/2$)—generative and classification—provide synthetic data and soft labels. A 1B student trains via MSE on raw logits ($\alpha = 0$), inheriting (ε, δ) through sequential composition [Flemings and Annavaram, 2024].

DP-Small: Direct DP-SGD training of the 1B model with binary cross-entropy loss and frequency-weighted positive class weights, evaluating direct private training without architectural modifications.

LoRA-No-DP: LoRA fine-tuning without gradient clipping or noise, serving as an upper bound on utility and testing LoRA's implicit privacy properties [Malekmohammadi and Farnadi, 2025].

2.5 Evaluation

Utility. We report Micro- and Macro-F_1, Micro-AUPRC, and Hamming loss, with thresholds optimised on the validation set. Micro-F_1 aggregates predictions across all labels, whilst Macro-F_1 computes per-label F_1 scores and averages them, giving equal weight to rare codes.

Privacy. Following Carlini et al. [2021a] and Shokri et al. [2017], we train logistic-regression membership-inference attacks on five features derived from model logits: cross-entropy loss, maximum confidence, entropy, confidence margin, and L_2 norm. Attacks are trained on a balanced dataset of 5,586 training members versus 5,586 test non-members. AUC = 0.5 indicates random guessing.

3 Results

3.1 Overall performance

Table 1 summarises classification accuracy across all privacy budgets $\varepsilon \in \{2, 4, 6\}$ ($\delta = 10^{-5}$). At the strictest budget ($\varepsilon = 2$), DP-Small performs best among DP methods (Micro-$F_1 \approx 0.26$), reflecting the advantage of allocating the full privacy budget to a single model. As ε increases, DP-Distil overtakes it, reaching 0.33 Micro-F_1 at $\varepsilon = 6$, an 8.7% gain over DP-Small and a 48% improvement over DP-Synthetic. The non-private LoRA-No-DP baseline achieves 0.52, defining the practical upper bound on utility.

Table 1: Classification performance across pipelines and privacy budgets ($\delta = 10^{-5}$ for all DP methods). Arrows indicate direction of improvement.

Pipeline	ε	Micro-F_1 ↑	Macro-F_1 ↑	Micro-AUPRC ↑	Ham ↓
DP-Distil	2	0.217	0.232	0.260	0.533
DP-Small	2	**0.258**	**0.257**	**0.285**	**0.343**
DP-Synthetic	2	0.220	0.207	0.198	0.402
DP-Distil	4	**0.289**	**0.280**	**0.321**	**0.273**
DP-Small	4	0.279	0.275	0.305	0.286
DP-Synthetic	4	0.228	0.216	0.212	0.442
DP-Distil	6	**0.330**	**0.303**	**0.347**	**0.216**
DP-Small	6	0.304	0.298	0.334	0.247
DP-Synthetic	6	0.222	0.213	0.225	0.476
LoRA-No-DP	∞	**0.521**	**0.499**	**0.565**	**0.099**

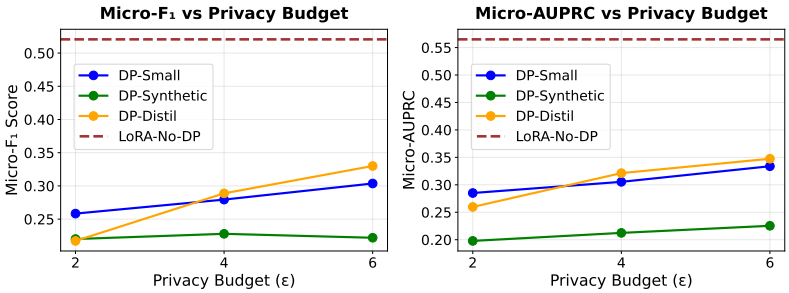

Figure 2: Utility-privacy trade-off for Micro-F_1 (left) and Micro-AUPRC (right). DP-Distil and DP-Small improve monotonically with privacy budget, whilst DP-Synthetic plateaus regardless of ε. DP-Distil dominates at $\varepsilon \geq 4$; DP-Synthetic fails to scale.

Performance scales monotonically with privacy budget for DP-Small and DP-Distil but remains flat for DP-Synthetic (Figure 2, ≈ 0.22 F_1 across ε). At $\varepsilon = 6$, DP-Distil recovers 63% of the non-private performance whilst maintaining formal DP guarantees, demonstrating that architectural design can offset much of the utility loss induced by DP noise.

3.2 Privacy evaluation

Table 2 reports membership-inference attack (MIA) results. All DP methods yield AUCs between 0.49 and 0.51, effectively random guessing, confirming strong empirical privacy. The LoRA-No-DP baseline, whilst partially protective (AUC ≈ 0.57), remains measurably vulnerable, notably to loss-based attacks (AUC ≈ 0.54).

Figure 3 shows ensemble MIA performance across configurations (left) and individual feature decomposition at $\varepsilon = 4$ (right). For DP models, all features—loss, confidence, entropy, margin, and L_2 norm—cluster around 0.5 AUC, indicating that differential privacy neutralises multiple leakage channels simultaneously. Notably, empirical AUC shows no systematic correlation with ε, suggesting that, within this range, the presence of DP noise itself matters more than the precise choice of ε.

3.3 Training efficiency and deployment considerations

While DP-Distil achieves the best utility-privacy trade-off, this comes at a computational cost. Training times vary substantially across pipelines: DP-Distil requires 38–49 hours due to training two 3B teachers plus the 1B student, compared to 14–15 hours for DP-Synthetic, 5–10 hours for DP-Small, and just 3 hours for LoRA-No-DP (all on RTX 6000 Ada GPUs, see Table 4 for details).

However, this training overhead is a one-time cost. All pipelines produce architecturally identical 1B classifiers that require only 13ms per sample and 4.4GB VRAM at inference (Table 5). This uniformity means that choosing DP-Distil does not impose any deployment penalties—the same efficient 1B model runs in production regardless of how it was trained. For healthcare institutions

Table 2: Privacy evaluation via membership inference attacks across all pipelines and privacy budgets, showing results for ensemble attacks and individual attack features (loss, confidence, entropy, margin). Values closer to 0.5 indicate better privacy (random guessing). All DP methods achieve AUC ≈ 0.5; LoRA-No-DP shows vulnerability with ensemble AUC of 0.565.

Pipeline	ε	Ensemble AUC	Loss AUC	Conf. AUC	Entropy AUC	Margin AUC
DP-Distil	2	0.487	0.499	0.500	0.500	0.500
DP-Distil	4	0.485	0.504	0.497	0.502	0.501
DP-Distil	6	0.498	0.506	0.492	0.504	0.499
DP-Small	2	0.512	0.505	0.492	0.509	0.498
DP-Small	4	0.497	0.507	0.493	0.507	0.503
DP-Small	6	0.499	0.505	0.494	0.508	0.493
DP-Synthetic	2	0.506	0.500	0.498	0.503	0.503
DP-Synthetic	4	0.496	0.503	0.495	0.505	0.498
DP-Synthetic	6	0.504	0.501	0.502	0.503	0.507
LoRA-No-DP	∞	0.565	0.536	0.493	0.507	0.510

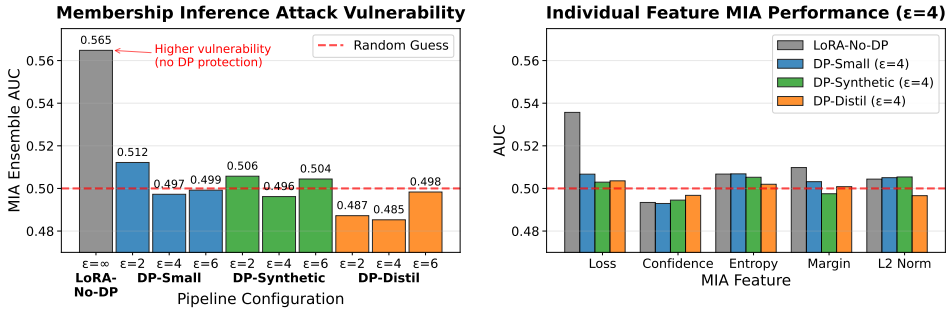

Figure 3: MIA vulnerability analysis. Left: Ensemble AUC showing LoRA-No-DP's higher vulnerability versus near-random performance for DP methods. Right: Individual feature contributions at $\varepsilon = 4$, demonstrating consistent protection across attack vectors for DP methods. All DP methods achieve AUC ≈ 0.5; LoRA-No-DP remains vulnerable.

processing thousands of daily predictions, the initial training investment in DP-Distil may be justified by its sustained 8.7% accuracy advantage over DP-Small at $\varepsilon = 6$.

3.4 Teacher-student distillation performance

Table 3 shows that 1B-parameter students consistently match or slightly outperform their 3B-parameter teachers in DP-Distil. At $\varepsilon = 6$, the student achieves Micro-F_1 of 0.3300 versus the teacher's 0.3212, with similar patterns for Macro-F_1. This result is consistent with previous distillation literature showing students can surpass teachers through regularisation effects [Furlanello et al., 2018, Nagarajan et al., 2024]. The student advantage is most pronounced at higher privacy budgets where teachers have more utility to transfer, while the gap narrows at stricter settings ($\varepsilon = 2$: +0.0004 for Micro-F_1), as visualised in Figure 4.

Table 3: Performance comparison between DP-Distil teachers (3B parameters) and their distilled students (1B parameters). Positive gaps indicate teacher superiority.

ε	Teacher Micro-F_1	Student Micro-F_1	Gap	Teacher Macro-F_1	Student Macro-F_1	Gap
2	0.2176	0.2172	0.0004	0.2303	0.2318	-0.0015
4	0.2824	0.2888	-0.0064	0.2765	0.2802	-0.0037
6	0.3212	0.3300	-0.0088	0.2988	0.3034	-0.0047

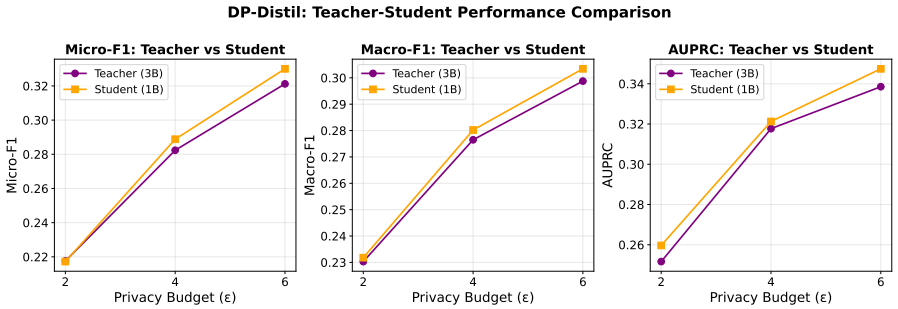

Figure 4: Performance comparison between 3B-parameter DP-Distil teachers and their 1B-parameter distilled students across privacy budgets ($\varepsilon \in \{2, 4, 6\}$).

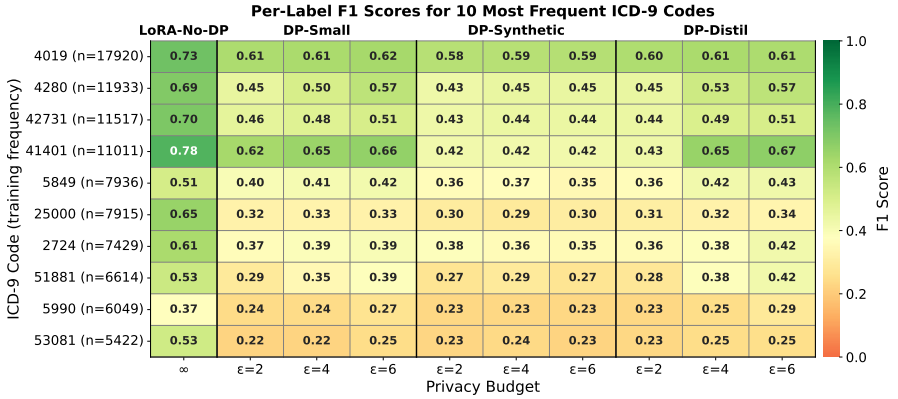

Figure 5: Per-label F_1 scores for the ten most frequent ICD-9 codes, sorted by decreasing training frequency (n in parentheses). DP-Synthetic consistently underperforms DP-Small and DP-Distil across nearly all codes, while LoRA-No-DP shows irregular memorisation patterns.

3.5 Per-label performance

Figure 5 displays F_1 scores for the ten most frequent ICD-9 codes, revealing how privacy mechanisms affect individual diagnoses. All three DP methods show general correlation with code frequency, with performance declining for rarer codes. DP-Small and DP-Distil exhibit similar patterns, with a notable peak at code 41401 (coronary atherosclerosis, fourth most frequent, n=11,011), achieving F_1 scores of 0.66 and 0.67 respectively at $\varepsilon = 6$.

DP-Synthetic shows consistent underperformance across nearly all codes compared to other DP methods. While it exhibits smooth frequency-dependent degradation—declining from 0.59 (code 401.9) to 0.23 (code 53081) at $\varepsilon = 6$—its per-code F_1 scores remain systematically lower than DP-Small and DP-Distil regardless of code frequency or privacy budget.

LoRA-No-DP shows the most irregular pattern, with substantial variation unexplained by frequency alone. It achieves 0.78 for code 41401 but only 0.37 for code 5990 (n=6,049).

3.6 Key observations

1. **Knowledge distillation outperforms direct DP-SGD ($\varepsilon \geq 4$).** High-capacity teachers retain higher utility under DP-SGD, and their soft labels allow students to learn cleaner decision boundaries than direct DP-SGD on the 1B model.
2. **DP-Synthetic fails due to low-fidelity text generation.** Even at relaxed budgets, synthetic data do not capture the clinical signal required for accurate coding.
3. **LoRA provides weak implicit privacy.** Low-rank updates slightly constrain memorisation but cannot substitute for formal DP guarantees.
4. **Privacy metrics saturate quickly.** Once DP noise is introduced, further relaxation of ε yields utility gains but negligible empirical privacy degradation.

4 Discussion

4.1 Why knowledge distillation wins

Among all differentially private methods, DP-Distil achieves the highest utility at moderate and relaxed privacy budgets. Its advantage stems from the ability of high-capacity teacher models to learn more robust representations under DP-SGD despite the injected noise, and to transfer these representations to a smaller student, consistent with prior observations that larger models can maintain higher utility under DP-SGD at fixed privacy budgets [Yu et al., 2021]. The use of soft labels plays a central role: rather than exposing the student to one-hot class assignments derived from noisy teacher predictions, the continuous logit distribution encodes relative class likelihoods, offering richer supervision and acting as a regulariser. This allows the student to recover cleaner decision boundaries even when the teachers' parameters are perturbed by privacy noise.

At very tight privacy budgets ($\varepsilon = 2$), DP-Small outperforms DP-Distil; the latter splits its budget between two teachers ($\varepsilon/2$ each), leaving insufficient signal for high-fidelity distillation. As ε increases, the teachers' larger capacity enables them to tolerate DP noise more effectively. By $\varepsilon = 6$, the 1B student achieves Micro-F_1 of 0.33, slightly exceeding the 3B teachers (0.32), consistent with regularisation benefits in prior distillation literature [Furlanello et al., 2018]. As shown in Table 3, students consistently match or outperform their teachers across all privacy budgets, with the performance gap widening at higher ε values where teachers have cleaner signals to transfer.

Taken together, these results highlight that deploying the 3B DP-trained teachers directly would offer no accuracy benefit over the 1B students while nearly doubling memory and latency (Table 5). Distillation therefore provides both higher utility and a more efficient deployed model, with lower latency and memory requirements, reinforcing its practicality for real-world clinical settings.

4.2 Why synthetic data fails

DP-Synthetic consistently underperforms, plateauing around Micro-F_1 of 0.22. Since its training procedure mirrors DP-Small but uses DP-generated text, this gap isolates synthetic data fidelity as the limiting factor. The DP generator struggles to reproduce the statistical and lexical diversity of true discharge summaries, leading to label-distribution drift and loss of rare diagnostic patterns.

As shown in Figure 5, this underperformance extends across nearly all individual codes—even the most frequent with ample training examples—confirming that synthetic data lack essential clinical signal rather than merely failing to cover rare patterns.

4.3 LoRA's modest implicit privacy

LoRA-No-DP achieves MIA AUC of 0.57, providing measurable but incomplete privacy protection. This represents a 13% relative increase in attack success over random guessing (AUC = 0.5), partially validating that low-rank constraints inherently limit memorisation [Malekmohammadi and Farnadi, 2025]. However, the gap to formal DP methods (AUC 0.49–0.51) confirms that LoRA alone cannot replace rigorous privacy guarantees in high-stakes clinical settings.

4.4 Limitations

Our study has several limitations. The 512-token truncation captures roughly 17% of each discharge summary, potentially omitting diagnostic context. Focusing on the 50 most frequent ICD-9 codes may not reflect performance on rare conditions. Our privacy audit evaluates only logit-based membership-inference attacks; other attack classes (e.g., attribute or extraction) remain outside scope. Finally, experiments used a single random seed and limited hyperparameter tuning. Despite these constraints, the observed trends remain robust and internally consistent across metrics and privacy budgets, supporting our central claim that architectural choices, alongside the privacy budget, substantially shape the privacy–utility trade-off.

4.5 Practical implications

Taken together, our results provide the first controlled evidence of how privacy mechanisms reshape model performance in clinical NLP at fixed capacity. When strict privacy is required ($\varepsilon \leq 2$), DP-Small remains the simplest and most compute-efficient option. For moderate budgets ($\varepsilon \geq 4$), DP-

Distil achieves the best balance of accuracy and protection, recovering up to 63% of the non-private baseline's performance whilst maintaining near-random MIA outcomes (AUC ≈ 0.5). DP-Synthetic is currently unsuitable for deployment, and LoRA should be treated purely as a non-DP upper bound.

Overall, knowledge distillation with differentially private teachers emerges as the most practical path toward deployable, privacy-preserving clinical language models, offering a clear direction for healthcare institutions seeking to combine diagnostic utility with rigorous confidentiality guarantees.

Future work should extend these comparisons to multi-hospital datasets and longer clinical narratives, establishing benchmarks for truly deployable privacy-preserving clinical language models.

Acknowledgments

This work was conducted as part of an MSc thesis at Imperial College London, supervised by Dr Andrew Duncan. We acknowledge Andy Thomas (Head of Research Computing, Imperial Mathematics) for providing access to computing resources, and the MIMIC-III team and PhysioNet for making the clinical database publicly available. Code is available at `https://github.com/mathieu-dufour/dp-clinical-coding`.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS'16. ACM, October 2016. doi: 10.1145/2976749.2978318. URL `http://dx.doi.org/10.1145/2976749.2978318`.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. *CoRR*, abs/2112.03570, 2021a. URL `https://arxiv.org/abs/2112.03570`.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models, 2021b. URL `https://arxiv.org/abs/2012.07805`.

Ali Dadsetan, Dorsa Soleymani, Xijie Zeng, and Frank Rudzicz. Can large language models be privacy preserving and fair medical coders?, 2024. URL `https://arxiv.org/abs/2412.05533`.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL `https://doi.org/10.1561/0400000042`.

James Flemings and Murali Annavaram. Differentially private knowledge distillation via synthetic text generation, 2024. URL `https://arxiv.org/abs/2403.00932`.

Tommaso Furlanello, Zachary C. Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks, 2018. URL `https://arxiv.org/abs/1805.04770`.

Aaron Grattafiori, Abhimanyu Dubey, Hugo Touvron, and et al. The llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL `https://arxiv.org/abs/2106.09685`.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):160035, 2016. doi: 10.1038/sdata.2016.35. URL `https://doi.org/10.1038/sdata.2016.35`.

Saber Malekmohammadi and Golnoosh Farnadi. Low-rank adaptation secretly imitates differentially private sgd, 2025. URL `https://arxiv.org/abs/2409.17538`.

Vaishnavh Nagarajan, Aditya Krishna Menon, Srinadh Bhojanapalli, Hossein Mobahi, and Sanjiv Kumar. On student-teacher deviations in distillation: does it pay to disobey?, 2024. URL https://arxiv.org/abs/2301.12923.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models, 2017. URL https://arxiv.org/abs/1610.05820.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in pytorch. *CoRR*, abs/2109.12298, 2021. URL https://arxiv.org/abs/2109.12298.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A. Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models. *CoRR*, abs/2110.06500, 2021. URL https://arxiv.org/abs/2110.06500.

Xiang Yue, Huseyin A. Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. Synthetic text generation with differential privacy: A simple and practical recipe, 2023. URL https://arxiv.org/abs/2210.14348.

A Supplementary material

A.1 Detailed dataset statistics

MIMIC-III discharge summaries were filtered to 49,426 admissions with at least one top-50 ICD-9 code, split 80/10/10% into training (44,778 notes from 39,540 admissions), validation (5,624 notes), and test (5,586 notes) sets. Notes average 2,993 tokens (before truncation to 512) and 4.5 codes. Top-5 codes: 401.9 (hypertension, n=17,920), 428.0 (heart failure, n=11,933), 427.31 (atrial fibrillation, n=11,517), 414.01 (coronary atherosclerosis, n=11,011), 584.9 (acute kidney failure, n=7,936).

A.2 Full training details

All models trained for maximum 18 epochs with early stopping (patience=3). DP methods used SGD with momentum 0.9 and learning rates of 1.5e-3 for 1B models and 2e-4 to 3e-4 for 3B models. Non-DP methods used AdamW with cosine scheduling and learning rates of 5e-4 to 7.5e-4. Gradient clipping norms: C=1.0 (1B), C=0.7 (3B). Batch sizes were maximised within 48GB GPU memory constraints. Synthetic generation used nucleus sampling (p=0.9, temperature=0.8).

A.3 Training efficiency and resource requirements

Table 4: Wall-clock training time by pipeline using a single RTX 6000 Ada GPU (48GB VRAM) with maximum batch sizes. Times in hours.

Pipeline	ε	Gen. Train	Synth. Gen.	Teacher Train	Student/ Classifier	Total (h)	Peak VRAM (GB)
LoRA-No-DP	∞	–	–	–	3.07	3.07	41.11
DP-Small	2	–	–	–	4.91	4.91	42.24
DP-Small	4	–	–	–	5.60	5.60	42.24
DP-Small	6	–	–	–	9.56	9.56	42.24
DP-Synthetic	2	8.45	3.33	–	2.67	14.45	41.11
DP-Synthetic	4	8.72	3.33	–	3.16	15.21	41.11
DP-Synthetic	6	8.53	3.33	–	2.68	14.55	41.11
DP-Distil	2	10.86	3.33	18.82	5.06	38.06	41.11
DP-Distil	4	11.05	3.33	21.59	2.34	38.31	41.11
DP-Distil	6	11.34	3.33	29.93	4.70	49.29	41.11

Table 5: Inference metrics (mean across all 1B classifiers and 3B DP-Distil teachers) measured on a single RTX 6000 Ada GPU with a batch size of 64 and a maximum sequence length of 512 tokens.

Model	Latency (ms/example)	Throughput (tokens/s)	Peak VRAM (MB)	Adapter (MB)
1B classifier	13.06	38,376	4,432	520
3B classifier (DP-Distil teacher)	35.65	14,054	8,462	1,529

A.4 Code availability

Complete code for data preprocessing, model training, and evaluation is available at `https://github.com/mathieu-dufour/dp-clinical-coding` under a CC-BY-4.0 licence. The repository includes a clear README describing how to reproduce results, and numbered scripts to run in sequence to perform the full research computations.