Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Towards privacy-safe synthetic French clinical documents

Riccardo Tripodi*, Simon Meoni†
*Assistance Publique – Hôpitaux de Paris, 75012 Paris, France
†ALMAnaCH team, Inria Paris, France

RÉSUMÉ

Les modèles de langues médicaux requierent de grands corpus annotés, rares en français et soumis à des restrictions en matière de confidentialité. Nous présentons un pipeline préservant la confidentialité qui produit un LLM médical français déployable sans utiliser de textes de patients pendant l'entraînement. À partir d'une correspondance établie entre les codes diagnostics CIM-10 et les mots-clés médicaux, un LLM médical génère des rapports synthétiques associés aux codes, formant ainsi un jeu de données de base exempt de dossiers cliniques réels. Ce jeu de données permet d'entraîner un modèle local plus petit, ensuite affiné de manière itérative au sein de l'hôpital sans qu'aucune donnée confidentielle ne quitte l'établissement. Pour évaluer l'utilité clinique de cette méthodologie, nous entraînons des classificateurs CIM-10 exclusivement sur les rapports synthétiques produits à chaque étape de raffinement. L'amélioration de la qualité de génération se traduit par de meilleures performances sur des données réelles, traduisant une pertinence médicale accrue. Ces résultats montrent que des connaissances structurées et un retour d'information sécurisé permettent le transfert de compétence clinique à un LLM français léger tout en produisant un ensemble de données synthétiques partageables bienvenues pour un NLP médical manquant de ressources.

MOTS-CLÉS : Génération de données synthétiques, Codage CIM-10, NLP médical, Apprentissage par renforcement, Confidentialité des données, Rapports de sortie.

KEYWORDS: Synthetic Data Generation, ICD-10 Coding, Clinical NLP, Reinforcement Learning, Data Privacy, Discharge Summaries.

1 Introduction

Electronic Health Records contain detailed clinical narratives describing patient conditions, medical reasoning, and care decisions. These documents are essential for billing, epidemiology, and clinical decision support (Rosenbloom *et al.*, 2011), yet their free-text structure complicates automated analysis. In particular, assigning ICD-10 codes from narrative reports remains a central but labor-intensive task in hospital workflows (Stanfill *et al.*, 2010).

The International Classification of Diseases, Tenth Revision (ICD-10), is a standardized diagnostic taxonomy maintained by the World Health Organization (Organization, 2004). It provides a hierarchical coding system used worldwide for disease reporting, reimbursement, and statistical analysis. Each clinical encounter may be associated with multiple ICD-10 codes, reflecting comorbidities and diagnostic complexity, making the coding process inherently multi-label and often ambiguous. Table 1 presents some representative ICD-10 categories together with illustrative clinical keywords associated with each code.

Actes des 19e Rencontres Jeunes Chercheurs en RI (RJCRI) et 28ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL), pages 68–79.

ICD-10 Code	Diagnosis	Example Keywords
I21	Acute myocardial infarction	chest pain, troponin, ST elevation
J18	Pneumonia, unspecified organism	fever, cough, pulmonary infiltrate
E11	Type 2 diabetes mellitus	hyperglycemia, insulin therapy
K35	Acute appendicitis	CRP elevation , ultrasound
C50	Malignant neoplasm of breast	breast mass, biopsy, chemotherapy

Table 1: Illustrative examples of ICD-10 codes and associated clinical terminology.

Recent progress in large language models has significantly improved clinical text understanding and generation (Agrawal *et al.*, 2022; Singhal *et al.*, 2023). However, directly deploying large proprietary models in healthcare environments raises several practical barriers. First, patient data cannot be transmitted outside secure infrastructures. Second, institutional practices vary, requiring locally adapted models. Third, many hospitals lack the computational resources needed to operate multi-billion-parameter systems in production settings.

Synthetic data generation has emerged as a promising strategy to mitigate data scarcity while preserving confidentiality (Melamud & Shivade, 2019). Previous approaches generate clinical narratives guided by keywords and iteratively refine them through preference-based alignment (Meoni *et al.*, 2024). While effective, these methods still rely on real anonymized reports to bootstrap training and primarily aim at improving generation quality.

In this work, our objective is not to replace clinical coding systems or assist physicians in drafting reports, but to generate privacy-preserving synthetic clinical datasets that can be used to train downstream medical NLP models. We focus specifically on whether synthetic reports can transfer diagnostically meaningful information to models trained without access to real patient text.

To achieve this, we transfer clinical knowledge from a powerful multilingual medical language model into a compact deployable generator while avoiding the need to construct or curate a shareable corpus of anonymized clinical records. Although real clinical documents are accessed within the secure hospital environment, they are never exposed or directly used for model training. Instead, generation begins from structured medical knowledge. We construct prompts from ICD-10-associated keywords and use a large medical language model to synthesize clinical reports paired with diagnostic codes. These automatically generated report–code pairs form a privacy-safe seed dataset used to instruction-tune a compact local model.

To bridge the remaining domain gap between synthetic and real hospital narratives, we introduce a privacy-bounded refinement stage. The compact model generates candidate reports locally, which are compared against private clinical documents entirely inside the hospital infrastructure. Only aggregate scalar preference signals are exported, enabling iterative preference optimization without exposing patient text or embeddings.

In summary, this work introduces a privacy-preserving approach for building a French clinical language model from structured medical knowledge rather than patient records. Our contributions are:

- **Knowledge-driven synthetic corpus creation:** we generate report–ICD-10 pairs from a curated keyword collection, avoiding the use of clinical text.

- **Distillation into a lightweight model:** the synthetic corpus enables training of a compact generator that can be deployed on standard hospital hardware.
- **Secure iterative refinement:** model outputs are compared with private clinical data inside the hospital environment, and only numerical feedback is returned, ensuring no data leakage.
- **Downstream clinical validation:** ICD-10 classifiers trained solely on synthetic reports show increasing accuracy on real data as the generator improves.
- **Resources for French medical NLP:** we release the trained model and a large synthetic dataset to support research in an under-resourced clinical language.

2 Related Work

Synthetic Data Generation. The generation of synthetic clinical text has emerged as a practical approach to mitigate data access restrictions and privacy constraints in medical NLP. Recent work explores the use of large language models (LLMs) to produce artificial clinical corpora suitable for downstream supervision. For instance, Kweon *et al.* (2024) generate synthetic clinical notes using online LLMs to enable public model development without exposing sensitive patient information. In parallel, Xie *et al.* (2024) propose AUG-PE, a framework for differentially private synthetic text generation built on top of foundation model APIs, explicitly incorporating formal privacy guarantees into the generation pipeline.

Beyond straightforward data augmentation, alternative strategies focus on structured generation processes. Li *et al.* (2024) introduce Generalized Instruction Tuning (GLAN), which constructs large-scale instruction datasets from a curated taxonomy of human capabilities rather than relying exclusively on pre-existing corpora. This taxonomy-driven perspective highlights the role of structured knowledge in guiding synthetic data creation. In the clinical setting, synthetic datasets have also been developed to model domain-specific phenomena under controlled conditions. Li *et al.* (2023a) generate Alzheimer's-related synthetic datasets grounded in expert-defined taxonomies to capture clinically meaningful symptom structures. For information extraction tasks, Hiebel *et al.* (2023) and Xie *et al.* (2024) examine the effectiveness of synthetic clinical corpora for named entity recognition (NER), evaluating performance trade-offs between artificial and authentic electronic health record data.

Self-Training and Preference Optimization. A complementary research direction investigates self-training and alignment techniques that leverage model-generated supervision. Reinforced Self-Training (ReST) (Gulcehre *et al.*, 2023) formulates self-alignment as an offline reinforcement learning procedure, iteratively improving a policy using trajectories sampled from the model itself. Instruction back-translation (Li *et al.*, 2023b) scales supervision by automatically generating instruction–response pairs from limited seed annotations combined with large web corpora.

More recent work employ language models as evaluators to produce reward signals for optimization. Yuan *et al.* (2024) propose self-rewarding language models in which an LLM generates both outputs and associated reward signals through structured prompting. Reinforcement Learning from AI Feedback (RLAIF) (Lee *et al.*, 2024) similarly replaces human preference annotations with model-generated preference labels, demonstrating competitive performance relative to traditional RLHF

pipelines. Direct Preference Optimization (DPO) (Rafailov *et al.*, 2023) further simplifies preference-based alignment by directly optimizing over ranked output pairs, eliminating the need for an explicit reward model while retaining the advantages of preference learning.

3 Methodology

Our objective is to develop a lightweight French clinical report generator without requiring the collection or anonymization of clinical documents for training. Instead of constructing a privacy-free corpus, we rely on synthetic data generation and secure in-hospital evaluation to transfer clinical knowledge while maintaining strict privacy guarantees. Real patient reports are never used as training data and never leave the secure environment. The pipeline comprises four sequential stages: medically-driven keyword prompting, supervised instruction tuning of a compact model, iterative refinement using privacy-bounded scoring, and downstream validation through ICD-10 classification. An example of a synthetic medical report generated at the end of this pipeline is provided in Box 3.

Medically-driven keyword prompting. We construct prompts from a curated database $\mathcal{D}_{\text{ICD}}$ linking ICD-10 codes $c \in \mathcal{C}$ to medically grounded French keywords describing symptoms, anatomy, treatments, and procedures. The database contains only privacy-safe terminology derived from public medical knowledge bases. Codes are sampled according to empirical hospital frequency distributions $p(c)$, after which a constrained subset of associated keywords $K_c \subseteq \mathcal{K}(c)$ with $|K_c| \leq K_{\max}$ is selected to maintain instruction adherence.

A large medical language model generates clinical reports in French following a fixed narrative structure comprising eight sections: *Motif d'hospitalisation*, *Antécédents*, *Mode de vie*, *Histoire de la maladie*, *Examen clinique*, *Examens complémentaires*, *Évolution pendant l'hospitalisation*, and *Conclusion*. The model must incorporate all keywords in K_c while maintaining medical coherence. This produces a seed dataset $\mathcal{D}_{\text{seed}} = \{(r_i, \mathbf{c}_i, K_{\mathbf{c}_i})\}_{i=1}^{N_{\text{seed}}}$ where each sample consists of a synthetic report r_i, a set of ICD-10 codes $\mathbf{c}_i = \{c_i^{(1)}, \ldots, c_i^{(m_i)}\} \subseteq \mathcal{C}$ (with m_i the number of codes for sample i), and the set of keywords $K_{\mathbf{c}_i}$ comprising all keywords associated with any code in $\mathbf{c}_i$. This corpus $\mathcal{D}_{\text{seed}}$ is entirely independent of real patient records and used to initiate a generator model M_0.

Supervised instruction tuning of a compact generator. We train a deployment-efficient generator by instruction-tuning a base model on $\mathcal{D}_{\text{seed}}$ (Wei *et al.*, 2021). During training, the model receives simplified prompts conditioned on the same keyword sets $K_{\mathbf{c}_i}$ for each sample, reducing context-window consumption. The base model initially produces inconsistent and non-medical French text. Instruction tuning transfers domain vocabulary and document conventions from the standardized synthetic reports, yielding an initial generator M_0 that maps keyword prompts to structured medical reports in French.

Iterative refinement with privacy-bounded scoring. Following supervised fine-tuning, we iteratively improve the generator using feedback derived from private hospital reports $\mathcal{R}_{\text{priv}}$. The critical constraint is that no private document content ever crosses the privacy boundary: raw reports remain strictly local, and only privacy-bounded aggregate signals are exposed to the training loop.

Within the hospital infrastructure, medical keywords are extracted from each private reference report $r_{\text{ref}} \in \mathcal{R}_{\text{priv}}$ using QuickUMLS (Soldaini & Goharian, 2016), producing a collection of privacy-safe prompt pairs $\mathcal{K}_{\text{priv}} = \{(r_{\text{ref}}, K)\}$. For each pair $(r_{\text{ref}}, K) \in \mathcal{K}_{\text{priv}}$, the current generator model M_t produces N candidate reports $\{r_1, \ldots, r_N\} \sim M_t(\cdot \mid K)$.

Each candidate r_i is evaluated against its corresponding private reference report r_{ref} using a composite score s_i computed entirely within the secure boundary. The total score is defined as the average of two components: a cosine similarity score derived from embedding representations (Reimers & Gurevych, 2019) and a utility score from an LLM-as-a-judge (Zheng *et al.*, 2023) rating the medical correctness and content alignment of the generated candidate with the reference report.

Preference pairs (r^+, r^-) are constructed by selecting the highest-scoring candidate as r^+ and a lower-scoring candidate as r^-. The generator is updated using Direct Preference Optimization (DPO) (Rafailov *et al.*, 2023), yielding M_{t+1}. This procedure of generating candidates, scoring them against private references, and applying DPO is repeated for T iterations, producing a sequence of refined generator models. In our experiments, we set $T = 3$. Algorithm 1 formalizes the complete procedure.

Algorithm 1 Complete Iterative Privacy-Aware Refinement Pipeline

Require: ICD-10 keyword database $\mathcal{D}_{\text{ICD}}$, large generator G_{large}, base generator G_{base}, private hospital reports $\mathcal{R}_{\text{priv}}$, max keywords $K_{\max}$, seed size N_{seed}, iterations T, candidates per prompt N, code distribution p, keyword mapping $\mathcal{K}$

Ensure: Refined generator M_T

Step 1: Seed synthesis with G_{large}
$\mathcal{D}_{\text{seed}} \leftarrow \emptyset$
for $i = 1$ to N_{seed} **do**
  Sample code set $\mathbf{c}_i \subseteq \mathcal{C}$ from $p(\mathbf{c})$
  $K_{\mathbf{c}_i} \leftarrow \{K_c \sim \mathcal{U}(\mathcal{K}(c)) \mid c \in \mathbf{c}_i, |K_c| \leq K_{\max}\}$
  $P_i \leftarrow \textsc{MakePrompt}(\mathbf{c}_i, K_{\mathbf{c}_i})$; $r_i \sim G_{\text{large}}(\cdot \mid P_i)$
  $\mathcal{D}_{\text{seed}} \leftarrow \mathcal{D}_{\text{seed}} \cup \{(r_i, \mathbf{c}_i, K_{\mathbf{c}_i})\}$
end for
Step 2: Supervised fine-tuning for initial generator model M_0
$M_0 \leftarrow \text{SFT}(G_{\text{base}}, \mathcal{D}_{\text{seed}})$
Step 3: Private prompt extraction
$\mathcal{K}_{\text{priv}} \leftarrow \{(r_{\text{ref}}, \textsc{ExtractKeywords}(r_{\text{ref}})) : r_{\text{ref}} \in \mathcal{R}_{\text{priv}}\}$ ▷ computed in-hospital
Step 4: Iterative refinement of M_t
for $t = 0$ to $T - 1$ **do**
  $\mathcal{P} \leftarrow \emptyset$
  for each $(r_{\text{ref}}, K) \in \mathcal{K}_{\text{priv}}$ **do**
    Generate $r_1, \ldots, r_N \sim M_t(\cdot \mid K)$ ▷ computed in-hospital
    Score $s_i \leftarrow \textsc{Score}(r_i, r_{\text{ref}})$ ▷ computed in-hospital
    $r^+ \leftarrow r_{\arg\max_i s_i}$; $r^- \leftarrow$ sample from $\{r_i : s_i < \max_j s_j\}$
    $\mathcal{P} \leftarrow \mathcal{P} \cup \{(K, r^+, r^-)\}$
  end for
  $M_{t+1} \leftarrow \text{DPO}(M_t, \mathcal{P})$
end for
return M_T

Downstream ICD-10 classification validation. We assess clinical utility through extrinsic evaluation on a downstream diagnostic coding task. For each refinement stage $t \in \{0, \ldots, T\}$, we sample keyword sets from $\mathcal{K}(c)$ and use generator M_t to produce reports, yielding a synthetic dataset $\mathcal{S}_t = \{(r, \mathbf{c})\}$. A multi-label ICD-10 classifier C_t is trained exclusively on $\mathcal{S}_t$ and evaluated on held-out real hospital documents $\mathcal{R}_{\text{test}}$. Improvements in classification metrics across stages indicate that synthetic reports increasingly encode diagnostically relevant information and that preference-based alignment improves clinically meaningful content.

Example generation

```
Motif d'hospitalisation: Mme Dupont, 68 ans, est admise en
hospitalisation de jour le 15/03/2024 pour traitement d'une
angiodermite nécrotique extensive surinfectée. Cette lésion cutanée
chronique, initialement diagnostiquée il y a plusieurs années, s'est
aggravée de manière significative, nécessitant une prise en charge
intensive et une surveillance rapprochée.

Histoire de la maladie: L'angiodermite nécrotique initiale de Mme
Dupont a connu des épisodes récurrents et des aggravations
progressives au fil des ans. Depuis environ 6 mois, une lésion
cutanée extensive, initialement localisée au niveau des membres
inférieurs, s'est étendue progressivement et s'est caractérisée par
une nécrose cutanée et une infection bactérienne sévère. La lésion
est devenue très douloureuse, avec des signes d'inflammation marquée
(érythème, chaleur, œdème) et une ulcération profonde. La patiente a
consulté son dermatologue qui a initié un traitement antibiotique
(Ceftriaxone) et une antibiothérapie topique, mais l'aggravation
persiste. Face à l'aggravation clinique et l'extension de la lésion,
une hospitalisation a été jugée nécessaire pour une prise en charge
multidisciplinaire et un traitement intensifié.

... le reste du compte rendu ...

Mots-clés: Angiodermite nécrotique extensive surinfectée, personne
vivant seule à son domicile.
```

4 Experiments

4.1 Experimental Setup

We first construct the keyword–code collection $\mathcal{D}_{\text{ICD}}$, a curated database linking ICD-10 diagnostic categories to medically grounded French keywords describing symptoms, anatomy, treatments, examinations, and procedures. The content of $\mathcal{D}_{\text{ICD}}$ is compiled from internal hospital resources together with publicly available medical documentation and terminology sources.

The seed corpus $\mathcal{D}_{\text{seed}}$ with size $N_{\text{seed}} = 20\,000$ is generated using `medgemma-27b-it` (Sellergren *et al.*, 2025), a large multilingual medical language model optimized for clinical text understanding and reasoning. MedGemma has been trained on diverse medical data, including clinical narratives, medical question–answer pairs, and structured health record information. The 27B text-only instruction-tuned

variant is specifically designed for high-quality medical text generation and inference-time reasoning, making it well suited for synthesizing clinically coherent reports in French.

The initial base generator is `Qwen3-4B-Instruct` (Yang *et al.*, 2025). It is fine-tuned on the seed corpus using LoRA (Hu *et al.*, 2022) with rank $r = 32$ and scaling factor $\alpha = 64$, applied to all attention projection layers (query, key, value, and output). Training runs for up to 10 epochs with an effective batch size 8. We use bfloat16 mixed precision with a maximum sequence length of 512 tokens, AdamW optimizer (Loshchilov & Hutter, 2017), a learning rate of 1×10^{-5} with cosine scheduling and 10% warmup, and early stopping based on validation cross-entropy loss.

During each refinement iteration, candidate reports are generated to construct preference pairs. For each keyword prompt, we generate $N = 4$ candidates using a mixed temperature strategy: one sample with fixed temperature $\tau = 0.2$ to favor high-likelihood coherent outputs, and three samples with temperatures independently drawn from a uniform distribution, $\tau \sim \mathcal{U}(0.6, 0.9)$, to encourage diversity. This setup provides contrast between reliable and varied generations, strengthening the alignment signal. Each candidate's score is computed as the average of two components: the cosine similarity obtained from `BioLORD-2023-M` embeddings (Remy *et al.*, 2024) and a medical correctness and content alignment score produced by `Qwen3-30B-A3B-Instruct-2507` acting as an LLM-as-a-judge.

While the refinement framework is designed to operate on private clinical documents $\mathcal{R}_{\text{priv}}$, in this study $\mathcal{R}_{\text{priv}}$ consists of fictive reports written by practicing physicians across multiple hospital departments and medical specialties. This ensures clinically realistic content while making the methodology fully reproducible and sharable, and eliminates any risk of privacy leakage due to anonymization or implementation errors.

Preference optimization is then performed for up to 10 epochs with AdamW optimizer and learning rate 5×10^{-6} using the same LoRA configuration ($r = 32$, $\alpha = 64$). Training uses bfloat16 precision with per-device batch size 8 and gradient accumulation over 4 steps, and the maximum sequence length is increased to 3000 tokens to accommodate the reports in their entirety. The best checkpoint is selected according to evaluation loss.

We evaluate the proposed synthetic data generation pipeline through a downstream ICD-10 multi-label multi-class classification task, assessing whether iterative preference-based refinement improves the clinical utility of generated reports.

For classification, we fine-tune `ModernBERT-large` (Warner *et al.*, 2025), a modernized bidirectional encoder-only Transformer containing 395 million parameters and designed for long documents. Training uses a peak learning rate of $1\text{e}{-4}$ with cosine learning-rate scheduling and early stopping based on validation macro-F1.

Synthetic training sets are generated at three scales: 10,000, 20,000, and 50,000 samples. At each scale, datasets are produced across four refinement stages: the initial supervised fine-tuned generator (Step 0) followed by three successive preference optimization iterations (DPO1, DPO2, DPO3). Classification targets are defined by the top-k most frequent ICD-10 codes, using only the initial letter and two digits that specify the core diagnostic category, with $k \in 20, 50, 100$.

Increasing k results in progressively more challenging classification settings: the label space expands to include rarer diagnoses, reducing the number of training examples per class and amplifying label imbalance. Because ICD-10 codes are sampled from a fixed distribution at generation time across all refinement stages, per-code frequencies scale linearly with dataset size. In the 10k setting, this yields

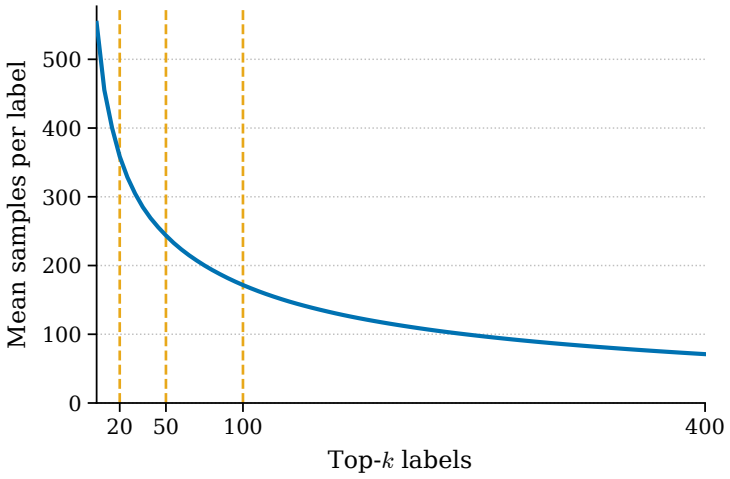

Figure 1: Mean training samples per ICD-10 code across top-k subsets for the DPO3 10k synthetic dataset.

approximately 350 samples per code for top-20 but only about 150 for top-100. Larger datasets (20k and 50k) increase these counts proportionally, while preserving the relative imbalance induced by higher k values. Concurrently, the average number of codes per document increases with k due to filtering: documents in the top-20 split contain 1.52 ± 0.82 codes on average, rising to 2.10 ± 1.14 for top-100 (from 3.37 ± 1.36 in the full unfiltered corpus). Across refinement stages, the average report length is approximately 800 ± 240 words. Figure 1 illustrates the samples-per-code distribution across the three top-k configurations..

Evaluation uses macro-averaged F1 scores computed on a strictly held-out private test set of real clinical reports annotated by experts. The test set is fully disjoint from all data used in the previously described pipeline including supervised training, validation, preference construction, scoring, and any synthetic or fictive reports. This ensures that performance gains reflect genuine transfer to authentic clinical documents rather than adaptation to artifacts of the training or refinement process.

4.2 Results

Table 2 reports classification performance across dataset sizes and refinement stages.

As a reference, a simple rule-based baseline using QuickUMLS (Soldaini & Goharian, 2016) obtains macro-F1 scores of 0.0151, 0.0091, and 0.0103 for the top-20, top-50, and top-100 label settings, respectively. The baseline operates through fuzzy matching between report text and our curated ICD-10 keyword collection: whenever one or more keywords associated with a diagnostic code are detected in a document, the corresponding ICD-10 code is assigned as a prediction.

Three patterns are immediately evident. First, iterative refinement consistently improves downstream performance. Across all settings, DPO iterations outperform the initial supervised model, with the largest gains observed after the full refinement cycle. This confirms that preference-based alignment enhances the clinical usefulness of the generated reports rather than merely improving surface quality.

Second, the benefit of refinement increases with task difficulty. For top-20 classification, improvements remain modest (2–7%), reflecting the relative ease of predicting frequent diagnoses. In contrast, top-100 classification shows substantial gains (36–46%), indicating that refinement is particularly ef-

Size	Top-k	Step 0	DPO1	DPO2	DPO3	Gain
10k	20	0.389	**0.422**	0.422	0.418	+7.4%
	50	0.250	0.294	0.302	**0.317**	**+26.9%**
	100	0.175	0.197	0.202	**0.238**	**+36.5%**
20k	20	0.438	**0.457**	0.448	0.447	+2.1%
	50	0.299	0.326	0.334	**0.347**	**+16.2%**
	100	0.195	0.230	0.210	**0.285**	**+45.8%**
50k	20	0.445	**0.461**	0.461	0.455	+2.3%
	50	0.326	0.351	**0.385**	0.383	**+17.6%**
	100	0.230	0.268	0.300	**0.316**	**+37.2%**

Table 2: ICD-10 classification macro F1 scores across refinement iterations and dataset sizes. Best score per row in bold. Gain indicates relative improvement of DPO3 over Step 0.

fective for rare and underrepresented codes. This suggests that alignment improves the discriminative content of synthetic reports, especially for long-tail conditions where label sparsity is most severe.

Third, dataset size primarily affects absolute performance rather than relative gains. Larger synthetic corpora yield higher F1 scores at every stage, confirming that data volume alone contributes to classification quality. However, refinement remains beneficial even at 50k samples, particularly for top-100 codes, demonstrating that generation quality and alignment are not fully substituted by scale.

Performance trends across refinement steps further support this interpretation. For easier tasks, improvements plateau after one or two iterations, whereas harder tasks continue to benefit from later refinement. This indicates diminishing returns for frequent codes but sustained gains for rare diagnoses.

We do not compare against fully supervised training on real clinical reports because the available private dataset is relatively small and highly imbalanced across ICD-10 categories, particularly for rare diagnoses. Under these conditions, performance estimates from real-only training would be unstable and difficult to generalize. Instead, our objective is to evaluate whether large-scale synthetic supervision can provide diagnostically meaningful signals that transfer to real clinical documents despite limited access to annotated hospital data.

Overall, the results show that synthetic data generation, when combined with secure preference alignment, can effectively support ICD-10 multi-label classification. While absolute performance remains dependent on dataset size and label distribution, the consistent improvements on held-out real clinical documents demonstrate that privacy-preserving synthetic pipelines can produce training data that transfers meaningfully to real-world diagnostic coding.

5 Conclusion

We presented a privacy-aware framework for training a French clinical language model without constructing or anonymizing patient corpora. By combining large-model synthetic generation with secure in-hospital alignment, the approach transfers medical knowledge into a compact, deployable generator while preserving strict data confidentiality. Evaluation on ICD-10 multi-label classification shows that

iterative refinement improves downstream performance, particularly for rare and underrepresented codes, demonstrating that aligned synthetic data can encode clinically meaningful information beyond surface fluency.

Several limitations suggest directions for future work. Our evaluation is restricted to ICD-10 classification, and extending validation to additional clinical NLP tasks would provide a broader assessment of generalization. The refinement process currently uses a fixed number of iterations, and adaptive stopping criteria could improve efficiency. The controlled generation of report–code pairs may also enable reinforcement-based training of language models directly as diagnostic classifiers, leveraging synthetic supervision to further bridge the gap between generation and coding.

Overall, the proposed framework shows that privacy-constrained synthetic data generation is a viable strategy for developing clinical NLP systems in mid-resource settings. By decoupling model training from direct access to patient text, it offers a practical pathway for advancing French clinical language technologies where data sharing is limited.

To facilitate reproducibility, we release the synthetic datasets, models[1] and full codebase[2].

References

AGRAWAL M., HEGSELMANN S., LANG H., KIM Y. & SONTAG D. (2022). Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, p. 1998–2022.

GULCEHRE C., PAINE T. L., SRINIVASAN S., KONYUSHKOVA K., WEERTS L., SHARMA A., SIDDHANT A., AHERN A., WANG M., GU C. *et al.* (2023). Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*.

HIEBEL N., FERRET O., FORT K. & NÉVÉOL A. (2023). Can synthetic text help clinical named entity recognition? a study of electronic health records in French. In A. VLACHOS & I. AUGENSTEIN, Éds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, p. 2320–2338, Dubrovnik, Croatia: Association for Computational Linguistics. DOI : 10.18653/v1/2023.eacl-main.170.

HU E. J., SHEN Y., WALLIS P., ALLEN-ZHU Z., LI Y., WANG S., WANG L., CHEN W. *et al.* (2022). Lora: Low-rank adaptation of large language models. *Iclr*, **1**(2), 3.

KWEON S., KIM J., KIM J., IM S., CHO E., BAE S., OH J., LEE G., MOON J. H., YOU S. C. *et al.* (2024). Publicly shareable clinical large language model built on synthetic clinical notes. In *Findings of the Association for Computational Linguistics: ACL 2024*, p. 5148–5168.

LEE H., PHATALE S., MANSOOR H., MESNARD T., FERRET J., LU K. R., BISHOP C., HALL E., CARBUNE V., RASTOGI A. & PRAKASH S. (2024). RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. In R. SALAKHUTDINOV, Z. KOLTER, K. HELLER, A. WELLER, N. OLIVER, J. SCARLETT & F. BERKENKAMP, Éds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 de *Proceedings of Machine Learning Research*, p. 26874–26901: PMLR.

[1] https://huggingface.co/collections/3podi/synth-fr
[2] https://github.com/3podi/synth_fr

LI H., DONG Q., TANG Z., WANG C., ZHANG X., HUANG H., HUANG S., HUANG X., HUANG Z., ZHANG D. *et al.* (2024). Synthetic data (almost) from scratch: Generalized instruction tuning for language models. *arXiv preprint arXiv:2402.13064*.

LI R., WANG X. & YU H. (2023a). Two directions for clinical data generation with large language models: Data-to-label and label-to-data. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, p. 7129–7143, Singapore: Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.474.

LI X., YU P., ZHOU C., SCHICK T., LEVY O., ZETTLEMOYER L., WESTON J. & LEWIS M. (2023b). Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259*.

LOSHCHILOV I. & HUTTER F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

MELAMUD O. & SHIVADE C. (2019). Towards automatic generation of shareable synthetic clinical notes using neural language models. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, p. 35–45.

MEONI S., DE LA CLERGERIE É. & RYFFEL T. (2024). Generating english synthetic documents with clinical keywords: A privacy-sensitive methodology. *LREC-COLING 2024*, p. 115.

ORGANIZATION W. H. (2004). *International Statistical Classification of Diseases and related health problems: Alphabetical index*, volume 3. World Health Organization.

RAFAILOV R., SHARMA A., MITCHELL E., MANNING C. D., ERMON S. & FINN C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, **36**, 53728–53741.

REIMERS N. & GUREVYCH I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, p. 3982–3992.

REMY F., DEMUYNCK K. & DEMEESTER T. (2024). BioLORD-2023: semantic textual representations fusing large language models and clinical knowledge graph insights. *Journal of the American Medical Informatics Association*, p. ocae029. DOI : 10.1093/jamia/ocae029.

ROSENBLOOM S. T., DENNY J. C., XU H., LORENZI N., STEAD W. W. & JOHNSON K. B. (2011). Data from clinical notes: a perspective on the tension between structure and flexible documentation. *Journal of the American Medical Informatics Association*, **18**(2), 181–186.

SELLERGREN A., KAZEMZADEH S., JAROENSRI T., KIRALY A., TRAVERSE M., KOHLBERGER T., XU S., JAMIL F., HUGHES C., LAU C. *et al.* (2025). Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

SINGHAL K., AZIZI S., TU T., MAHDAVI S. S., WEI J., CHUNG H. W., SCALES N., TANWANI A., COLE-LEWIS H., PFOHL S. *et al.* (2023). Large language models encode clinical knowledge. *Nature*, **620**(7972), 172–180.

SOLDAINI L. & GOHARIAN N. (2016). Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, p. 1–4.

STANFILL M. H., WILLIAMS M., FENTON S. H., JENDERS R. A. & HERSH W. R. (2010). A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association*, **17**(6), 646–651.

WARNER B., CHAFFIN A., CLAVIÉ B., WELLER O., HALLSTRÖM O., TAGHADOUINI S., GALLAGHER A., BISWAS R., LADHAK F., AARSEN T., ADAMS G. T., HOWARD J. & POLI I. (2025). Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In W. CHE, J. NABENDE, E. SHUTOVA & M. T. PILEHVAR, Éds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 2526–2547, Vienna, Austria: Association for Computational Linguistics. DOI : 10.18653/v1/2025.acl-long.127.

WEI J., BOSMA M., ZHAO V. Y., GUU K., YU A. W., LESTER B., DU N., DAI A. M. & LE Q. V. (2021). Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

XIE C., LIN Z., BACKURS A., GOPI S., YU D., INAN H. A., NORI H., JIANG H., ZHANG H., LEE Y. T., LI B. & YEKHANIN S. (2024). Differentially private synthetic data via foundation model APIs 2: Text. In R. SALAKHUTDINOV, Z. KOLTER, K. HELLER, A. WELLER, N. OLIVER, J. SCARLETT & F. BERKENKAMP, Éds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 de *Proceedings of Machine Learning Research*, p. 54531–54560: PMLR.

YANG A., LI A., YANG B., ZHANG B., HUI B., ZHENG B., YU B., GAO C., HUANG C., LV C. *et al.* (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

YUAN W., PANG R. Y., CHO K., LI X., SUKHBAATAR S., XU J. & WESTON J. E. (2024). Self-rewarding language models. In R. SALAKHUTDINOV, Z. KOLTER, K. HELLER, A. WELLER, N. OLIVER, J. SCARLETT & F. BERKENKAMP, Éds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 de *Proceedings of Machine Learning Research*, p. 57905–57923: PMLR.

ZHENG L., CHIANG W.-L., SHENG Y., ZHUANG S., WU Z., ZHUANG Y., LIN Z., LI Z., LI D., XING E. *et al.* (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, **36**, 46595–46623.