# Évaluation de la confidentialité des textes cliniques synthétiques générés par des modèles de langue

Foucauld Estignard[1] Sahar Ghannay[1] Julien Girard-Satabin[2]
Nicolas Hiebel[1] Aurélie Névéol[1]
(1) Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France
(2) Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
[1]prenom.nom@lisn.upsaclay.fr, [2]julien.girard2@cea.fr

## RÉSUMÉ

Les grands modèles de langue (LLM) peuvent être utilisés pour produire des documents synthétiques similaires à des documents réels dont la disponibilité est limitée pour des raisons de confidentialité ou de droits d'auteur. Dans cet article, nous étudions les risques en lien avec la confidentialité dans les documents générés automatiquement. Nous utilisons des textes synthétiques générés à partir d'un modèle pré-entraîné et affiné sur des cas cliniques en français afin d'évaluer ces risques selon trois critères : (1) la similarité entre un corpus d'entraînement réel et le corpus synthétique (2) les corrélations entre les caractéristiques cliniques dans le corpus d'entraînement et le corpus synthétique et (3) une attaque par inférence d'appartenance (*MIA*, en anglais) utilisant un modèle affiné sur le corpus synthétique. Nous identifions des associations de caractéristiques cliniques qui suggèrent que le filtrage du corpus d'entraînement pourrait contribuer à la préservation de la confidentialité. Les attaques par inférence d'appartenance n'ont pas été concluantes.

## ABSTRACT

**Evaluating the Confidentiality of Synthetic Clinical Texts Generated by Language Models**

Large Language Models (LLMs) can be used to produce synthetic documents that mimic real documents when these are not available due to confidentiality or copyright restrictions. Herein, we investigate potential privacy breaches in automatically generated documents. We use synthetic texts generated from a pre-trained model fine-tuned on French clinical cases to evaluate potential privacy breaches according to three directions : (1) similarity between real, training corpus and synthetic corpus (2) strong correlations between clinical features in training and synthetic corpus and (3) Membership Inference Attack (MIA) using a fined tuned model on the synthetic corpus. We identify clinical feature associations that suggest strategies for filtering training corpus that could contribute to privacy preservation. Membership attacks were not conclusive.

MOTS-CLÉS : Confidentialité, Textes cliniques synthétiques, LLM.

KEYWORDS: Confidentiality, Synthetic Clinical Texts, LLMs.

ARTICLE : **Accepté à** AIME 2025 : https://hal.science/hal-05046326.