Document de travail — Preprint

LaPlume : un pipeline hybride de pseudonymisation des récits sociaux et médico-sociaux français

Conception, architecture et évaluation empirique

Patrick Danto

ConfidensIA SAS

patrick.danto@confidensia.fr

Juin 2026

Statut du document

Document de travail. L'arbitrage inter-annotateur est clos : les trois annotateurs ont rendu leurs décisions et les scores présentés sont définitifs. La taxonomie et l'architecture du pipeline sont stabilisées.

Abstract (English)

We describe **LaPlume**, a hybrid NER pseudonymisation pipeline for the French social and *medico-social* sector (ESSMS), combining a distilled CamemBERT model (*titibongbong* v2, 11L→9L, 172 MB FP16), a rule engine, sectoral gazetteers, and a 101-category taxonomy with a four-level criticality hierarchy. LaPlume specifically addresses the coverage gap of generic pipelines on *indirectly identifying* entities — named institutions, sector-specific pathologies, and administrative identifiers — which are prevalent in social work reports but systematically missed by general-purpose tools. Human annotation on 10 complex real reports yields a criticality-weighted F1 of 96.7% (all categories, consolidated by expert arbitration of 313 disagreements), with CRIT at 98.4% and ELEV at 97.3%. An exploratory benchmark against OpenAI Privacy Filter (OPF) documents: Latence 3–11× lower for LaPlume on CPU, 87–89 % span coverage agreement on direct PII, and a gap in sector-specific entity coverage that generic models cannot address without specialisation. The distilled NER model (*titibongbong* v2) and the distillation corpus are publicly available on Hugging Face under DOI 10.5281/zenodo.17689395. The full pipeline (rules, gazetteers, resolver) remains proprietary.

Keywords: pseudonymisation, NER, social work, medico-social sector, CamemBERT, GDPR, indirectly identifying entities, ESSMS.

Résumé

Nous décrivons **LaPlume**, un pipeline de pseudonymisation NER hybride pour le secteur social et médico-social français (ESSMS), combinant un modèle CamemBERT distillé, un moteur de règles, des gazetteers sectoriels et une taxonomie à 101 catégories avec hiérarchie de criticité. LaPlume traite spécifiquement le défaut de couverture des pipelines généralistes sur les entités *indirectement identifiantes* — institutions nommées, pathologies sectorielles, identifiants administratifs — fréquentes dans les rapports sociaux mais systématiquement manquées par les outils généralistes. L'évaluation humaine sur 10 rapports réels complexes donne un F1 pondéré par criticité de 96,7 % (consolidé par arbitrage de 313 désaccords), avec CRIT à 98,4 % et ELEV à 97,3 %. Une étude comparative exploratoire avec OpenAI Privacy Filter (OPF) documente une latence 3–11× plus faible pour LaPlume en CPU, un accord à 87–89 % sur les entités directes communes, et une absence de couverture sectorielle coté OPF que seul un pipeline spécialisé peut combler. Le modèle NER distillé (*titibongbong* v2) et le corpus de distillation sont disponibles sur Hugging Face sous DOI 10.5281/zenodo.17689395. Le pipeline complet (règles, gazetteers, resolver) demeure propriétaire.

Mots-clés : pseudonymisation, NER, travail social, ESSMS, CamemBERT, RGPD, entités indirectement identifiantes.

Contents

1 Introduction **3**

2 Travaux existants **3**

2.1 Dé-identification clinique 3
2.2 NER médical en langue française 4
2.3 Anonymisation et pseudonymisation 4
2.4 Travaux antérieurs sur LaPlume 5
2.5 Limites des approches généralistes pour les ESSMS 5

3 Architecture de LaPlume **5**

3.1 Vue d'ensemble 5
3.2 Le modèle CamemBERT distillé 6
3.3 Taxonomie à 101 catégories 6
3.4 Hiérarchie de criticité 7
3.5 Moteur de règles et gazetteers 8
3.6 Post-traitement adresses 8
3.7 Comportement en production 8

4 Corpus et protocole d'évaluation **9**

4.1 Deux régimes d'évaluation 9
4.2 Régime 1 : Annotation humaine 9
4.3 Régime 2 : Audit LLM sur corpus de distillation 10

5 Résultats **10**

5.1 Scores F1 par annotateur 10
5.2 Décomposition par niveau de criticité 11
5.3 Accord inter-annotateur 11
5.4 Analyse des erreurs 12

6 Discussion **12**

6.1 Forces 12
6.2 Limites 13

7 Étude comparative exploratoire : LaPlume vs OpenAI Privacy Filter **13**

7.1 Contexte et objectif 13
7.2 Protocole 13
7.3 Résultats 14
7.4 Analyse comparative 15
7.5 Conclusion de l'analyse comparative 15

8 Conclusion **16**

Disponibilité **16**

Références **16**

1. Introduction

Les établissements et services sociaux et médico-sociaux (ESSMS) français produisent quotidiennement un volume considérable de documents narratifs : rapports d'évaluation sociale, notes de suivi, lettres de liaison institutionnelle, comptes rendus de réunion, rapports de tutelle. Ces textes sont au cœur des usages émergents de l'intelligence artificielle dans le secteur : systèmes RAG pour la mémoire organisationnelle, assistants à la formalisation de l'écrit professionnel, outils d'analyse des pratiques, préparation aux évaluations HAS.

Or, pour que ces usages soient légalement et éthiquement recevables, les données personnelles des personnes accompagnées doivent être traitées avec les garanties requises par le RGPD (article 89, considérant 26) et par le référentiel sectoriel de la CNIL. La pseudonymisation est l'instrument central de cette protection : elle permet de traiter des documents sans exposî les informations identifiantes, sous réserve que le risque résiduel reste systématiquement évalué.

Le problème est que les outils de pseudonymisation existants sont conçus pour des corpus généralistes ou médicaux anglophones. Ils ne couvrent pas les entités spécifiques aux rapports sociaux français : noms d'établissements ESSMS, désignations de mesures de protection judiciaire, pathologies du champ psycho-social, identifiants administratifs sectoriels (numéro de dossier MDPH, identifiant SIAO, numéro RSA). Ces entités sont pourtant *indirectement identifiantes* : leur combinaison peut situer une personne dans un réseau local de professionnels, même en l'absence de tout nom propre.

LaPlume a été conçu pour combler ce déficit. Le présent article décrit son architecture, sa taxonomie, et ses premiers résultats d'évaluation. Il constitue la contribution technique fondatrice d'un programme de recherche plus large sur le risque narratif de réidentification dans les rapports ESSMS [15].

2. Travaux existants

2.1. Dé-identification clinique

La dé-identification de textes cliniques a été l'objet de nombreux travaux, principalement sur des corpus en langue anglaise. Les challenges i2b2 (2006, 2014) ont produit des systèmes à haute performance sur les PHI (*Protected Health Information*) : noms, dates, adresses, numéros de sécurité sociale [1, 2]. Ces systèmes atteignent des F1 supérieurs à 97 % sur les identifiants directs, mais ne couvrent pas les entités institutionnelles et quasi-identificatoires spécifiques à un secteur.

Ford et al. [3] soulignent que, même sur des textes cliniques anglophones bien couverts par la dé-identification de surface, le risque résiduel de réidentification demeure non nul dès que des entités indirectes (diagnostics rares, parcours institutionnel) sont présentes.

Pour le français, El Azzouzi et al. [11] présentent une approche de dé-identification des dossiers médicaux électroniques par supervision distante et modéles profonds (Bi-LSTM + CRF, Flair + FastText) atteignant un F1 de 96,96 % sur un corpus d'EHR de l'AP-HP. Leur architecture hybride — combinant apprentissage automatique et règles — préfigure l'approche adoptée par LaPlume, mais sur un corpus médical généraliste et une taxonomie limitée aux PHI standards, sans couverture des entités spécifiques au secteur social. Ce constat motive directement la conception de LaPlume pour un contexte sectoriel et linguistique spécifique.

Lim et Kim [4] présentent Anonpsy, un système de dé-identification des récits psychiatriques par réécriture guidée par graphe. Leur travail illustre la limite des approches de masquage PHI : la structure clinique du récit reste identifiable après masquage, ce qui constitue précisément la problématique que LaPlume traite en amont via une taxonomie étendue.

2.2. NER médical en langue française

Plusieurs travaux portent sur la reconnaissance d'entités dans des documents médicaux français (CAS, QUAERO, CLEF eHealth). Ces systèmes ciblent principalement les entités cliniques standardisées (diagnos-tics, médicaments, actes) et restent peu adaptés aux entités administratives et institutionnelles propres au secteur social.

CamemBERT [5] et ses dérivés spécialisés constituent la colonne vertébrale des approches NER françophones actuelles. La distillation (knowledge distillation) a permis de réduire significativement la taille des modèles tout en maintenant des performances comparables sur les tâches cibles [6].

2.3. Anonymisation et pseudonymisation

Sweeney [7] a montré dès 1997 que la suppression des identifiants directs ne garantit pas la confidentialité : les quasi-identifiants (genre, date de naissance, code postal) permettent de réidentifier des individus par recoupement. Narayanan et Shmatikov [8] ont étendu ce résultat aux données structurées sparses, et Rocher et al. [9] ont montré qu'en utilisant seulement 15 attributs démographiques, 99,98 % des Américains sont réidentifiables dans un jeu de données anonymisé.

Ces travaux ont été produits sur des données structurées. La pseudonymisation des textes narratifs pose des problèmes distincts : l'entité identifiante n'est pas un champ isolé mais un fragment contextuel dont la portée dépend de sa combinaison avec d'autres fragments.

2.4. Travaux antérieurs sur LaPlume

Un premier travail a présenté l'architecture initiale du système sous le nom ConfidensIA [13], évaluée sur un corpus de 330 phrases annotées par un expert unique, avec un F1 global de 86,1 % et 95,6 % sur les entités critiques. Le présent article représente une 'evolution substantielle : (a) 'evaluation sur 10 vrais rapports sociaux complexes avec protocole multi-annotateurs et arbitrage ; (b) mise à disposition publique du modèle NER distillé et du corpus de distillation ; (c) benchmark comparatif avec OpenAI Privacy Filter ; (d) taxonomie portée à 101 catégories avec niveau AUC. Le pipeline complet demeure propriétaire.

2.5. Limites des approches généralistes pour les ESSMS

Les outils commerciaux disponibles (OpenAI Privacy Filter, Microsoft Presidio, spaCy NLP pipelines) présentent trois limites structurelles pour les documents ESSMS :

1. Ils couvrent les entités directement identifiantes (PER, LOC, DATE, PHONE) mais ignorent les entités institutionnelles spécifiques : type d'établissement, service sectoriel, mesure de protection judiciaire.
2. Ils ne disposent pas de gazetteers couvrant la nomenclature ESSMS française (MECS, IME, SESSAD, ESAT, CHRS, CADA, MDPH, ASE…).
3. Ils ne sont pas conçus pour la taxonomie hiérarchisée par criticité que requièrt une pseudonymisation adaptée au RGPD évaluée par le niveau de risque.

3. Architecture de LaPlume

3.1. Vue d'ensemble

LaPlume est un pipeline de pseudonymisation NER hybride conçu spécifiquement pour les rapports sociaux et médico-sociaux français. Il combine quatre composantes complémentaires :

1. **Modèle CamemBERT distillé** (*titibongbong* v2) : détection probabiliste des entités textuelles.
2. **Moteur de règles** (`rules/rules.yaml`) : détection déterministe par patterns regex et contexte local.
3. **Gazetteers sectoriels** : reconnaissance par correspondance exacte sur des listes contrôlées.
4. **Post-traitement adresses** : résolution et fusion des fragments d'adresse en entités complètes.

L'architecture hybride repose sur un principe de complémentarité : le modèle NLP gère la généralisation, les règles traitent les patterns structurés (numéros NIR, codes CAF, téléphones), les gazetteers assurent la couverture des entités nominatives spécifiques au secteur. Le post-traitement adresses évite le fractionnement des adresses complètes en entités partielles.

Le pipeline fonctionne entièrement **en mémoire**, sans persistance de données côté ConfidensIA. Le modèle NER distillé (*titibongbong* v2) et le corpus de distillation sont publiés sur Hugging Face (`huggingface.co/jmdanto`) sous DOI `10.5281/zenodo.17689395` [14]. Le pipeline complet (moteur de règles, gazetteers, resolver) demeure propriétaire : ConfidensIA SAS en assure le développement et la maintenance.

3.2. Le modèle CamemBERT distillé

Le modèle de base est *titibongbong* v2, une distillation du modèle enseignant CamemBERT de 11 couches vers un modèle élève de 9 couches, résultant en un modèle de 172 Mo au format FP16. Le F1 de détection du modèle distillé atteint 86,8 %, à comparer aux 86,4 % du modèle enseignant — gain de 0,4 points pour une réduction de taille de 22 %. Ce modèle a été entraîné sur un corpus de documents médico-sociaux français annotés couvrant les 101 catégories de la taxonomie.

3.3. Taxonomie à 101 catégories

La taxonomie (`TAXONOMIE_PSEUDONYMISATION_v1.1.yaml`) organise les 101 catégories d'entités en neuf domaines :

- **Domaine 1 : Identifiants et coordonnées** (11 catégories, toutes CRIT) — NIR, PHONE, EMAIL, CAF_NUM, ID_DOSSIER, ID_RSA, ID_SIAO, ID_DPH…
- **Domaine 2 : Personnes** (2 catégories) — PER, PER_INITIALE.
- **Domaine 3 : Adresses et localisations** (9 catégories) — `ADDR_FULL`, `ADDR_STREET`, `LOC`, `LOC_CITY`, `LOC_DEPARTEMENT`, `LOC_POI`…
- **Domaine 4 : Organismes publics** (21 catégories) — ORG_CAF, ORG_MDPH, ORG_ASE, ORG_PJJ, ORG_SIAO, ORG_JUSTICE…
- **Domaine 5 : Établissements médico-sociaux** (37 catégories) — ETAB_MECS, ETAB_IME, ETAB_ESAT, ETAB_CHRS, ETAB_EHPAD, ETAB_CADA…
- **Domaine 6 : Associations** (15 catégories) — ORG_ASSOC_HANDICAP, ORG_ASSOC_PROTECTION_ENFANCE, ORG_ASSOC_REFUGIES…
- **Domaine 7 : Unités et services** (4 catégories) — `UNITE_NOMMEE`, `SERVICE_NOMME`, `PAVILLON_NOMME`…

- **Domaine 8 : Dates et temps** (3 catégories) — `DATE`, `DATE_NAISSANCE` (CRIT), `TIME`.
- **Domaine 9 : Pathologies** (1 catégorie, ELEV) — PATHOLOGY, couvrant 764 termes du champ psycho-social, psychiatrique, neurodéveloppemental et somatique.

La richesse du Domaine 5 (37 catégories d'établissements ESSMS) constitue la spécificité la plus distinctive de LaPlume par rapport aux outils généralistes, qui ne distinguent généralement qu'une catégorie `ORG` générique.

3.4. Hiérarchie de criticité

Les 101 catégories sont distribuées en quatre niveaux de criticité :

Table 1: Niveaux de criticité, poids, cibles de rappel et exemples

Niveau	**Poids**	**Cible rappel**	**Exemples**
CRIT (4)	Identifiants directs	98–100 %	PER, DATE_NAISSANCE, NIR, ID_DOSSIER, PHONE, EMAIL
ELEV (2)	Quasi-identifiants forts	90–95 %	ADDR_FULL, ADDR_STREET, PATHOLOGY, tout ETAB_* ou ORG_* nommé
MOY (1)	Quasi-identifiants faibles	80–90 %	ORG_* générique, LOC_CITY, LOC_DEPARTEMENT
FAIB (0.5)	Contexte résiduel	70–80 %	Durées relatives, LOC_POI, profils génériques
AUC (0)	Ne doit pas être pseudonymisé	0 %	Titres de fonctions génériques, acronymes institutionnels non identifiants

Ce schéma de pondération reflète les enjeux métier : les entités CRIT et ELEV sont déterminantes pour la protection effective des personnes ; FAIB est structurellement volatile et peu discriminante. Il fonde le score de référence :

$$\mathrm{F1}_{\mathrm{pondr}} = \frac{4\,\mathrm{F1}_{\mathrm{CRIT}} + 2\,\mathrm{F1}_{\mathrm{ELEV}} + 1\,\mathrm{F1}_{\mathrm{MOY}} + 0{,}5\,\mathrm{F1}_{\mathrm{FAIB}}}{7{,}5} \tag{1}$$

3.5. Moteur de règles et gazetteers

Le moteur de règles (`rules/rules.yaml`) traite les patterns structurés que le modèle NLP gère moins bien :

- numéros NIR (format 13 chiffres + clé),
- numéros de téléphone (formats français et internationaux),
- codes CAF et numéros d'allocataire,
- dates de naissance en format explicite,
- numéros RPPS et FINESS.

Les gazetteers sectoriels couvrent :

- les noms d'établissements ESSMS référencés dans la base FINESS (11 000 gestionnaires),
- les communes et codes postaux français,
- un lexique de 764 termes de pathologies du champ psycho-social,
- les sigles et acronymes sectoriels courants.

3.6. Post-traitement adresses

Le mode `equilibre` fusionne les fragments de rue (`ADDR_STREET`) et de localité (`LOC_CITY`) détectés dans une même fenêtre de contexte en une entité `ADDR_FULL` (ELEV). Ce post-traitement évite qu'une adresse complète apparaisse pseudonymisée en parties — une limite connue des systèmes NER appliqués naïvement à des adresses françaises.

3.7. Comportement en production

Les entités détectées sont remplacées par des balises structurées de la forme `<TYPE_N>`, où `TYPE` est la catégorie taxonomique et `N` un index de corréférence. Exemple :

> *Texte brut* : M. Dupont, né le 12/03/1987, suivi au CHRS L'Escale depuis janvier 2023, porte un trouble du spectre autistique avec déficience intellectuelle. Il est suivi à l'Hôpital de Creil et père de deux enfants placés à l'ASE de l'Oise.
>
> *Après LaPlume* : M. `<PER_1>`, né le `<DATE_NAISSANCE_1>`, suivi au `<ETAB_CHRS_1>` `<DATE_1>`, porte un `<PATHOLOGY_1>` avec `<PATHOLOGY_2>`. Il est suivi à l'`<ORG_CH_CHU_1>` et père de deux enfants placés à l'`<ORG_ASE_1>` `<LOC_DEPARTEMENT_1>`.

Le sens, la structure narrative et les relations entre entités sont entièrement préservés ; seuls les identifiants sont occultés.

4. Corpus et protocole d'évaluation

4.1. Deux régimes d'évaluation

L'évaluation repose sur deux régimes complémentaires, qui mesurent des propriétés différentes :

1. **Annotation humaine indépendante** : mesure la généralisation sur des documents réels de difficulté élevée. C'est l'indicateur de référence pour la robustesse métier.
2. **Audit LLM sur corpus de distillation** : mesure la cohérence interne du pipeline sur un large corpus de documents synthétiques. Indicateur de stabilité, non de généralisation.

4.2. Régime 1 : Annotation humaine

Constitution du corpus

Dix vrais rapports ESSMS ont été sélectionnés par un praticien expérimenté du secteur pour leur difficulté représentative : adresses imbriquées, entités multiples dans une même phrase, acronymes sectoriels ambiguës, contextes administratifs denses (CADA, CHRS, contentieux MDPH, mesures tutelles cumulées). La sélection préfère délibérément les cas difficiles aux cas ordinaires — l'objectif est de mesurer la robustesse aux situations limites, pas la performance sur des documents banals.

Protocole d'annotation

Le protocole repose sur un schéma en double aveugle suivi d'un arbitrage. Trois annotateurs indépendants, possédant une connaissance du domaine du travail social, ont travaillé sans coordination avec l'équipe ConfidensIA et sans lien de subordination entre eux. Chacun a reçu le même guide d'annotation décrivant la taxonomie, les niveaux de criticité et les conventions de délimination des spans. Les annotateurs ont été affectés par paires selon un plan croisé (chaque document annoté par deux annotateurs distincts).

Au cours du processus, les annotateurs ont introduit spontanément la notion de *pathologie indirecte* (mention d'un état de santé par allusion plutôt que par diagnostic explicite). Cette catégorie ne faisant pas partie de la taxonomie officielle, elle a été exclue du calcul présenté ici et fait l'objet d'un travail de recherche complémentaire.

Arbitrage. 313 cas de désaccord ont été soumis à un troisième annotateur expert jouant le rôle d'arbitre pour chaque paire. Les arbitres avaient pour mission unique de trancher parmi : *Valider A*, *Valider B*, *Valider rapprochement*, *Fusionner les deux*, *Rejeter les deux*. Les 26 cas initialement suspendus ont été résolus par analyse contextuelle approfondie. Le gold final est reconstruit comme suit : les erreurs explicites

proviennent des 323 lignes validées après arbitrage ; toute entité pipeline non signalée dans l'audit final est considérée correcte.

Sur les 10 rapports, le pipeline détecte 2 018 entités réparties comme suit :

Table 2: Entités détectées par criticité sur les 10 rapports

Criticité	**Entités**
CRIT	150
ELEV	860
MOY	576
FAIB	267
INCONNUE	165
Total	**2 018**

Un troisième annotateur (Sophie) a assuré l'arbitrage des 313 cas de désaccord. Les décisions sont intégrées : le gold reconstruit présenté dans ce document est le référentiel final.

4.3. Régime 2 : Audit LLM sur corpus de distillation

Un audit automatisé a été conduit sur 592 rapports issus du corpus de distillation, avec un LLM auditeur indépendant du LLM générateur. Le pipeline détecte 25 921 entités sur ce corpus. Le F1 pondéré dépasse 99 % — indicateur de cohérence interne, pas de généralisation : le LLM générateur avait accès à la taxonomie.

5. Résultats

5.1. Scores F1 par annotateur

Table 3: Scores F1 consolidés après arbitrage (corpus d'audit humain)

Périmètre	**F1 global**	**$F1_p$**
Toutes catégories (avec FAIB)	92,5 %	96,7 %
Sans FAIB	93,9 %	97,7 %

$F1_p$ désigne le F1 pondéré par criticité selon l'équation (1). Le score consolidé est unique : il résulte de l'arbitrage du troisième annotateur, qui a tranché l'ensemble des 313 cas de désaccord. Le gold reconstruit comprend 2 016 entités (2 018 prédites, 323 erreurs explicites validées après arbitrage).

5.2. Décomposition par niveau de criticité

Table 4: F1 consolidé par niveau de criticité (après arbitrage)

Criticité	**F1**	**Interprétation**
CRIT	98,4 %	Très forte robustesse sur les identifiants directs
ELEV	97,3 %	Très forte robustesse sur personnes et adresses
MOY	95,6 %	Très bon niveau sur établissements et organismes nommés
FAIB	83,5 %	Niveau plus faible, structurellement attendu

Les trois niveaux opérationnels CRIT, ELEV et MOY sont tous supérieurs à 95 %, traduisant un niveau de protection robuste sur les entités à risque réel d'identification. Un pipeline qui rate une entité CRIT laisse un identifiant direct en clair ; une entité ELEV manquée laisse un quasi-identifiant fort susceptible de permettre une réidentification indirecte.

La catégorie FAIB (83,5 %) regroupe des entités de contexte faible (lieux génériques, horodatages, services sans nom propre) dont la délimination est intrinsèquement subjective. Exclure les catégories FAIB du calcul est cohérent : elles ne contribuent pas significativement au risque RGPD et les objectifs opérationnels du pipeline sont centrés sur les entités à risque d'identification directe.

5.3. Accord inter-annotateur

Table 5: Métriques d'accord inter-annotateur

Indicateur	**Valeur**
Paires appariées	113
Left only (A)	144
Right only (B)	105
Couverture d'appariement	0,312
Accord extraction	0,814
Accord complet	0,434
Accord inter-annotateur ajusté	0,852
Accord ajusté sans missing suspects	0,865

La faible couverture d'appariement (0,312) reflète des différences de segmentation : les deux annotateurs délimitent parfois différemment les frontières d'une même en-

tité. L'accord ajusté de 0,852 est défendable pour une tâche de cette complexité [10]. L'arbitrage du troisième annotateur a tranché les cas litigieux et établi la référence gold définitive (voir section 5).

5.4. Analyse des erreurs

L'analyse qualitative des erreurs révèle trois catégories principales :

1. **Entités multi-tokens à frontières ambigües** : noms d'établissements composites, associations portant un acronyme et un nom développé dans la même phrase.
2. **Institutions hors gazetteers** : services locaux non référencés dans la base FINESS ou nommés de façon non standardisée ("Cellule d'Accueil et de Réinsertion de Périgu'eux", "Commission de coordination MDPH 45").
3. **Pathologies rares ou formulées atypiquement** : termes non couverts par le gazetteer de 764 entrées, formulations cliniques informelles.

Ces catégories définissent les axes d'amélioration prioritaires : enrichissement des gazetteers par extraction FINESS systématique, extension du lexique pathologies vers le champ psychiatrique et neurodéveloppemental.

6. Discussion

6.1. Forces

Couverture métier. La taxonomie à 101 catégories, et notamment ses 37 catégories d'établissements ESSMS et ses 15 catégories d'associations, constitue une couverture sans équivalent dans les outils disponibles pour le français. Elle permet de pseudonymiser des documents où l'établissement, le service et la mesure sont plus identifiants que le nom de la personne.

Traitement des quasi-identifiants. La hiérarchie de criticité distingue explicitement les entités selon leur pouvoir identificatoire. Cette distinction est absente de la plupart des pipelines NER, qui traitent toutes les entités de façon homogène.

Architecture hybride. La combinaison modèle NLP + règles + gazetteers produit une complémentarité : le modèle gère la généralisation contextuelle, les règles traitent les formats structurés, les gazetteers assurent la couverture sectorielle. Aucune composante seule n'atteindrait les mêmes performances.

Exécution locale. Le pipeline fonctionne entièrement en mémoire sans persistance de données. Il peut être déployé sur infrastructure locale, ce qui est compatible avec les exigences RGPD des ESSMS traitant des données de santé.

6.2. Limites

Taille du corpus d'évaluation. Dix rapports constituent un corpus d'évaluation minimal. Les scores présentés sont définitifs (arbitrage clos, gold reconstruit stabilisé). L'établissement d'une référence gold sur un corpus élargi reste la priorité principale.

Généralisation sectorielle. Le corpus d'évaluation représente une sélection de documents difficiles ; les performances sur des documents ordinaires pourraient être différentes. Par ailleurs, certains sous-secteurs ESSMS (petite enfance, protection maternelle) sont sous-représentés.

Drift du vocabulaire sectoriel. Les institutions, services et pathologies évoluent. Les gazetteers requièrent une mise à jour périodique pour maintenir leur couverture. Cette maintenance est un coût récurrent.

Labels génériques résiduels. Dans 41 % des documents du corpus élargi, les entités institutionnelles sont détectées avec les labels génériques `ORG` ou `ETAB` au lieu de leur sous-catégorie fine. Ce phénomène reflète les limites de couverture des gazetteers et constitue un axe d'amélioration prioritaire pour la version v3.

7. Étude comparative exploratoire : LaPlume vs OpenAI Privacy Filter

7.1. Contexte et objectif

OpenAI Privacy Filter (OPF) [12] est un modèle de classification de tokens publié sur Hugging Face, basé sur une architecture bidirectionnelle préentraînée de façon autorégressive (apparenteé à gpt-oss), fine-tunée comme classificateur de tokens sur 8 catégories PII standard : `private_person`, `private_phone`, `private_email`, `private_address`, `private_date`, `account_number`, `private_url`, `secret`. Il est dité avec 1,5B paramètres totaux pour environ 50M paramètres actifs, une fenêtre de contexte de 128 000 tokens, et une licence Apache 2.0.

Nous présentons ici une **étude comparative exploratoire** sur 3 rapports réels du corpus d'annotation, visant non á établir un classement absolu mais à documenter les différences d'approche entre un pipeline hybride spécialisé (LaPlume) et un transformer PII pur (OPF).

7.2. Protocole

Trois rapports courts mais variés ont été sélectionnés du corpus d'annotation :

- `courrier_ssiad_interpellation_cms_tulle` : nombreux noms, dates, téléphone, email, adresse complète ;

- `doc_polyvalence_aide_financiere_001_provence` : contexte administratif, numéros, adresses, allocations ;
- `doc_cada_001_brest` : rapport narratif avec institutions, famille, adresses, contexte d'asile.

La comparaison directe porte sur une **taxonomie commune projetée** : seules les catégories PII standard couvertes par les deux systèmes sont comparées (`PER`/`PER_INITIALE`, `PHONE`, `EMAIL`, `ADDR_*`, `DATE`, identifiants numériques). L'ensemble des entités spécifiques ESSMS de LaPlume (37 catégories ETAB_*, 21 ORG_*, PATHOLOGY, mesures…) est hors périmètre de cette comparaison.

L'exécution est locale en CPU (MacBook Pro M3 Max) avec un warmup par pipeline et une passe mesurée par document.

7.3. Résultats

Latence et empreinte mémoire

Table 6: Latence de traitement par document (CPU local)

Document	LaPlume	OPF	Ratio
Courrier SSIAD	2 284 ms	7 473 ms	3,3×
Polyvalence aide financière	1 966 ms	19 775 ms	10,1×
Rapport CADA	2 062 ms	23 074 ms	11,2×

LaPlume reste stable autour de 2 secondes quelle que soit la densité du document. OPF est significativement plus lent sur les rapports narratifs longs, avec un rapport pouvant dépasser 11×. Cet écart reflète une différence architecturale : le pipeline hybride LaPlume ne charge qu'un modèle 172 Mo en mémoire vive, quand OPF mobilise une empreinte de 1 665 Mo après warmup (RSS observé).

Recouvrement sur la taxonomie commune

Table 7: Spans rapprochés sur la taxonomie commune projetée

Document	Spans LaPlume	Spans OPF	Rapprochés
Courrier SSIAD	48	53	42 (87,5 %)
Polyvalence aide financière	63	66	56 (88,9 %)
Rapport CADA	74	80	9 (12,2 %)

La convergence est forte sur les deux premiers documents (87–89 % de spans rapprochés) : les deux systèmes détectent de façon cohérente les PII directs standards.

La divergence radicale sur le rapport CADA (12,2 %) s'explique principalement par un phénomène technique : OPF produit un avertissement *Token indices sequence length... 551 > 512* lors du traitement local, suggérant que l'intégration Hugging Face testée n'exploite pas de façon transparente la fenêtre de contexte long de 128 000 tokens annoncée dans la fiche du modèle.

7.4. Analyse comparative

Approche hybride vs transformer pur. OPF est un transformer pur de token classification ; LaPlume combine NER distillé, règles et gazetteers. L'intérêt de l'architecture hybride apparaît clairement sur les entités structurées : les règles regex de LaPlume traitent les formats NIR, CAF et téléphone de façon déterministe, sans risque d'omission lié à la distribution des données d'entraînement. OPF produit quelques faux positifs lexicaux typiques (le fragment `la` classé comme entité, un préfixe de téléphone `05` classé en `account_number`).

Densité et spécificité des catégories. Sur les catégories communes, les deux systèmes sont comparables. Mais LaPlume produit au total — sur les mêmes documents — un volume d'entités sensiblement supérieur grâce à ses 37 catégories d'établissements ESSMS, ses 21 organismes publics et son gazetteer de pathologies. Ces entités sont absentes de la taxonomie OPF et constituent la valeur ajoutée spécifique du pipeline pour le secteur médico-social français.

Interprétabilité et contrôle. L'architecture hybride de LaPlume offre une interprétabilité directe : les règles et gazetteers sont lisibles et modifiables, ce qui permet un audit humain et un enrichissement incrémental. OPF est une boîte noire de classification ; ses échecs ne sont pas facilement diagnosticables sans accès à ses données d'entraînement.

Déploiement local. Les deux systèmes sont compatibles avec un déploiement local : LaPlume à 172 Mo FP16, OPF à 50M paramètres actifs (avec une empreinte plus élevée en pratique). Sur une machine avec budget mémoire de 8 Go simulé (RSS observé), les deux systèmes restent dans les limites compatibles avec une infrastructure ESSMS ordinaire.

7.5. Conclusion de l'analyse comparative

Ce benchmark exploratoire n'a pas vocation à établir une hiérarchie générale. Il illustre un choix d'architecture : pipeline hybride spécialisé et interprétable vs transformer pur généraliste. Pour le secteur médico-social français, le choix en faveur de LaPlume repose sur : (a) la couverture des entités indirectement identifiantes sectorielle ; (b) la stabilité de latence sur CPU ; (c) l'interprétabilité via règles et gazetteers ; (d) l'absence de dépendance à une taxonomie PII standard généraliste. OPF constitue une référence

pertinente pour la détection des PHI directs universels, sans répondre à la spécificité métier.

8. Conclusion

LaPlume apporte une infrastructure de pseudonymisation conçue pour les spécificités du secteur ESSMS : couverture des entités indirectement identifiantes, taxonomie sectorielle à 101 catégories, hiérarchie de criticité alignée sur les enjeux de protection des personnes, architecture hybride déployable localement.

Les résultats d'évaluation sur 10 rapports réels complexes (F1 pondéré 96,7 %, CRIT 98,4 %, ELEV 97,3 %, gold reconstruit de 2 016 entités, 323 erreurs arbitrées) confirment la robustesse du pipeline sur les niveaux critiques CRIT et ELEV. L'arbitrage du troisième annotateur a établi la référence gold définitive ; les améliorations prioritaires (enrichissement des gazetteers, extension du lexique pathologies) sont dès lors clairement orientées.

LaPlume constitue une infrastructure de pseudonymisation destinée à rendre possibles des usages de mémoire organisationnelle, d'analyse documentaire et d'assistance par IA dans le secteur ESSMS.

Disponibilité

Modèle NER distillé (*titibongbong* v2, 172 Mo FP16) et corpus de distillation : https://huggingface.co/jmdanto.
DOI : 10.5281/zenodo.17689395.
Le pipeline complet (moteur de règles, gazetteers, resolver d'adresses) demeure propriétaire. Pour toute demande de collaboration ou d'accès acheminale : patrick.danto@confidensia.fr.

Références

References

[1] Stubbs, A. & Uzuner, Ö. (2015). Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *Journal of Biomedical Informatics*, 58, S20–S29.

[2] Dernoncourt, F., Lee, J. Y., Uzuner, O. & Szolovits, P. (2017). De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3), 596–606. https://doi.org/10.1093/jamia/ocw156

[3] Ford, E. et al. (2025). What is the patient re-identification risk from using de-identified clinical free text data for health research? *AI and Ethics*. https://doi.org/10.1007/s43681-025-00681-0

[4] Lim, K. H. & Kim, B.-H. (2026). Anonpsy: A graph-based framework for structure-preserving de-identification of psychiatric narratives. *arXiv* : 2601.13503.

[5] Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D. & Sagot, B. (2020). CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the ACL*, pp. 7203–7219. https://doi.org/10.18653/v1/2020.acl-main.645

[6] Sanh, V. et al. (2019). DistilBERT, a distilled version of BERT. *NeurIPS 2019 Workshop*.

[7] Sweeney, L. (1997). Weaving technology and policy together to maintain confidentiality. *Journal of Law, Medicine & Ethics*, 25, 98–110.

[8] Narayanan, A. & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. *IEEE Symposium on Security and Privacy*, pp. 111–125.

[9] Rocher, L., Hendrickx, J. M. & de Montjoye, Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10, 3069.

[10] Artstein, R. & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4), 555–596.

[11] El Azzouzi, M., Coatrieux, G., Bellafqira, R., Delamarre, D., Riou, C., Oubenali, N., Cabon, S., Cuggia, M. & Bouzillé, G. (2024). Automatic de-identification of French electronic health records: a cost-effective approach exploiting distant supervision and deep learning models. *BMC Medical Informatics and Decision Making*, 24, 54. https://doi.org/10.1186/s12911-024-02422-5

[12] OpenAI (2025). OpenAI Privacy Filter. Modèle de classification de tokens pour la détection de PII, 1,5B paramètres totaux / 50M paramètres actifs, fenêtre de contexte 128 000 tokens, licence Apache 2.0. https://huggingface.co/openai/privacy-filter

[13] Danto, P. (2025). ConfidensIA : un système hybride de pseudonymisation fine-grained pour les documents médico-sociaux français. Preprint, novembre 2025. Zenodo. DOI : 10.5281/zenodo.17689395

[14] Danto, P. (2026). LaPlume v2 : modèle de pseudonymisation NER pour le secteur médico-social français. Zenodo. DOI : 10.5281/zenodo.17689395

[15] Danto, P. (2026). Mémoire organisationnelle, récits professionnels et risque de ré-attribution narrative dans les ESSMS. Document de travail, ConfidensIA SAS.