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

OncoBERT : un modèle de langue clinique français dédié à l'oncologie pour la structuration automatique des données

Quentin Filori[1, 3]
(1) Centre Léon Bérard, Lyon, France
(3) Université Claude Bernard Lyon 1, Lyon, France
quentin.filori@lyon.unicancer.fr

RÉSUMÉ

Nous avons développé et validé un modèle de langage français spécialisé, entraîné sur des narratifs cliniques en oncologie afin d'améliorer la structuration automatique des biomarqueurs et des événements indésirables. Le modèle, OncoBERT, résulte d'un pré-entraînement continu de CamemBERT sur 2,7 millions de comptes rendus médicaux dé-identifiés issus d'un Centre de lutte contre le cancer entre 2000 et 2023. Il a été évalué sur deux tâches : la classification des biomarqueurs (OncoBERT-ANA) et l'extraction de toxicités (OncoBERT-TOX). OncoBERT obtient des scores F1 de 0,96 et 0,92, dépassant DrBERT et CamemBERT-bio de 1 à 5 points. Une validation externe confirme sa généralisabilité avec un F1 de 0,93. Ces résultats montrent qu'OncoBERT constitue une avancée importante dans l'adaptation des modèles biomédicaux aux particularités linguistiques et cliniques de l'oncologie française.

ABSTRACT

OncoBERT : A French Clinical Language Model for Oncology and Data Structuring

We developed and validated a French domain-specific language model trained on oncology clinical narratives to improve the automatic structuring of biomarkers and adverse events. Building on CamemBERT, we performed continued pre-training using 2.7 million de-identified medical reports collected between 2000 and 2023. The resulting model, OncoBERT, was evaluated on biomarker classification (OncoBERT-ANA) and toxicity extraction (OncoBERT-TOX). It achieved F1-scores of 0.96 for ANA and 0.92 for TOX, consistently outperforming DrBERT and CamemBERT-bio by 1 to 5 points. External validation further confirmed its generalizability, with an F1-score of 0.93 on an independent dataset. These findings indicate that OncoBERT constitutes a significant step toward adapting biomedical language models to the specific linguistic and clinical characteristics of French oncology practice.

MOTS-CLÉS : Traitement automatique du langage naturel, oncologie, modèle de langage, pré-entraînement continu, extraction d'information clinique.

KEYWORDS: Natural language processing, oncology, language model, continual pretraining, clinical information extraction.

1 Introduction

Le volume croissant de textes cliniques narratifs dans les dossiers médicaux électroniques (DME) représente une opportunité majeure pour l'utilisation secondaire des données en oncologie. Les

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 41–51.

rapports générés par les anatomo-pathologistes, les radiologues et les cliniciens contiennent des informations essentielles sur la biologie tumorale, les traitements et les événements indésirables. Cependant, leur hétérogénéité intrinsèque rend l'analyse à grande échelle particulièrement complexe. Actuellement, l'abstraction manuelle des données est une tâche chronophage, coûteuse et sujette à une variabilité inter-observateur importante, ce qui limite la disponibilité de données structurées de haute qualité pour la recherche clinique, l'épidémiologie et les systèmes d'aide à la décision.

Les avancées récentes dans le domaine du traitement automatique du langage naturel (TALN) ont transformé notre capacité à extraire et structurer des informations à partir de textes non structurés. Les architectures basées sur les transformeurs, et plus particulièrement le modèle BERT *(Bidirectional Encoder Representations from Transformers)* (Devlin *et al.*, 2019), ont démontré une capacité remarquable de compréhension contextuelle à travers de multiples langues et domaines, y compris la santé (Labrak *et al.*, 2023). Cependant, la grande majorité des modèles de langage pré-entraînés ont été développés initialement pour la langue anglaise, en utilisant des corpus issus du web ou de la littérature biomédicale ouverte (PubMed). Par conséquent, ces modèles sont souvent mal adaptés aux spécificités linguistiques, syntaxiques et stylistiques de la documentation clinique française.

En France, plusieurs initiatives ont vu le jour pour développer des modèles fondamentaux. On peut citer des modèles généralistes comme CamemBERT (Martin *et al.*, 2020) et FlauBERT (Le *et al.*, 2020), entraînés sur des données ouvertes diversifiées (Wikipédia, presse), ainsi que des modèles spécialisés comme DrBERT (Labrak *et al.*, 2023) et CamemBERT-bio (Touchent *et al.*, 2023), qui ont exploité des corpus biomédicaux. Bien que ces modèles constituent des jalons importants, leurs corpus d'entraînement restent majoritairement composés de textes scientifiques, réglementaires ou encyclopédiques plutôt que de narratifs cliniques réels rédigés par des professionnels de santé en situation de soin. En conséquence, ces modèles peinent à capturer la terminologie hospitalière spécifique, les abréviations complexes et la variabilité linguistique propre à la pratique clinique quotidienne.

Dans le domaine de l'oncologie, cette limitation est particulièrement critique. La prise en charge du cancer repose lourdement sur l'interprétation de documents en texte libre, qui incluent des descriptions complexes de résultats histologiques, de statuts de biomarqueurs et de toxicités liées aux traitements. Aujourd'hui, des centaines de biomarqueurs sont utilisés en routine pour guider les décisions thérapeutiques, évaluer le pronostic et identifier les patients éligibles aux essais cliniques (Scott *et al.*, 2023). Par exemple, l'amplification de HER2 et l'expression des récepteurs hormonaux sont des biomarqueurs clés dans le cancer du sein, mais ils apparaissent sous de multiples variations textuelles (ex : « HER2 3+ », « Her-2 positif », « surexpression de HER2 »). De même, les rapports de toxicité suivent des formats idiosyncratiques où les symptômes, les grades et les traitements sont mentionnés de manière non standardisée, rendant l'extraction automatisée extrêmement difficile pour des modèles non spécialisés.

Pour relever ces défis, le développement de modèles de langage entraînés directement sur des textes cliniques d'oncologie est apparu comme une nécessité. Des travaux antérieurs en anglais (BioBERT (Lee *et al.*, 2020), ClinicalBERT (Huang *et al.*, 2019)) et en allemand (FS-BERT (Bressem *et al.*, 2020)) ont démontré que le pré-entraînement continu sur des données spécifiques au domaine améliore significativement les performances sur les tâches cliniques. Cependant, de telles approches restent peu explorées dans le contexte clinique français, malgré la disponibilité de vastes entrepôts de données de santé dé-identifiées et un intérêt institutionnel croissant pour la structuration des données assistée par l'intelligence artificielle.

C'est dans ce contexte que nous avons développé OncoBERT, un modèle de langage français spé-

cifiquement adapté à l'oncologie. OncoBERT repose sur l'architecture CamemBERT et a bénéficié d'un pré-entraînement continu (CPT) sur un corpus massif de 2,7 millions de rapports d'oncologie dé-identifiés, couvrant la période 2000–2023. L'objectif principal de ce travail est d'évaluer si cette adaptation spécifique permet d'obtenir des performances supérieures aux modèles biomédicaux français existants sur des tâches cliniques réelles.

Nous avons focalisé notre évaluation sur deux applications majeures : (1) l'extraction de biomarqueurs à partir de rapports d'histopathologie (OncoBERT-ANA) et (2) l'extraction de toxicités à partir de notes de consultation (OncoBERT-TOX), en suivant la nomenclature CTCAE V5. À travers ces tâches, nous visons à démontrer la faisabilité, la valeur clinique et l'efficacité computationnelle de notre approche, ouvrant la voie à un déploiement opérationnel au sein des structures hospitalières.

2 Matériels et Méthodes

2.1 Source des Données et Pré-traitement

Cette étude a été réalisée en utilisant les dossiers médicaux électroniques (DME) dé-identifiés d'un centre de lutte contre le cancer (CLCC) français de référence. Le corpus initial comprenait une grande diversité de documents narratifs : comptes-rendus d'hospitalisation, rapports de consultation, lettres de sortie et comptes-rendus d'anatomo-pathologie, générés entre 2000 et 2023. Les documents purement administratifs ou non cliniques ont été rigoureusement exclus de l'analyse.

La protection de la vie privée des patients a été une priorité absolue. Tous les rapports ont été traités par un pipeline de dé-identification interne robuste. Ce système utilise des expressions régulières et des règles métier pour supprimer les identifiants personnels (noms, prénoms, dates de naissance précises, adresses, numéros de téléphone). Ce processus a été validé en interne pour garantir que le contenu médical essentiel était préservé tout en assurant l'anonymat, conformément aux standards de la littérature (Vakili & Dalianis, 2022). Seuls les patients ne s'étant pas opposés à l'utilisation secondaire de leurs données à des fins de recherche ont été inclus dans l'étude.

Le corpus final de pré-entraînement se compose d'environ 2,7 millions de documents, répartis entre 1,4 million de documents internes au Centre et 1,3 million de documents externes intégrés au dossier patient. Ces derniers, souvent issus de numérisations (OCR), ont fait l'objet d'un filtrage qualité strict. Nous avons développé une métrique de confiance OCR basée sur la proportion de caractères non reconnus pour exclure les documents trop bruités. Au total, le corpus représente un volume de 7 Go de texte brut, ce qui est comparable ou supérieur aux corpus utilisés pour DrBERT (7 Go) (Labrak *et al.*, 2023) ou CamemBERT-bio (2,7 Go) (Touchent *et al.*, 2023).

2.2 Architecture et Entraînement du Modèle

OncoBERT est construit sur la base du modèle CamemBERT-base (Martin *et al.*, 2020), qui implémente l'architecture RoBERTa (Liu *et al.*, 2019). Cette configuration comprend 12 couches de transformeurs, 768 dimensions cachées, 12 têtes d'attention, pour un total de 110 millions de paramètres.

La stratégie adoptée est celle du **Pré-entraînement Continu** (CPT) (Gu *et al.*, 2021). Contrairement

à un entraînement à partir de zéro (*from scratch*), le CPT consiste à poursuivre l'entraînement d'un modèle déjà pré-entraîné sur un corpus généraliste en utilisant des données spécifiques à un domaine. Cette méthode permet d'adapter les poids du modèle à la sémantique et à la syntaxe spécialisées sans perdre les capacités linguistiques générales acquises initialement. Il a été démontré que le CPT offre souvent de meilleures performances sur les tâches spécialisées (Douka *et al.*, 2021; Bressem *et al.*, 2023).

L'objectif d'entraînement utilisé est la Modélisation du Langage Masqué (MLM). Dans chaque séquence de texte, 15 % des jetons sont sélectionnés aléatoirement : 80 % sont remplacés par un jeton spécial `[MASK]`, 10 % restent inchangés et 10 % sont remplacés par un jeton aléatoire du vocabulaire. Le modèle doit alors prédire les jetons originaux en minimisant une fonction de perte d'entropie croisée.

L'entraînement a été réalisé sur un unique GPU NVIDIA A40 pendant 50 époques, avec une taille de lot (*batch size*) de 16 et un taux d'apprentissage de 5×10^{-5}. Ce processus a duré environ 170 heures. Il est important de souligner que cette configuration est relativement légère par rapport à des modèles comme DrBERT, qui a nécessité plus de 25 000 heures de calcul sur le supercalculateur Jean Zay. Cette efficacité démontre que l'adaptation spécifique au domaine est accessible avec des ressources computationnelles modérées au sein d'une structure hospitalière.

Le modèle final, OncoBERT, a ainsi intégré les spécificités sémantiques, les abréviations médicales et les structures syntaxiques propres à l'oncologie française, tout en conservant la robustesse de l'architecture CamemBERT.

2.3 Tâches d'Évaluation descendantes

Afin de mesurer l'apport réel du pré-entraînement continu, nous avons évalué OncoBERT sur deux tâches cliniques représentatives des besoins de structuration de données en oncologie. Ces tâches ont été formulées comme des problèmes d'apprentissage supervisé et comparées aux performances de CamemBERT (généraliste), CamemBERT-bio et DrBERT (biomédicaux).

2.3.1 Extraction de Biomarqueurs (OncoBERT-ANA)

Cette tâche consiste à identifier et classifier le statut de biomarqueurs tumoraux (ex : HER2, ALK, P16, ER, PR) à partir des conclusions de rapports d'anatomo-pathologie et de biologie moléculaire (NGS). La méthodologie a suivi plusieurs étapes :

- **Identification** : Les phrases contenant des mentions de biomarqueurs ont été extraites à l'aide d'expressions régulières et d'un lexique exhaustif incluant les variantes orthographiques (ex : « Her-2 », « HER 2 », « Her2 »).
- **Anonymisation des entités** : Pour éviter que le modèle ne se focalise sur des motifs textuels spécifiques à un biomarqueur, chaque mention a été remplacée par un jeton neutre `[BM]`.
- **Classification** : Chaque phrase a été classée selon quatre catégories : Positif, Négatif, Inconnu ou Équivoque. La catégorie « Équivoque » est particulièrement cruciale pour le biomarqueur HER2, où les résultats d'immunohistochimie peuvent être indéterminés (score 2+).

Deux sous-modèles ont été entraînés : un modèle spécifique à la détection de l'ambiguïté HER2 (1200 phrases annotées) et un modèle général de classification des biomarqueurs (11 000 phrases annotées).

2.3.2 Extraction de Toxicités et Relations (OncoBERT-TOX)

Cette tâche vise à identifier les événements indésirables (toxicités) mentionnés dans les notes de consultation et à établir leur lien de causalité avec les traitements administrés. Nous avons sélectionné aléatoirement 500 rapports médicaux contenant au moins deux événements indésirables. L'annotation manuelle, réalisée par des experts, a porté sur les traitements, les symptômes, les dates et les grades de toxicité selon la nomenclature CTCAE v5.0.

La tâche a été décomposée en trois sous-problèmes :

1. **Reconnaissance d'Entités Nommées (NER)** : Identification des noms de médicaments, des symptômes et des dates. Nous avons utilisé le schéma d'étiquetage BIOUL (*Beginning, Intermediate, Last, Unique*) pour gérer les entités composées de plusieurs mots, portant l'espace des étiquettes à 13 tags.
2. **Extraction de Grade** : Réalisée par expressions régulières (ex : « nausée de grade 2 » $\rightarrow$ Grade 2) pour garantir une extraction fiable sans interprétation subjective du texte.
3. **Classification de Relation** : Identification du lien causal entre un symptôme et un traitement. Les paires traitement-symptôme ont été marquées avec des balises `<e1>` et `<e2>` au sein de la phrase et classées en trois catégories : Toxicité, Autre ou Incertain.

2.4 Éthique et Gouvernance

L'ensemble du projet a été mené dans un cadre éthique strict. L'étude a reçu l'approbation du Comité de Gouvernance des Données du Centre Léon Bérard. Elle est conforme au Règlement Général sur la Protection des Données (RGPD) et aux recommandations de la CNIL. L'utilisation des données repose sur le principe de non-opposition des patients, informés préalablement de la possibilité de réutilisation de leurs données de santé à des fins de recherche.

3 Résultats

3.1 Description des Cohortes d'Évaluation

Pour la tâche OncoBERT-ANA, le jeu de données comprenait 11 000 phrases pour la classification générale et un sous-groupe de 1200 phrases pour l'analyse spécifique de HER2.

Pour la tâche OncoBERT-TOX, le corpus d'évaluation était constitué de 500 rapports de consultation annotés, totalisant plus de 163 000 jetons. Ce corpus contenait 2013 mentions de symptômes et 2371 mentions de médicaments. Le jeu de données de relation comprenait 15 847 paires traitement-symptôme, avec un fort déséquilibre (seulement 6 % de relations de type « Toxicité ») La répartition des pathologies par cohortes est illustrée dans les figures 1 et 2.

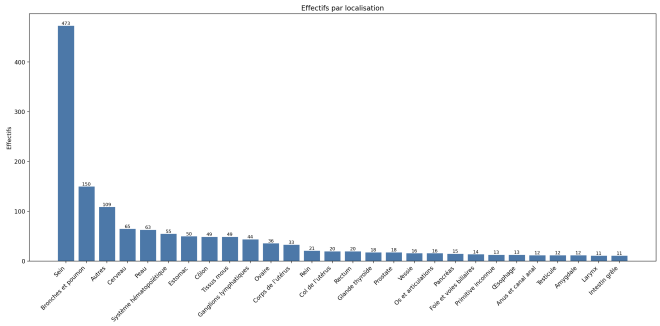

FIGURE 1 – Descriptif cohorte OncoBERT-ANA

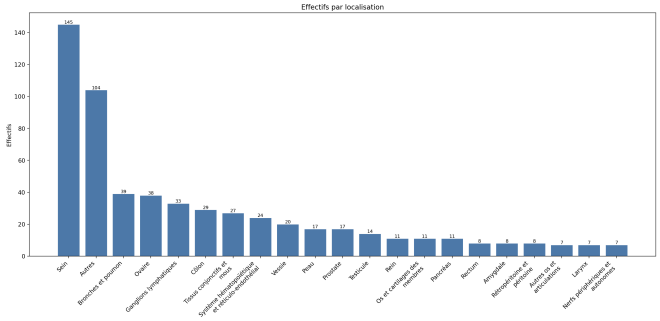

FIGURE 2 – Descriptif cohorte OncoBERT-TOX

3.2 Efficacité du Pré-entraînement

Le pré-entraînement continu a montré une convergence stable. L'inspection qualitative des prédictions sur les jetons masqués a révélé une nette supériorité d'OncoBERT pour prédire des termes techniques et des abréviations oncologiques (ex : « Ki-67 > 30 % », « neutropénie G3 », « HER2 3+ ») par rapport au modèle CamemBERT original. Le coût computationnel réduit (170 h sur un GPU) confirme que cette approche est viable pour une institution hospitalière souhaitant développer ses propres outils souverains.

3.3 Performances sur les Tâches descendantes

3.3.1 Résultats OncoBERT-ANA (Biomarqueurs)

Les performances de classification des biomarqueurs montrent qu'OncoBERT surpasse légèrement mais systématiquement les modèles de référence (Table 1).

L'amélioration est particulièrement notable sur les cas ambigus de HER2, où OncoBERT parvient à mieux distinguer les expressions d'immunohistochimie non concluantes (ex : « 2+ / FISH en attente ») des véritables positifs.

Modèle	Précision	Rappel	Score F1	Exactitude
CamemBERT	0,95	0,95	0,95	0,95
CamemBERT-bio	0,95	0,95	0,95	0,95
DrBERT-7GB	0,95	0,95	0,95	0,95
OncoBERT	0,96	0,95	0,96	0,96

TABLE 1 – Performances de classification des biomarqueurs sur le jeu de test interne.

3.3.2 Résultats OncoBERT-TOX (Toxicités)

Sur la tâche de reconnaissance d'entités nommées (NER), OncoBERT affiche une progression significative, notamment en termes de rappel (table 2).

Modèle	Précision	Rappel	Score F1	Exactitude
CamemBERT	0,94	0,93	0,93	0,93
CamemBERT-bio	0,93	0,94	0,92	0,92
DrBERT-7GB	0,91	0,91	0,92	0,91
OncoBERT	0,96	0,98	0,97	0,97

TABLE 2 – Performances NER sur OncoBERT-TOX.

Pour la classification de relation, OncoBERT a obtenu un score F1 de 0,87, identique à DrBERT mais avec une meilleure précision qualitative. Le modèle a généré moins de faux positifs dans les contextes de polychimiothérapie, où l'attribution d'un symptôme à un médicament spécifique est complexe.

3.4 Validation Externe et Généralisabilité

La robustesse d'OncoBERT a été testée par une validation externe dans deux autres centres de lutte contre le cancer (Table 3).

Centre	F1 Système local (Règles)	F1 OncoBERT-ANA
Institut Curie	0,87	0,94
Centre Oscar Lambret	0,85	0,93

TABLE 3 – Résultats de la validation externe.

À l'Institut Curie, sur un « Gold Standard » de 1000 patients, OncoBERT a atteint un score F1 de 0,94. Au Centre Oscar Lambret, sur une cohorte de 12 612 patients, le modèle a obtenu un score F1 de 0,93. Dans les deux cas, OncoBERT a surpassé les systèmes experts basés sur des expressions régulières utilisés localement. Ces résultats confirment que les connaissances acquises par OncoBERT sur les données du Centre Léon Bérard sont transférables à d'autres institutions, malgré les variations de styles de rédaction entre cliniciens.

4 Discussion

Cette étude démontre que le pré-entraînement continu sur des données d'oncologie réelles améliore significativement les performances des modèles de langage français pour les tâches de TALN clinique. En adaptant le modèle CamemBERT à l'aide de narratifs cliniques spécifiques, OncoBERT a obtenu des résultats supérieurs aux modèles biomédicaux existants, tels que DrBERT et CamemBERT-bio, sur deux tâches critiques : l'extraction de biomarqueurs et la structuration des toxicités. Ces résultats confirment que l'adaptation au domaine est non seulement faisable mais essentielle pour l'informatique clinique dans un contexte non anglophone.

4.1 Interprétation des Résultats Principaux

Le premier enseignement majeur est qu'OncoBERT surpasse les modèles biomédicaux de pointe tout en ayant été entraîné avec des ressources computationnelles nettement inférieures. L'obtention d'un score F1 de $0{,}96$ pour les biomarqueurs et de $0{,}97$ pour la reconnaissance d'entités (NER) souligne que la proximité linguistique entre les données de pré-entraînement et les textes cliniques cibles est un facteur de performance plus déterminant que la taille brute du corpus ou la puissance de calcul engagée.

Deuxièmement, nos résultats indiquent que la nature du corpus de pré-entraînement — documents cliniques réels versus littérature scientifique ouverte — a un impact critique sur la capacité de généralisation. Des modèles comme DrBERT, bien qu'entraînés sur de vastes corpus biomédicaux, capturent la terminologie savante mais peinent face à la syntaxe fragmentée, aux abréviations hospitalières et à la structure narrative des dossiers patients. L'exposition d'OncoBERT à des rapports authentiques lui a permis d'intégrer des signaux contextuels spécifiques (ex : « ER+ », « HER2++ », « toxicité G2 »), omniprésents dans le discours oncologique quotidien.

Enfin, l'efficacité computationnelle d'OncoBERT (170 heures sur un seul GPU) est un argument fort pour la démocratisation de ces technologies. Elle prouve que les centres hospitaliers peuvent développer et maintenir leurs propres modèles spécialisés, garantissant ainsi une meilleure souveraineté sur leurs données et leurs outils d'analyse, sans dépendre de supercalculateurs nationaux ou de solutions propriétaires coûteuses.

4.2 Comparaison avec les Travaux Antérieurs

Nos conclusions rejoignent les observations faites dans d'autres langues, comme BioBERT pour l'anglais (Lee *et al.*, 2020) ou FS-BERT pour l'allemand (Bressem *et al.*, 2020), confirmant que l'adaptation au domaine est une stratégie gagnante. OncoBERT étend cette preuve de concept au domaine de l'oncologie en langue française, un domaine où aucun modèle de cette envergure n'avait été décrit publiquement jusqu'à présent.

Contrairement aux modèles anglais souvent entraînés sur des bases de données multi-institutionnelles (comme MIMIC-III), OncoBERT a été initialement entraîné sur les données d'un seul centre. Pourtant, sa validation externe réussie à l'Institut Curie et au Centre Oscar Lambret suggère que les régularités sémantiques de l'oncologie (terminologie des biomarqueurs, gradation CTCAE, schémas thérapeutiques) sont suffisamment stables à l'échelle nationale pour permettre un transfert d'apprentissage

robuste.

La tâche de classification des biomarqueurs (OncoBERT-ANA) a montré une variabilité inter-modèle relativement faible, ce qui s'explique par la standardisation croissante du langage entourant les récepteurs hormonaux. Cependant, le gain de performance d'OncoBERT réside dans sa capacité à traiter les cas complexes et les notations ambiguës, là où les modèles généralistes échouent par manque de contexte clinique.

À l'inverse, la tâche OncoBERT-TOX est linguistiquement beaucoup plus ardue. Elle nécessite de lier des entités souvent dispersées dans le texte. L'amélioration de 5 points du score F1 pour la NER démontre l'importance de la familiarité avec le phrasé clinique réel (ex : références implicites aux cycles de chimiothérapie comme « FEC100 C2 »).

4.3 Limites de l'Étude

Plusieurs limites doivent être soulignées. Premièrement, bien que la validation externe soit encourageante, le pré-entraînement initial repose sur une institution unique. Des variations régionales ou des différences de systèmes d'information pourraient influencer les performances lors d'un déploiement à plus large échelle. Un futur pré-entraînement multi-institutionnel, potentiellement via l'apprentissage fédéré, permettrait de lever cette limite.

Deuxièmement, le pipeline de dé-identification, bien que rigoureux, peut introduire des artefacts de tokenisation ou supprimer par erreur des termes contextuels utiles. Bien que l'impact semble minime selon la littérature (Vakili & Dalianis, 2022), une étude d'ablation plus formelle permettrait de quantifier précisément cet effet.

Troisièmement, la taille du jeu de données pour la classification des relations de toxicité (500 documents) reste modeste. Le déséquilibre important des classes (seulement 6 % de toxicités réelles) impose une limite aux performances du modèle. La chute de performance observée sur certains jeux de données externes (61 % F1 pour les relations) souligne que le modèle dépend fortement de l'explicitation textuelle du lien de causalité. Or, en pratique clinique, ce lien est souvent implicite et repose sur le raisonnement médical de l'oncologue, ce qui échappe à une analyse purement textuelle.

4.4 Travaux Futurs et Perspectives

Nous prévoyons d'étendre OncoBERT à d'autres tâches de structuration clinique, telles que la détection du statut métastatique, l'extraction des comorbidités ou l'identification des allergies médicamenteuses. Ces informations sont essentielles pour obtenir une vision complète et intégrée du parcours patient.

Un axe important de nos travaux porte sur la possibilité de publier ce modèle en open source. Inspirés par les travaux de Boutet et al. (Boutet *et al.*, 2025), nous explorons actuellement différentes approches de désapprentissage afin de garantir une anonymisation suffisante, nous permettant ainsi de diffuser un modèle entraîné sur de véritables données cliniques tout en respectant les contraintes de confidentialité.

Une autre perspective consiste à moderniser l'architecture d'OncoBERT en développant une nouvelle version fondée sur l'architecture CamemBERTa V2 (Antoun *et al.*, 2024). Les différences dans les

stratégies d'entraînement proposées par cette architecture devraient nous offrir un meilleur contrôle sur la mémorisation potentielle d'informations identifiantes par le modèle.

Enfin, des études d'impact organisationnel seront nécessaires afin de quantifier le temps économisé par les attachés de recherche clinique (ARC) et les médecins lors de la curation des données et de la constitution de cohortes pour les essais cliniques.

5 Conclusion

Cette étude présente OncoBERT, le premier modèle de langage français pré-entraîné de manière continue sur des narratifs cliniques d'oncologie. En exploitant 2,7 millions de rapports médicaux, OncoBERT a réussi à adapter l'architecture CamemBERT aux spécificités sémantiques et syntaxiques de la cancérologie. Le modèle a atteint des performances de pointe sur l'extraction de biomarqueurs et la structuration des toxicités, surpassant les modèles biomédicaux français existants tout en restant économe en ressources de calcul.

Nos résultats confirment que le pré-entraînement continu sur des données cliniques réelles est une stratégie efficace pour combler le fossé entre la recherche en IA et la pratique hospitalière. Au-delà des gains quantitatifs, OncoBERT démontre une capacité réelle à capturer les nuances du langage médical français, ouvrant la voie à une exploitation plus systématique et performante des données textuelles massives contenues dans les dossiers patients.

En conclusion, OncoBERT offre une base solide et souveraine pour la structuration automatique des données en oncologie, facilitant ainsi la recherche clinique, le suivi de la qualité des soins et, à terme, une prise en charge plus personnalisée et efficace des patients atteints de cancer.

Références

ANTOUN W., KULUMBA F., TOUCHENT R., ÉRIC DE LA CLERGERIE, SAGOT B. & SEDDAH D. (2024). Camembert 2.0 : A smarter french language model aged to perfection.

BOUTET A., MAGNANA L., SÉNÉCHAL J. & ZIMMERMANN H. (2025). Towards the anonymization of the language modeling.

BRESSEM K. K., ADAMS L. C., ERXLEBEN C., HAMM B., DEWEY M. & FELDHAUS F. (2020). Highly accurate classification of chest radiographic reports using a deep learning natural language model. *Bioinformatics*.

BRESSEM K. K., PAPAIOANNOU J.-M., GRUNDMANN P., BORCHERT F., ADAMS L. C., LIU L., BUBER F., GEORGENS J., SEEGRÄBER M., KHIL L. *et al.* (2023). MEDBERT.de : A comprehensive German BERT model for the medical domain. arXiv : 2303.08179.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. arXiv : 1810.04805.

DOUKA S., MEEHAN B., KALLIS N., BIKAKIS N. & DALIANIS H. (2021). JuriBERT : A masked-language model adaptation for French legal text. arXiv : 2110.01485.

GU Y., TINN R., CHENG H., LUCAS M., USUYAMA N., LIU X., NAUMANN T., GAO J. & POON H. (2021). Domain-specific language model pretraining for biomedical natural language processing. In *ACM Conference on Health, Inference, and Learning (CHIL)*.

HUANG K., ALTOSAAR J. & RANGANATH R. (2019). ClinicalBERT : Modeling clinical notes and predicting hospital readmission. arXiv : 1904.05342.

LABRAK Y., BAZOGE A., DUFOUR R., DAILLE B., GOURRAUD P.-A., MORIN E. & ROUVIER M. (2023). DrBERT : A robust pre-trained model in French for biomedical and clinical domains. arXiv : 2304.00958.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). FlauBERT : Unsupervised language model pre-training for French. arXiv : 1912.05372.

LEE J., YOON W., KIM S., KIM D., KIM S., SO C. H. & KANG J. (2020). BioBERT : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). RoBERTa : A robustly optimized BERT pretraining approach. arXiv : 1907.11692.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE , SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

SCOTT E. C., BAINES A. C., GONG Y., MOORE R., PAMBER G. S., HERBER S., GOLDBERG K. B., PAZDUR R., SRIDHARA R., IBRAHIM A. *et al.* (2023). Trends in the approval of cancer therapies by the FDA in the twenty-first century. *Nature Reviews Drug Discovery*.

TOUCHENT R., LABRAK Y., BAZOGE A., MORIN E., DUFOUR R. & ROUVIER M. (2023). CamemBERT-bio : Un modèle de langue français savoureux et meilleur pour la santé. HAL : hal-04085419.

VAKILI T. & DALIANIS H. (2022). Downstream task performance of BERT models pre-trained using automatically de-identified clinical data. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.