OncoDEBERTa : adaptation d'un modèle DeBERTa-v3 au domaine oncologique clinique français

Quentin Filori[1,3] Thomas Checchin[2] Hugo Crochet[1] Pierre Heudel[1]
Loïc Verlingue[1] Julien Jacques[2] Adrien Guille[2] Jean-Yves Blay[1]

[1] Centre Léon Bérard, Lyon, France
[2] Université Lumière Lyon 2, Université Claude Bernard Lyon 1, ERIC, 69007, Lyon, France
[3] Université Claude Bernard Lyon 1, France

quentin.filori@lyon.unicancer.fr

RÉSUMÉ

Nous présentons OncoDEBERTa, un modèle de langue français adapté au domaine oncologique en combinant l'architecture DeBERTa-v3 (CamemBERTav2) et une stratégie d'entraînement de type ELECTRA. OncoDeBERTa est le résultat d'un pré-entraînement continu sur le même corpus de 2,7 millions de comptes-rendus oncologiques dé-identifiés que son prédécesseur OncoBERT (CamemBERT + MLM), mais atteint la convergence en une seule itération contre 50, illustrant l'efficience supérieure du cadre ELECTRA en pré-entraînement. Évalué sur deux tâches cliniques en validation croisée stratifiée à 5 plis, OncoDEBERTa établit de nouveaux niveaux de performance : F1-macro de $0{,}88$ en classification binaire du statut métastatique ($+5$ pts vs OncoBERT, $p = 0{,}013$) et F1 de $0{,}81$ sur l'entité rare *Médicament* en reconnaissance d'entités nommées (NER) des toxicités médicamenteuses ($+6$ pts, $p = 0{,}001$). Nos résultats montrent que les gains proviennent conjointement de l'architecture DeBERTa-v3 et de l'adaptation au domaine oncologique ; un pré-entraînement biomédical général (CamemBERT-bio) ne suffit pas à reproduire ces gains.

ABSTRACT

OncoDEBERTa : Domain Adaptation via ELECTRA-style Training for French Clinical Oncology NLP.

We present OncoDEBERTa, a French language model adapted to the oncology domain by combining the DeBERTa-v3 architecture (CamemBERTav2) with an ELECTRA-style pre-training strategy. OncoDEBERTa is trained with continued pre-training on the same corpus of 2.7 million de-identified oncological reports used to train OncoBERT (CamemBERT + MLM), but reaches convergence in a single itération versus 50, illustrating the higher pre-training efficiency of the ELECTRA framework. Evaluated on two clinical tasks under 5-fold stratified cross-validation, OncoDEBERTa achieves state-of-the-art performance : F1-macro of 0.88 for binary metastatic status classification ($+5$ pts vs OncoBERT, $p = 0.013$) and F1 of 0.81 for the rare *Drug* entity in toxicity NER ($+6$ pts, $p = 0.001$). Our results show that gains jointly stem from the DeBERTa-v3 architecture and oncological domain adaptation ; generic biomedical pre-training (CamemBERT-bio) does not reproduce these gains.

MOTS-CLÉS : modèle de langue, adaptation de domaine, ELECTRA, DeBERTa, oncologie, NER, classification, NLP clinique, français.

KEYWORDS: language model, domain adaptation, ELECTRA, DeBERTa, oncology, NER, classification, clinical NLP, French.

1 Introduction

Le volume croissant de textes cliniques narratifs dans les dossiers médicaux électroniques représente une opportunité majeure pour la structuration automatique des données en oncologie. Les comptes-rendus générés par les anatomo-pathologistes, les radiologues et les oncologues contiennent des informations décisives sur la biologie tumorale, l'extension de la maladie et les événements indésirables liés aux traitements. Cependant, leur hétérogénéité linguistique rend l'exploitation automatique à grande échelle difficile, et l'extraction manuelle demeure chronophage et coûteuse.

Les architectures de type transformeur, et en particulier BERT (Devlin *et al.*, 2019), ont démontré leur capacité à traiter des tâches de compréhension textuelle dans de nombreux domaines, y compris la santé (Labrak *et al.*, 2023). Pour le français, des modèles généralistes comme CamemBERT (Martin *et al.*, 2020) et des modèles biomédicaux comme DrBERT (Labrak *et al.*, 2023) ou CamemBERT-bio (Touchent *et al.*, 2023) ont constitué des jalons importants. Néanmoins, ces modèles sont entraînés principalement sur de la littérature scientifique ou réglementaire et peinent à capturer la syntaxe fragmentée, les abréviations et la terminologie des narratifs cliniques réels.

Dans ce contexte, nous nous appuyons sur OncoBERT, un modèle faisant l'objet d'un article en cours de publication, obtenu par pré-entraînement continu de CamemBERT sur 2,7 millions de comptes-rendus oncologiques dé-identifiés et ayant démontré des gains significatifs sur l'extraction de biomarqueurs et la structuration de toxicités. OncoBERT repose sur l'objectif de Modélisation du Langage Masqué (MLM), dans lequel seuls 15 % des tokens d'une séquence contribuent au signal d'apprentissage à chaque pas d'entraînement ; les 85 % restants ne sont pas supervisés.

Des architectures plus récentes offrent une alternative plus efficace. ELECTRA (Clark *et al.*, 2020) introduit un cadre générateur-discriminateur dans lequel le modèle principal (le discriminateur) apprend à distinguer les tokens originaux des tokens remplacés par un petit générateur ; le modèle est ainsi supervisé sur 100 % des tokens à chaque pas d'entraînement. DeBERTa-v3 (He *et al.*, 2021) combine cette stratégie ELECTRA avec un mécanisme d'attention découplant explicitement la représentation de contenu et la représentation de position, pour une meilleure modélisation des relations contextuelles. CamemBERTav2 (Antoun *et al.*, 2024) adapte cette architecture au français.

Nous présentons OncoDEBERTa, qui étend OncoBERT selon deux axes :

1. **Architecture** : remplacement de CamemBERT par CamemBERTav2 (DeBERTa-v3), doté d'une attention découplée et d'une représentation positionnelle relative.
2. **Stratégie de pré-entraînement** : passage du MLM à une approche ELECTRA, plus efficiente en pré-entraînement, ce qui permet d'atteindre la convergence en une seule itération contre 50 pour OncoBERT.

Nous évaluons OncoDEBERTa sur deux tâches cliniques représentatives : la classification binaire du statut métastatique (présence ou absence de métastases) et la NER des toxicités médicamenteuses. Nous comparons OncoDEBERTa à OncoBERT, à deux modèles non spécialisés (CamemBERT-base, CamemBERTav2) et à un modèle spécialisé sur un corpus biomédical libre (CamemBERT-bio), afin d'isoler les apports respectifs de l'architecture et de l'adaptation au domaine.

2 Données

2.1 Corpus de pré-entraînement

OncoBERT et OncoDEBERTa partagent le même corpus de pré-entraînement, constitué d'environ 2,7 millions de comptes-rendus médicaux dé-identifiés issus d'un Centre de Lutte Contre le Cancer (CLCC) français de référence, couvrant la période 2000–2023. La répartition par type de document est détaillée au tableau 1. Les documents sont approximativement répartis à parts égales entre documents **internes** (produits au sein du CLCC : 51,9 % des documents) et documents **externes** (intégrés au dossier patient à partir d'autres établissements ou prestataires, souvent par numérisation OCR : 48,1 % des documents, mais 83,4 % des mots du fait de leur longueur supérieure). Les documents purement administratifs ont été exclus. Le corpus total représente **1,22 milliard de mots** (environ 7 Go de texte brut) après un filtrage qualité reposant sur un seuil de confiance OCR (proportion de caractères non reconnus).

Type de document	Documents	Mots
Documents internes		
Comptes-rendus de consultation	792 881	130 870 465
Comptes-rendus de séjour	328 525	28 245 254
Texte libre	194 843	12 055 223
Début d'affection	57 470	21 583 518
Comptes-rendus d'anatomopathologie	26 291	9 595 263
Documents externes		
Comptes-rendus de laboratoire	504 608	470 941 791
Comptes-rendus de consultation	342 512	225 473 425
Courriers	200 979	93 485 103
Comptes-rendus de séjour	135 517	160 083 968
Comptes-rendus d'histologie	116 373	67 215 927
Total	**2 700 000**	**1 219 549 937**

TABLE 1 – Composition du corpus de pré-entraînement, ventilé par type et provenance des documents.

La dé-identification a été assurée par un pipeline interne combinant expressions régulières et règles métier, validé en interne conformément aux standards de la littérature (Vakili & Dalianis, 2022). L'utilisation secondaire des données repose sur le principe de non-opposition des patients, dans le cadre d'une approbation par le Comité de Gouvernance des Données du centre, en conformité avec le RGPD et les recommandations de la CNIL.

2.2 Classification du statut métastatique

La tâche est formulée comme un problème de classification **binaire** : prédire la présence ou l'absence de métastases à distance à partir d'un compte-rendu oncologique. Le corpus d'évaluation comprend **1 030 comptes-rendus**, à raison d'**un document par patient**, tirés aléatoirement dans le dossier patient informatisé sur les dix dernières années. La taille de ce corpus a été contrainte par la capacité d'annotation manuelle disponible sur une durée fixe de l'attaché de recherche clinique (ARC) en

charge de l'annotation ; aucun équilibrage *a priori* n'a été imposé, afin que la répartition observée reflète la prévalence dans la cohorte. Les classes obtenues sont de **612 cas métastatiques (59,4 %)** et **418 cas non métastatiques (40,6 %)**. Les critères d'inclusion de l'atteinte métastatique ont été établis par l'ARC, spécialisé en oncologie, à partir des mentions explicites de lésions secondaires (extension à distance documentée par imagerie ou histologie).

Ces documents sont inclus dans le corpus de pré-entraînement décrit en §2.1 ; en revanche, leurs **labels de classification n'ont jamais été exposés aux modèles**, puisque le pré-entraînement (MLM ou ELECTRA) est purement non supervisé. La portée de cette inclusion est discutée en §6.2.

Cette tâche est représentative d'un besoin fréquent en oncologie : la constitution automatique de cohortes stratifiées par stade de la maladie pour la recherche clinique, la veille épidémiologique et l'identification des patients éligibles aux essais thérapeutiques.

2.3 NER des toxicités

Définition de la tâche. La NER des toxicités vise à extraire automatiquement, depuis les comptes-rendus oncologiques, les mentions structurées qui permettent de documenter les **événements indésirables liés aux traitements** selon la nomenclature CTCAE v5.0. Concrètement, le modèle reconnaît trois types d'entités : **Médicament** (dénominations communes internationales et noms de protocoles), **Symptôme** (signes cliniques de toxicité) et **Date** (datation des événements). Ces extractions alimentent en aval la constitution de triplets *(traitement, symptôme, date)* utilisés pour le suivi pharmacovigilance et la veille épidémiologique sur les effets secondaires des thérapies oncologiques.

Corpus. Le corpus d'évaluation est issu de la même infrastructure que celui utilisé pour OncoBERT-TOX (article en cours de publication). Il comprend **500 comptes-rendus de consultation oncologique** annotés manuellement, totalisant **234 232 tokens**, dont **4 474 mentions de Date**, **2 956 mentions de Symptôme** et **1 414 mentions de Médicament**. L'annotation utilise le schéma BILOU (*Beginning, Intermediate, Last, Outside, Unique*), portant l'espace des étiquettes à 13 tags. L'entité **Médicament est environ deux fois moins fréquente que Symptôme et trois fois moins fréquente que Date**, ce qui en fait la classe la plus discriminante pour évaluer la robustesse des modèles aux déséquilibres de distribution.

Comme pour la classification (§2.2), ces documents sont inclus dans le corpus de pré-entraînement, mais les annotations BILOU n'ont jamais été exposées aux modèles avant l'étape d'affinage supervisé. Cette inclusion est discutée en §6.2.

3 Modèles et pré-entraînement

3.1 OncoBERT : référence domaine (MLM)

OncoBERT est construit sur CamemBERT-base (Martin *et al.*, 2020), qui implémente l'architecture RoBERTa : 12 couches de transformeurs, 768 dimensions cachées, 12 têtes d'attention, soit 110 millions de paramètres. La stratégie adoptée est le pré-entraînement continu (CPT, *continual pre-training*) sur le corpus décrit en §2.1, avec l'objectif MLM standard : à chaque pas d'entraînement, 15 % des tokens sont sélectionnés, dont 80 % remplacés par `[MASK]`, 10 % laissés inchangés et 10 %

remplacés aléatoirement. Le modèle prédit les tokens originaux par minimisation de l'entropie croisée. Le pré-entraînement a été mené sur **50 itérations** sur un GPU NVIDIA A40, avec une taille de lot de 16 et un taux d'apprentissage de 5×10^{-5}.

3.2 OncoDEBERTa : adaptation domaine par stratégie ELECTRA

OncoDEBERTa repose sur CamemBERTav2 (Antoun *et al.*, 2024), adaptation française de l'architecture DeBERTa-v3 (He *et al.*, 2021). Par rapport à CamemBERT, CamemBERTav2 introduit deux innovations majeures :

— **Attention découplée** : les représentations de contenu et de position relative sont apprises séparément et combinées lors du calcul de l'attention, offrant une meilleure modélisation des relations positionnelles à longue portée.
— **Représentation positionnelle relative** : la position d'un token est encodée relativement aux autres tokens de la séquence plutôt qu'absolument, ce qui améliore la généralisation sur des séquences de longueurs variées, fréquentes dans les textes cliniques.

La stratégie d'entraînement suit le cadre ELECTRA (Clark *et al.*, 2020) : un petit modèle générateur produit des remplacements plausibles de certains tokens, et le modèle discriminateur principal détermine, pour chaque token de la séquence, s'il est original ou s'il a été remplacé. Cette approche, plus efficiente en pré-entraînement que le MLM (Clark *et al.*, 2020), combinée à l'architecture DeBERTa-v3, permet à OncoDEBERTa d'atteindre la convergence en **une seule itération** sur le corpus oncologique, contre 50 pour OncoBERT.

Le pré-entraînement a été conduit sur un GPU NVIDIA A40 (identique à l'infrastructure d'OncoBERT), avec une taille de lot de 4 en entraînement et 8 en évaluation. L'optimiseur est AdamW (Loshchilov & Hutter, 2019) (implémentation fusionnée, $\beta_1 = 0{,}9$, $\beta_2 = 0{,}999$, $\varepsilon = 10^{-8}$), avec un taux d'apprentissage de 5×10^{-5} et une pondération L_2 de $0{,}01$. Le taux d'apprentissage suit une décroissance linéaire après une phase d'échauffement de 1 000 pas. La durée totale du pré-entraînement est d'environ **100 heures de calcul**.

3.3 Modèles de référence

Pour isoler les effets respectifs de l'architecture et du pré-entraînement oncologique, nous incluons trois modèles de référence :

— **CamemBERT-base** (Martin *et al.*, 2020) : modèle MLM généraliste, même architecture qu'OncoBERT mais sans pré-entraînement continu sur données cliniques.
— **CamemBERT-bio** (Touchent *et al.*, 2023) : adaptation biomédicale de CamemBERT, pré-entraînée en continu sur un corpus biomédical français libre de 2,7 Go (publications scientifiques, principalement issues de PubMed). Cette référence permet de tester si un pré-entraînement biomédical *général* peut se substituer à un pré-entraînement *clinique oncologique*.
— **CamemBERTav2** (Antoun *et al.*, 2024) : modèle ELECTRA généraliste, même architecture qu'OncoDEBERTa mais sans pré-entraînement continu sur données cliniques.

Avec OncoBERT et OncoDEBERTa, ces cinq modèles permettent à la fois de dresser un panorama comparatif et de constituer une **comparaison croisée** sur les deux dimensions étudiées (architecture :

BERT vs DeBERTa ; adaptation domaine : avec vs sans CPT oncologique), comme illustré au tableau 2.

	Sans CPT oncologique	Avec CPT oncologique
Architecture BERT	CamemBERT-base	OncoBERT
Architecture DeBERTa	CamemBERTav2	**OncoDEBERTa**

TABLE 2 – Comparaison croisée des modèles évalués selon l'architecture et l'adaptation de domaine.

4 Affinage supervisé et protocole d'évaluation

4.1 Classification du statut métastatique

Le corpus d'affinage et d'évaluation est celui décrit en §2.2 (1 030 comptes-rendus). Pour chaque modèle, une tête de classification linéaire est ajoutée au-dessus du transformer, alimentée par une représentation agrégée obtenue par **moyenne pondérée par le masque d'attention** (*mean pooling*) sur les embeddings de la dernière couche. Le modèle de base est entièrement affiné conjointement à la tête de classification.

L'affinage utilise la **Focal Loss** (Lin *et al.*, 2017) avec $\gamma = 0{,}85$, afin de compenser le léger déséquilibre entre classes (59,4 % vs 40,6 %) sans sur-pénaliser les exemples correctement classés avec une forte confiance. Une stratégie de **décroissance du taux d'apprentissage par couche** (LLRD, *Layer-wise Learning Rate Decay*, facteur $0{,}749$) est appliquée : les couches supérieures reçoivent un taux d'apprentissage plus élevé que les couches inférieures, ce qui préserve les représentations linguistiques générales acquises lors du pré-entraînement tout en autorisant une adaptation plus rapide des couches spécifiques à la tâche.

Les hyperparamètres ont été sélectionnés par optimisation automatique : taux d'apprentissage $= 4{,}64 \times 10^{-5}$, pondération $L_2 = 0{,}071$, ratio de la phase d'échauffement $= 0{,}144$, γ de la Focal Loss $= 0{,}85$, dropout $= 0$. L'entraînement est réalisé sur 10 itérations, avec une taille de lot de 4 et une longueur maximale de séquence de 1 024 tokens, exploitant la capacité de CamemBERTav2 à traiter des séquences longues — atout important pour les comptes-rendus oncologiques dont la longueur médiane dépasse fréquemment 512 tokens. L'optimiseur est AdamW avec décroissance linéaire du taux d'apprentissage après échauffement.

Les performances sont mesurées par l'**accuracy** et le **F1-macro** afin de tenir compte du déséquilibre entre classes.

4.2 NER des toxicités

Le corpus d'affinage et d'évaluation est celui décrit en §2.3 (500 comptes-rendus, 13 étiquettes BILOU). Une couche de classification au niveau token est ajoutée au-dessus du transformeur. Le modèle prédit pour chaque token l'une des 13 étiquettes. L'alignement entre tokens de sous-mots et étiquettes est géré via l'API `word_ids()` du tokeniseur, évitant les erreurs d'alignement connues avec `offset_mapping` et les tokeniseurs SentencePiece.

L'affinage est réalisé sur 10 itérations, avec un taux d'apprentissage de 1×10^{-5}, une taille de lot de 4 en entraînement et 8 en validation. La longueur de séquence maximale est fixée à la **capacité native de chaque architecture** : 1 024 tokens pour les modèles fondés sur DeBERTa-v3 (CamemBERTav2, OncoDEBERTa) qui exploitent une représentation positionnelle relative, et 512 tokens pour les modèles fondés sur RoBERTa (CamemBERT-base, CamemBERT-bio, OncoBERT). Ce choix, identique à celui retenu pour la classification, reflète l'usage des modèles tels qu'ils seraient déployés en pratique, plutôt que de brider artificiellement les modèles DeBERTa à la limite de BERT. L'optimiseur est AdamW avec décroissance linéaire du taux d'apprentissage précédée d'une phase d'échauffement représentant 10 % des pas totaux ; un écrêtage de gradient est appliqué avec une norme maximale de 1,0. Aucune stratégie de sur-échantillonnage n'est employée : la distribution naturelle des entités est préservée, ce qui rend l'entité rare *Médicament* particulièrement discriminante entre modèles.

Les performances sont rapportées en **F1-macro**, **F1-weighted** et F1 par entité, calculés au niveau entité (et non token) pour refléter la performance effective d'extraction.

4.3 Protocole commun

Pour les deux tâches, le protocole d'évaluation repose sur une **validation croisée stratifiée à 5 plis** (stratification sur les labels cibles), garantissant une distribution équilibrée des classes dans chaque pli. Les performances rapportées correspondent aux moyennes sur les 5 plis, accompagnées de leur écart-type. La significativité des écarts entre modèles est évaluée par t-test apparié unilatéral et test de Wilcoxon apparié unilatéral sur les 5 plis. Les expériences sont conduites avec des graines aléatoires fixes pour assurer la reproductibilité, et tracées via MLflow.

5 Résultats

5.1 Classification du statut métastatique

Le tableau 3 présente les performances des cinq modèles.

Modèle	Accuracy	F1-macro
CamemBERT-base	$0{,}58 \pm 0{,}24$	$0{,}51 \pm 0{,}30$
CamemBERT-bio	$0{,}67 \pm 0{,}24$	$0{,}61 \pm 0{,}30$
CamemBERTav2	$0{,}82 \pm 0{,}04$	$0{,}80 \pm 0{,}04$
OncoBERT	$0{,}84 \pm 0{,}02$	$0{,}83 \pm 0{,}03$
OncoDEBERTa	$\mathbf{0{,}89 \pm 0{,}03}$	$\mathbf{0{,}88 \pm 0{,}03}$

TABLE 3 – Classification du statut métastatique : moyenne $\pm$ écart-type sur 5 plis.

OncoDEBERTa atteint 0,88 en F1-macro, soit **+5 points** sur OncoBERT (t-test apparié unilatéral, $p = 0{,}013$), **+8 points** sur CamemBERTav2 non spécialisé ($p = 0{,}002$) et **+27 points** sur CamemBERT-bio. Le gain est confirmé par les deux tests appariés (Wilcoxon et t-test), tout en gardant à l'esprit la limite intrinsèque du Wilcoxon avec $n = 5$ plis (p-valeur minimale atteignable : 0,031).

Plusieurs observations se dégagent. Les deux modèles généralistes (CamemBERT-base et CamemBERT-bio) affichent des **écart-types très élevés** ($\sigma \approx 0{,}30$ sur F1-macro), témoignant d'une variabilité d'apprentissage importante d'une exécution à l'autre. Cette instabilité disparaît dès lors qu'au moins un des deux leviers est activé : architecture moderne (CamemBERTav2, $\sigma = 0{,}04$) ou pré-entraînement oncologique (OncoBERT, $\sigma = 0{,}03$). **Chacun des deux leviers stabilise individuellement l'optimisation**, mais seule leur combinaison produit la meilleure performance moyenne.

Le résultat de CamemBERT-bio est particulièrement instructif : malgré un pré-entraînement biomédical sur un corpus PubMed français, il n'apporte qu'un gain modeste de 10 points par rapport à CamemBERT-base, restant 22 points en deçà d'OncoBERT et 27 points en deçà d'OncoDEBERTa. Ce résultat confirme que **le pré-entraînement sur la littérature scientifique biomédicale ne se substitue pas à l'exposition aux narratifs cliniques réels**.

5.2 NER des toxicités

Le tableau 4 présente les métriques agrégées sur la tâche de NER.

Modèle	F1-macro	F1-weighted	F1-Médicament
CamemBERT-base	$0{,}82 \pm 0{,}01$	$0{,}84 \pm 0{,}01$	$0{,}72 \pm 0{,}01$
CamemBERT-bio	$0{,}83 \pm 0{,}01$	$0{,}85 \pm 0{,}01$	$0{,}73 \pm 0{,}01$
CamemBERTav2	$0{,}73 \pm 0{,}02$	$0{,}75 \pm 0{,}01$	$0{,}63 \pm 0{,}03$
OncoBERT	$0{,}84 \pm 0{,}01$	$0{,}86 \pm 0{,}01$	$0{,}74 \pm 0{,}01$
OncoDEBERTa	$\mathbf{0{,}86 \pm 0{,}01}$	$\mathbf{0{,}87 \pm 0{,}01}$	$\mathbf{0{,}81 \pm 0{,}02}$

TABLE 4 – NER des toxicités : métriques agrégées (moyenne $\pm$ écart-type sur 5 plis).

OncoDEBERTa obtient les meilleures performances sur les trois métriques. Les écart-types sont faibles et homogènes pour tous les modèles sur cette tâche, contrairement à la classification.

Focus sur l'entité Médicament. L'entité *Médicament*, environ deux fois moins fréquente que *Symptôme* et trois fois moins fréquente que *Date*, constitue la cible la plus discriminante pour évaluer la robustesse aux classes rares (tableau 5).

Modèle	F1-Médicament	Δ vs OncoBERT	Significativité
CamemBERT-base	$0{,}72 \pm 0{,}01$	$-0{,}02$	—
CamemBERT-bio	$0{,}73 \pm 0{,}01$	$-0{,}01$	—
CamemBERTav2	$0{,}63 \pm 0{,}03$	$-0{,}11$	—
OncoBERT	$0{,}74 \pm 0{,}01$	—	—
OncoDEBERTa	$\mathbf{0{,}81 \pm 0{,}02}$	$\mathbf{+0{,}06}$	$p = 0{,}001$

TABLE 5 – F1-Médicament : analyse comparative.

OncoDEBERTa réalise un gain de **+6 points** sur l'entité *Médicament*, le gain relatif le plus élevé parmi toutes les entités. Le test apparié unilatéral confirme la significativité statistique de ce gain

(t-test $p = 0{,}001$; Wilcoxon $p = 0{,}031$). Deux facteurs peuvent expliquer ce résultat. D'une part, l'**attention découplée** de DeBERTa-v3 modélise séparément contenu et position, ce qui améliore la capture des mentions médicamenteuses qui apparaissent dans des contextes syntaxiques variés (début ou fin d'énumération, formes abrégées, co-références dans les cycles de chimiothérapie). D'autre part, l'**exposition au corpus oncologique** lors du pré-entraînement continu enrichit les représentations des dénominations communes internationales et de leurs variantes textuelles (abréviations, noms de protocoles) peu présentes dans les corpus généralistes.

Un résultat contre-intuitif mérite attention : CamemBERTav2 sans pré-entraînement continu ($0{,}63$) est en retrait de 11 points sur OncoBERT et de 9 points sur CamemBERT-base pour cette entité. Pour la détection d'entités rares dans un domaine très spécialisé, l'**adaptation de domaine prime sur le gain architectural** : CamemBERTav2 hors domaine est pénalisé par l'absence d'exposition aux mentions médicamenteuses cliniques lors du pré-entraînement, un déficit que l'architecture seule ne compense pas. À l'inverse, CamemBERT-bio (pré-entraîné sur PubMed) ne fait pas mieux que CamemBERT-base ($0{,}73$ vs $0{,}72$), confirmant là encore que le pré-entraînement biomédical général n'apporte pas de gain mesurable sur les mentions médicamenteuses cliniques.

5.3 Synthèse

L'entité *Médicament* permet d'illustrer clairement l'interaction entre les deux facteurs étudiés. Le tableau 6 présente les scores F1 selon la comparaison croisée architecture $\times$ adaptation de domaine.

	Sans CPT onco	Avec CPT onco	Δ Domaine
Architecture BERT	$0{,}72$	$0{,}74$	$\mathbf{+0{,}02}$
Architecture DeBERTa	$0{,}63$	$\mathbf{0{,}81}$	$\mathbf{+0{,}18}$
Δ Architecture	$\mathbf{-0{,}09}$	$\mathbf{+0{,}07}$	

TABLE 6 – F1-Médicament selon l'architecture et l'adaptation de domaine.

Trois observations centrales se dégagent :

1. L'architecture DeBERTa seule ne suffit pas. Sans adaptation au domaine, CamemBERTav2 régresse de 9 points par rapport à CamemBERT-base sur cette entité rare ($0{,}63$ contre $0{,}72$). La supériorité théorique de l'attention découplée ne se matérialise pas sans exposition au lexique spécialisé.

2. L'adaptation de domaine seule apporte peu sur cette classe. Appliqué à l'architecture BERT, le pré-entraînement continu oncologique n'améliore que marginalement la détection de l'entité *Médicament* (+2 points, de $0{,}72$ à $0{,}74$) : le plafond de performance de l'architecture BERT semble atteint sur cette classe rare.

3. La combinaison produit un gain disproportionné. Un modèle d'effets strictement additifs prédirait un score de $0{,}72 - 0{,}09 + 0{,}02 = \mathbf{0{,}65}$ pour OncoDEBERTa. Le score observé est $0{,}81$, soit **+16 points au-dessus de la prédiction additive**. Les deux facteurs se renforcent mutuellement : l'attention découplée exploite efficacement les représentations spécialisées acquises lors du pré-entraînement continu, tandis que ces représentations bénéficient en retour de la modélisation positionnelle relative pour les mentions en contexte syntaxique variable.

Ce schéma d'interaction non additive s'observe également sur le F1-macro de NER : la prédiction additive (0,76) est dépassée de 10 points par le score observé (0,86). Sur la tâche de classification, en revanche, les effets individuels sont si forts qu'une décomposition additive perd sa pertinence (la somme dépasserait 1,0) ; on retient simplement qu'OncoDEBERTa reste significativement supérieur au meilleur modèle mono-factoriel (+5 points sur OncoBERT, $p = 0{,}013$). **OncoDEBERTa n'est donc pas la simple addition des apports d'OncoBERT et de CamemBERTav2 : l'articulation entre architecture et corpus de domaine est déterminante pour les gains obtenus, en particulier sur les classes rares.**

6 Discussion

6.1 Efficacité computationnelle comme argument de déploiement

OncoDEBERTa atteint la convergence en une seule itération là où OncoBERT nécessitait 50 itérations sur le même corpus de 2,7 millions de documents et la même infrastructure matérielle (GPU NVIDIA A40). Cette efficience reflète une propriété centrale du cadre ELECTRA, déjà rapportée par (Clark *et al.*, 2020) et étendue au français par (Antoun *et al.*, 2024) : ELECTRA atteint des performances comparables ou supérieures à celles du MLM avec un coût de pré-entraînement significativement moindre.

Ce résultat a une conséquence pratique importante : un centre hospitalier disposant d'une infrastructure GPU modeste peut produire un modèle de langue oncologique de meilleure qualité en une fraction du temps requis par le MLM. Dans une perspective de **souveraineté des données**, où les comptes-rendus cliniques ne peuvent quitter l'infrastructure hospitalière, cette accessibilité computationnelle est un facteur clé de déployabilité.

6.2 Limites

Plusieurs limites doivent être soulignées.

Premièrement, les corpus d'évaluation sont internes à un seul CLCC. Une validation d'OncoDEBERTa sur des données multi-institutionnelles reste nécessaire pour confirmer que les gains observés se maintiennent malgré la variabilité inter-centres des styles rédactionnels.

Deuxièmement, le corpus annoté pour l'affinage est de taille modeste, notamment pour la NER (500 comptes-rendus) et la classification métastatique (1 030 comptes-rendus), la limite venant de la capacité d'annotation manuelle experte disponible. Les scores observés pourraient évoluer avec des jeux d'annotation plus larges ; en particulier, l'entité *Médicament* (F1 $= 0{,}81$) présente encore une marge de progression.

Troisièmement, notre évaluation porte sur deux tâches de structuration. D'autres tâches cliniques pertinentes (extraction de stade TNM, détection de réponse au traitement, identification de comorbidités) n'ont pas encore été évaluées pour OncoDEBERTa, et il n'est pas garanti que les gains observés se généralisent uniformément.

Quatrièmement, les corpus d'évaluation des deux tâches (§2.2 et §2.3) sont inclus dans le corpus de pré-entraînement de 2,7 millions de documents. Si les têtes de classification et de NER n'ont jamais

été exposées aux labels d'évaluation, le modèle de base a néanmoins rencontré le texte brut de ces documents lors du pré-entraînement non supervisé. Cette pratique reste standard dans la littérature de pré-entraînement de domaine (BioBERT, ClinicalBERT, DrBERT) et ne constitue pas une fuite de données au sens supervisé, mais elle pourrait introduire un avantage de mémorisation textuelle. Une évaluation sur des documents postérieurs au corpus de pré-entraînement, ou issus d'autres centres, permettrait de mieux quantifier cet effet. Notons néanmoins qu'un éventuel effet de mémorisation devrait avantager OncoBERT (50 itérations d'exposition au corpus) davantage qu'OncoDEBERTa (une seule itération), ce qui suggère que les gains d'OncoDEBERTa ne sont pas imputables à un avantage de mémorisation.

6.3 Perspectives

Plusieurs axes de travail prolongent naturellement ces résultats. D'une part, un pré-entraînement multi-institutionnel via apprentissage fédéré permettrait d'enrichir le corpus de pré-entraînement sans centraliser les données, renforçant à la fois la robustesse du modèle et sa généralisabilité. D'autre part, l'intégration d'OncoDEBERTa dans des chaînes de traitement opérationnelles, notamment pour la constitution automatique de cohortes en recherche clinique ou pour l'alimentation de registres oncologiques, représente un axe de transfert à fort impact.

Enfin, la question de la **diffusion du modèle en accès ouvert** est centrale pour maximiser l'impact de ce travail auprès de la communauté francophone en TAL appliqué à la santé . Notre objectif est de publier OncoDEBERTa sous licence ouverte afin de permettre son adoption par d'autres centres hospitaliers et équipes de recherche. Des discussions sont actuellement en cours avec la CNIL pour définir les conditions de cette diffusion, en garantissant que le modèle publié ne puisse être exploité pour la ré-identification des patients dont les données ont contribué au pré-entraînement. Des approches de désapprentissage machine (Boutet *et al.*, 2025), combinées à une évaluation formelle du risque de ré-identification via des attaques ciblées, sont explorées en parallèle pour préparer une diffusion conforme aux exigences réglementaires françaises et européennes.

7 Conclusion

Nous avons présenté OncoDEBERTa, un modèle de langue clinique français en oncologie combinant l'architecture DeBERTa-v3 (CamemBERTav2) et une stratégie de pré-entraînement ELECTRA. Sur le même corpus de 2,7 millions de comptes-rendus oncologiques dé-identifiés, OncoDEBERTa atteint la convergence en **une seule itération** contre 50 pour OncoBERT, sur la même infrastructure matérielle. Il établit de nouvelles meilleures performances sur la classification binaire du statut métastatique (F1-macro $0{,}88 \pm 0{,}03$, +5 pts, $p = 0{,}013$) et la NER des toxicités (F1-macro $0{,}86 \pm 0{,}01$, +2 pts ; F1-Médicament $0{,}81 \pm 0{,}02$, +6 pts, $p = 0{,}001$). L'analyse de l'entité rare *Médicament* révèle que ces gains résultent d'une interaction positive entre architecture DeBERTa et adaptation oncologique : l'un sans l'autre ne suffit pas à reproduire la performance combinée, et un pré-entraînement biomédical général (CamemBERT-bio) ne se substitue pas à l'exposition aux narratifs cliniques réels.

OncoDEBERTa s'impose comme nouveau modèle de référence pour le NLP clinique oncologique français. Son efficacité computationnelle, combinée à ses performances supérieures, en fait un candidat particulièrement adapté au déploiement en milieu hospitalier, dans une logique de traitement souverain des données de santé.

Références

ANTOUN W., KULUMBA F., TOUCHENT R., DE LA CLERGERIE E. V., SAGOT B. & SEDDAH D. (2024). CamemBERT 2.0 : A smarter french language model aged to perfection. arXiv : 2411.08868.

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

CLARK K., LUONG M.-T., LE Q. V. & MANNING C. D. (2020). ELECTRA : Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations (ICLR)*. arXiv : 2003.10555.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT)*, p. 4171–4186. DOI : 10.18653/v1/N19-1423.

HE P., GAO J. & CHEN W. (2021). DeBERTaV3 : Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. arXiv : 2111.09543.

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. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 16207–16221. DOI : 10.18653/v1/2023.acl-long.896.

LIN T.-Y., GOYAL P., GIRSHICK R., HE K. & DOLLÁR P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, p. 2980–2988. DOI : 10.1109/ICCV.2017.324.

LOSHCHILOV I. & HUTTER F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*. arXiv : 1711.05101.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE E. V., 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)*, p. 7203–7219. DOI : 10.18653/v1/2020.acl-main.645.

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. In *Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*. 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 Thirteenth Language Resources and Evaluation Conference (LREC)*, p. 4245–4252.