ModernCamemBERT-bio : un encodeur biomédical et clinique français à contexte long

Rian Touchent Éric Villemonte de la Clergerie
Inria, Sorbonne Université
48 rue Barrault 75013 Paris, 21 rue de l'école de médecine 75006 Paris
{rian.touchent,eric.de_la_clergerie}@inria.fr

RÉSUMÉ

Les encodeurs biomédicaux français existants sont limités à 512 tokens de contexte. Cela les rend peu adaptés aux documents cliniques longs tels que les comptes rendus d'hospitalisation. Nous présentons ModernCamemBERT-bio, un encodeur biomédical français avec un contexte de 8192 tokens, obtenu par pré-entraînement continu de ModernCamemBERT avec un budget de 10 milliards de tokens. Le jeu de données d'entraînement est constitué à partir de six sources francophones et annoté par un LLM selon quatre signaux de qualité ; une étude d'ablation identifie ceux qui améliorent les performances en aval. La recette de pré-entraînement inclut un détour par un objectif causal (CLM) avant un retour au MLM, dont nous mesurons le gain par rapport à un pré-entraînement continu MLM standard. Le modèle atteint 61,6% de F1 moyen sur 8 tâches biomédicales et cliniques françaises. Le détour CLM apporte +2,8pp par rapport à un contrôle MLM identique en données et en budget. ModernCamemBERT-bio dépasse aussi les encodeurs biomédicaux français existants, qui sont limités à 512 tokens : +23,3pp vs CamemBERT-bio, +11,5pp vs DrBERT. Une version Large (350M) porte la moyenne à 64,2% (+1,2pp sur le contrôle MLM). Nous publions les modèles sous licence libre.

ABSTRACT

ModernCamemBERT-bio : A Long-Context French Biomedical Encoder for Clinical Documents

Existing French biomedical encoders are limited to a 512-token context. This makes them poorly suited to long clinical documents such as discharge summaries and hospital reports. We present ModernCamemBERT-bio, a French biomedical encoder with an 8192-token context, obtained by continued pretraining of ModernCamemBERT with a 10-billion-token training budget. The training data is built from six French-language sources and annotated by an LLM along four quality signals ; an ablation study identifies those that improve downstream performance. The pretraining recipe includes a detour through a causal language modeling (CLM) objective before returning to MLM, whose gain we measure against standard MLM continued pretraining. The model reaches 61.6% average F1 on 8 French biomedical and clinical tasks. The CLM detour adds +2.8pp over a matched MLM control (same data and budget). ModernCamemBERT-bio also outperforms existing French biomedical encoders, which are restricted to a 512-token context : +23.3pp vs CamemBERT-bio, +11.5pp vs DrBERT. A Large variant (350M) brings the average to 64.2% (+1.2pp over the MLM control). We release the models under an open license.

MOTS-CLÉS : pré-entraînement continu, biomédical, clinique, encodeur, français, contexte long, curation de données.

KEYWORDS: continued pretraining, biomedical, clinical, encoder, French, long context, data

1 Introduction

Les deux modèles de référence pour le traitement automatique du langage biomédical français sont CamemBERT-bio (Touchent & de la Clergerie, 2024) et DrBERT (Labrak *et al.*, 2023). CamemBERT-bio adapte CamemBERT (Martin *et al.*, 2020) au domaine biomédical par pré-entraînement continu sur biomed-fr, un corpus de 413 millions de mots. DrBERT entraîne un modèle RoBERTa (Liu *et al.*, 2019) à partir d'une initialisation aléatoire sur NACHOS, un corpus médical francophone de 1,1 milliard de mots. Ces deux modèles restent largement utilisés au moment de la rédaction, avec environ 8 000 téléchargements mensuels pour CamemBERT-bio et 1 000 pour DrBERT sur HuggingFace. Ils présentent cependant plusieurs limites pour les usages cliniques.

Premièrement, ils sont limités à un contexte de 512 tokens. Cela tronque la plupart des documents cliniques longs : comptes rendus d'hospitalisation, lettres de sortie, protocoles, articles de revue. Sur les tâches cliniques multi-étiquettes que nous évaluons, CamemBERT-bio plafonne sous 42% de F1 et descend sous 15% sur plusieurs d'entre elles.

Deuxièmement, leurs corpus de pré-entraînement n'ont pas fait l'objet d'un filtrage par qualité avec une approche neuronale. Le pipeline de NACHOS et celui de biomed-fr appliquent un nettoyage classique (déduplication, filtres heuristiques, sélection de sources médicales), mais aucun ne note les documents par un classifieur de qualité appris. Or les travaux récents sur les données généralistes (Penedo *et al.*, 2024) montrent qu'un filtrage par score éducatif annoté par un LLM, distillé dans un classifieur léger, améliore significativement les modèles. Cette approche n'a pas encore été adaptée au biomédical français.

Troisièmement, leur objectif de pré-entraînement est le MLM standard avec 15% de masquage. Cet objectif a récemment été remis en question pour le pré-entraînement des encodeurs : Gisserot-Boukhlef *et al.* (2025) montrent qu'un schéma biphasique combinant modélisation causale (CLM) puis MLM peut surpasser le MLM seul à budget équivalent sur des tâches généralistes anglaises.

Nous présentons ModernCamemBERT-bio, un successeur de CamemBERT-bio qui répond à ces trois points. Nos contributions sont :

1. Le corpus biomédical français *biomed-fr-v2*, annoté par un LLM selon quatre signaux de qualité et neuf types de contenu. Une étude d'ablation quantifie l'effet de chaque signal et de chaque type sur les performances en aval.
2. ModernCamemBERT-bio, un encodeur biomédical français à contexte 8192 tokens en deux tailles (Base 149M, Large 350M), publié sous licence libre. Il atteint l'état de l'art sur 8 tâches biomédicales et cliniques françaises, avec les gains les plus marqués sur les tâches cliniques à documents longs.

2 Travaux connexes

Encodeurs à contexte long. ModernBERT (Warner *et al.*, 2025) introduit une architecture d'encodeur à contexte 8192 tokens (Flash Attention (Dao *et al.*, 2022), embeddings positionnels rotatifs (Su *et al.*, 2024), attention locale/globale alternée). ModernCamemBERT (Antoun *et al.*, 2025)

adapte cette architecture au français. BioClinical-ModernBERT (Sounack *et al.*, 2025) l'applique au biomédical anglais par pré-entraînement continu sur PubMed et MIMIC, en deux phases (30 % puis 15 % de masquage), sans filtrage par classifieur de qualité.

Curation de données de pré-entraînement. FineWeb (Penedo *et al.*, 2024) propose un pipeline de curation à grande échelle pour CommonCrawl ; son sous-ensemble FineWeb-Edu sélectionne les documents éducatifs au moyen d'un classifieur entraîné sur des annotations produites par Llama-3-70B-Instruct (Grattafiori *et al.*, 2024) (échelle additive 0–5). Biomed-Enriched (Touchent *et al.*, 2025) transpose cette approche au biomédical anglais avec une granularité au niveau du paragraphe sur PubMed Central, en distillant les annotations LLM dans un classifieur XLM-RoBERTa (Conneau *et al.*, 2020) appliqué au corpus complet. Aucun équivalent n'existe pour le biomédical français à notre connaissance.

Objectifs d'entraînement pour le pré-entraînement continu. Gisserot-Boukhlef *et al.* (2025) comparent MLM pur, CLM pur, et un schéma biphasique CLM puis MLM, lors d'un pré-entraînement de zéro sur de l'anglais généraliste. Ils montrent qu'un pré-entraînement séquentiel CLM puis MLM peut dominer le MLM pur à budget égal, et qu'initialiser un MLM depuis un *checkpoint* CLM est plus efficace que de continuer un *checkpoint* MLM. Leur protocole reste limité à l'entraînement de zéro sur des données généralistes ; le cas du pré-entraînement continu de domaine n'a, à notre connaissance, pas été étudié.

3 Corpus biomed-fr-v2

3.1 Sources

Nous avons constitué biomed-fr-v2 à partir de six sources francophones complémentaires. HAL, ISTEX et FineWiki-bio couvrent la littérature scientifique biomédicale. EMEA apporte les notices de médicaments, E3C des cas cliniques. Des manuels synthétiques générés à partir d'ontologies médicales complètent le corpus avec le vocabulaire des nomenclatures administratives françaises. La composition du mélange d'entraînement (10 milliards de tokens au total, après suréchantillonnage par qualité décrit en §3.4) est donnée dans le tableau 2.

— **HAL** : thèses médicales, articles de recherche et rapports déposés sur l'archive ouverte HAL, filtrés sur la section biomédecine.
— **ISTEX** : articles scientifiques français issus du corpus ISTEX, filtrés sur les domaines de la biologie et de la médecine.
— **FineWiki-bio** : sous-ensemble biomédical de FineWiki, le pendant Wikipédia de FineWeb (Penedo *et al.*, 2024), restreint au français et filtré pour le domaine biomédical.
— **EMEA** (Névéol *et al.*, 2014) : notices de médicaments publiées par l'Agence européenne des médicaments en français.
— **E3C** (Magnini *et al.*, 2020) : corpus clinique multilingue (couche 3, partition française), composé de cas cliniques extraits de revues médicales et de thèses.
— **Manuels synthétiques d'ontologie** : textes pédagogiques générés à partir de trois ontologies médicales françaises publiées par l'Agence du Numérique en Santé au format RDF/OWL :

CIM-10 (diagnostics, 19 161 codes), CCAM (actes, 38 191 codes) et ATC (médicaments, 6 950 codes). Des marches aléatoires dans le graphe d'ontologie (1,3 million de marches au total) parcourent les relations hiérarchiques et, pour CIM-10 FR PMSI, les relations causales, de manifestation et d'exclusion publiées par l'ANS (par exemple : E11 « diabète de type 2 » cause N08.3 « glomérulopathie diabétique »). Un LLM (Qwen3-235B-A22B-Instruct, Yang *et al.*, 2025) reformule chaque marche en un paragraphe de prose médicale tout en préservant les codes et relations d'origine. Le pipeline complet de génération est décrit dans Touchent & de la Clergerie (2026b).

HAL et ISTEX sont distribués sous des formats hétérogènes : certains documents sont accompagnés d'un XML structuré, d'autres uniquement de PDF, parfois issus de scans imprimés. Pour les documents sans XML propre, nous appliquons le pipeline d'extraction et de nettoyage décrit à la section 3.2. FineWiki-bio, EMEA, E3C et les manuels synthétiques sont déjà de haute qualité et intégrés directement au mélange final, sans passer par ce pipeline.

3.2 Pipeline de curation

Les documents HAL et ISTEX livrés sous forme de PDF (en particulier les articles antérieurs aux années 1990, issus de scans imprimés) contiennent, une fois convertis en texte, des artefacts d'océrisation, des références bibliographiques, des en-têtes répétitifs et du contenu non pertinent. Nous leur appliquons un pipeline d'extraction et de nettoyage par LLM.

Extraction et nettoyage. Chaque document est découpé en segments d'environ 5 000 tokens en coupant aux fins de paragraphes. Un LLM (Qwen3-Next-80B-A3B-Instruct, Yang *et al.*, 2025) reçoit chaque segment et produit une sortie JSON structurée contenant les passages sémantiquement complets, en retirant les références, métadonnées, en-têtes et légendes, et en corrigeant les erreurs d'océrisation (ligatures, encodage). Les passages extraits doivent contenir entre 30 et 600 mots et comporter au moins 60% de caractères alphabétiques. Une déduplication par similarité de Jaccard (seuil 0,80, fenêtre glissante de 5 passages) est ensuite appliquée, suivie de la chaîne de filtres FineWeb-2 pour le français : correction d'encodage, filtres de répétition et de qualité Gopher (Rae *et al.*, 2021) et FineWeb (Penedo *et al.*, 2024).

Vérification de fidélité. Pour s'assurer que le LLM n'introduit pas de contenu absent du source, nous mesurons la concordance caractère entre chaque passage extrait et son document source sur 300 passages ISTEX (après normalisation typographique). 94,4% des passages ont un recouvrement $\geq$ 90% et 84,1% un recouvrement $\geq$ 99%. Les mots du passage extrait absents du document source (3,4% en moyenne, 0% en médiane) proviennent essentiellement de corrections d'océrisation (ligatures, caractères substitués) et de la mise en prose de tableaux, et non d'hallucinations du LLM. Les distributions complètes et deux exemples sont donnés en annexe E.

Annotation qualité par LLM. Le même LLM annote chaque passage selon quatre signaux de qualité notés de 1 à 10 : score éducatif (`edu`, moyenne 6,5), richesse de contenu (`cont`, 6,8), qualité rédactionnelle (`writ`, 7,6) et précision terminologique (`term`, 7,6). Un type de contenu est également attribué parmi neuf catégories : `research_findings` (526k passages), `medical_knowledge` (388k), `clinical_guidance` (218k), `research_methodology`, `drug_information`,

TABLE 1 – Études d'ablation sur biomed-fr-v2 : (a) effet du retrait d'un type de contenu ; (b) effet des filtres par signaux de qualité. Toutes les ablations utilisent ModernCamemBERT-base (149M) entraîné en MLM standard (masquage 30 %) sur 10 milliards de tokens, sans phase de décroissance ni détour CLM, pour isoler l'effet des données.

(a) Ablation par type de contenu (3 tâches, 9 seeds). On retire un type et on mesure Δ par rapport au corpus complet.

Type retiré	F1	Δ
`drug_information`	64,35	+0,67
`research_findings`	64,24	+0,56
`policy_administrative`	64,07	+0,39
aléatoire	63,72	+0,04
complet (réf.)	63,68	réf.
`medical_knowledge`	63,55	−0,13
`research_methodology`	63,51	−0,17

(b) Ablation des signaux de qualité (6 tâches, 9 seeds). Δ par rapport à « sans filtre ».

Filtre	F1	Δ
edu≥8 ET cont≥8	58,50	+0,78
edu≥7 ET cont≥7	**58,39**	**+0,67**
edu≥6 ET cont≥6	58,01	+0,29
edu≥7 seul	58,09	+0,37
cont≥7 seul	58,07	+0,35
sans filtre (réf.)	57,72	réf.
term≥7 seul	57,49	−0,23
writ≥7 seul	57,35	−0,37

`background_review`, `policy_administrative`, `patient_case` (36k) et `other`. Au total, 2,16 millions de passages sont annotés.

3.3 Filtrage par type de contenu

Nous mesurons l'effet de chaque type de contenu par ablation. Le corpus est d'abord équilibré entre les neuf types pour qu'ils contiennent chacun le même nombre de tokens ; nous retirons ensuite un type complet et comparons au corpus équilibré sur trois tâches d'évaluation (DiaMed, FrACCO-30 et EMEA, 9 seeds). Le tableau 1a montre que retirer `drug_information` améliore les performances de $+0{,}67$pp, tandis que retirer `medical_knowledge` les dégrade de $-0{,}13$pp. Un contrôle aléatoire, retirant la même quantité de tokens au hasard (indépendamment du type), donne $+0{,}04$pp, confirmant que l'effet est spécifique au type de contenu.

Nous excluons `drug_information` (redondant avec EMEA) et `other` du corpus final. Les paragraphes de moins de 50 tokens sont également exclus.

3.4 Filtrage et suréchantillonnage par qualité

Nous testons ensuite l'effet du seuillage sur les signaux de qualité (tableau 1b). Le filtrage conjoint `edu`$\geq$7 ET `cont`$\geq$7 produit un gain de $+0{,}67$pp ; un seuil plus strict (edu≥8 ET cont≥8) atteint $+0{,}78$pp. Pris seuls, edu et cont apportent chacun un gain plus modeste ($+0{,}37$pp et $+0{,}35$pp), ce qui justifie leur combinaison. À l'inverse, la qualité rédactionnelle et la précision terminologique dégradent les performances lorsqu'elles sont utilisées comme filtres ($-0{,}37$ et $-0{,}23$pp). Cette dégradation provient probablement du fait que ces filtres éliminent des documents cliniques informels mais informatifs, tels que des notes de cas et des comptes rendus peu édités.

Plutôt qu'un filtrage strict, nous adoptons un suréchantillonnage inspiré de FineWeb (Penedo *et al.*, 2024). Chaque article reçoit un ratio de qualité, défini comme la proportion de ses paragraphes

TABLE 2 – Composition du mélange d'entraînement biomed-fr-v2 après suréchantillonnage par qualité.

Source	Tokens	%	Contenu
Articles biomédicaux curés	7B	70%	HAL + ISTEX + FineWiki-bio
Manuels d'ontologie (synth.)	2B	20%	CIM-10, CCAM, ATC
EMEA (notices de médicaments)	600M	6%	Pharmacologie
E3C (phrases cliniques)	400M	4%	Cas patients
Total	**10B**	100%	

satisfaisant edu$\geq$7 ET cont$\geq$7. Les articles sont ensuite suréchantillonnés proportionnellement à ce ratio (coefficients allant de 4,3$\times$ pour les articles à 1–25% de paragraphes de qualité jusqu'à 34,1$\times$ pour ceux à 100%, cf. annexe D). Les articles avec 0% de paragraphes de qualité sont exclus du corpus final.

3.5 Composition finale

Le mélange d'entraînement final, après nettoyage et suréchantillonnage par qualité, contient 10 milliards de tokens (tableau 2). Les articles HAL et ISTEX (passés dans le pipeline de la section 3.2) et FineWiki-bio (intégré tel quel) sont regroupés dans la ligne « articles biomédicaux curés ». Les trois autres sources (EMEA, E3C, manuels synthétiques) sont également intégrées telles quelles.

4 Pré-entraînement

4.1 Architecture

Nous partons de ModernCamemBERT (Antoun *et al.*, 2025), une adaptation française de l'architecture ModernBERT (Warner *et al.*, 2025). Nous entraînons deux tailles : Base (149M paramètres, 22 couches, 768 dimensions cachées, 12 têtes d'attention) et Large (350M paramètres, 28 couches, 1024 dimensions cachées, 16 têtes). Les deux partagent le même vocabulaire de 32 768 tokens et un contexte maximal de 8192 tokens. L'architecture combine Flash Attention (Dao *et al.*, 2022), embeddings positionnels rotatifs (Su *et al.*, 2024), et attention locale (fenêtre de 128) alternée avec une attention globale toutes les trois couches.

4.2 Recette : un détour par un objectif causal

Le pré-entraînement continu d'un encodeur utilise presque toujours l'objectif de modélisation de langage masqué (MLM), identique à celui du pré-entraînement initial. Ce choix conserve l'attention bidirectionnelle et s'applique directement à l'architecture d'un encodeur. Gisserot-Boukhlef *et al.* (2025) ont cependant montré qu'un schéma biphasique CLM puis MLM, appliqué *de zéro*, surpasse le MLM seul sur des tâches généralistes en anglais. Nous transposons cette idée au pré-entraînement

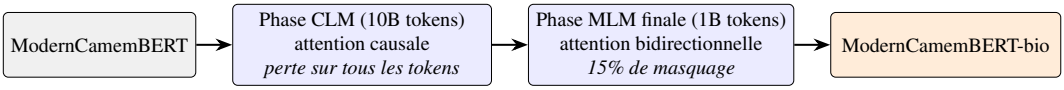

FIGURE 1 – Pipeline de pré-entraînement continu : phase CLM (détour) puis phase MLM finale qui rétablit l'attention bidirectionnelle. Détails en §4.2.

continu sur un domaine cible : nous insérons une phase CLM comme *détour* temporaire dans le pré-entraînement continu d'un encodeur MLM existant, avant un retour au MLM (figure 1).

Le pré-entraînement se déroule en deux phases (figure 1). La **phase CLM** (10B tokens, 90% du budget) remplace l'attention bidirectionnelle par une attention causale et entraîne tous les tokens, contre 15 à 30 % pour MLM. La **phase MLM finale** (1B tokens, 10% du budget, masquage 15%, taux d'apprentissage décroissant) rétablit l'attention bidirectionnelle et l'interface d'évaluation standard des encodeurs sans effacer les représentations acquises pendant la phase CLM.

Une explication plausible : le signal d'entraînement plus dense du CLM modifie davantage les représentations du modèle initial. Dans le cas généraliste de Gisserot-Boukhlef *et al.* (2025), ce remaniement est peu utile, mais l'adaptation à un nouveau domaine en bénéficie. Pour mesurer ce gain, nous entraînons un modèle de contrôle avec un pré-entraînement MLM standard sur les mêmes données et au même budget (10B + 1B), avec un masquage de 30% en phase 1 puis 15% en phase finale.

L'entraînement utilise Composer (The Mosaic ML Team, 2021). Hyperparamètres : AdamW décalé ($\beta_1 = 0{,}9$, $\beta_2 = 0{,}98$, $\epsilon = 10^{-6}$, weight decay 10^{-5} hors biais et normalisations), taille de lot globale de 384 séquences de 8192 tokens ($\sim$ 3,1M tokens par pas), précision mixte bf16, 4 GPU H100 par exécution. Les documents sont concaténés en séquences de 8192 tokens (*sequence packing*) pour éliminer les tokens de *padding*. La version Base utilise le budget de 10B+1B tokens décrit ci-dessus et prend 4 heures. La version Large utilise un budget plus élevé de 50B+5B tokens et prend environ 81 heures.

5 Évaluation

5.1 Tâches

Nous évaluons sur 8 tâches biomédicales et cliniques françaises couvrant la classification, la classification multi-étiquette et la reconnaissance d'entités nommées (NER) :

- **DiaMed** (Labrak *et al.*, 2024) : classification de spécialités médicales (cas cliniques annotés par chapitres CIM-10)
- **FrACCO-30 / FrACCO-100** (Pignat *et al.*, 2025) : classification multi-étiquette d'entités oncologiques (30 et 100 classes)
- **CANTEMIST** (Miranda-Escalada *et al.*, 2020) : classification multi-étiquette de tumeurs
- **DisTEMIST** (Miranda-Escalada *et al.*, 2022) : classification multi-étiquette de maladies
- **MedDialog-FR** (Liu *et al.*, 2024) : classification multi-étiquette de dialogues médicaux (santé intime féminine)
- **EMEA** (Névéol *et al.*, 2014) : NER sur notices de médicaments
- **Medline** (Névéol *et al.*, 2014) : NER sur titres PubMed

Les six tâches de classification (monolabel ou multi-étiquettes) portent sur des documents cliniques longs (comptes rendus, cas patients, dialogues), pour lesquels un contexte de plus de 512 tokens est nécessaire. Les deux tâches NER ont des entrées plus courtes : Medline porte sur des titres d'articles PubMed, tandis qu'EMEA porte sur des notices de médicaments de longueur intermédiaire.

Chaque modèle est évalué avec 9 amorces aléatoires (*seeds*, 42 à 50) et nous rapportons la moyenne sur les seeds. Nous utilisons le F1 macro pour la classification monolabel (DiaMed), le F1 micro pour les tâches multi-étiquettes (FrACCO, CANTEMIST, DisTEMIST, MedDialog) et le F1 strict d'entité pour la NER (EMEA, Medline) en suivant le protocole de Bannour *et al.* (2024). L'affinage se fait avec une tête linéaire sur le jeton `[CLS]` (classification et multi-étiquettes) ou sur chaque position (NER), optimisée par AdamW avec arrêt précoce sur le jeu de validation. Pour les modèles à contexte 512 (CamemBERT, CamemBERT-bio, DrBERT), les documents dépassant cette limite sont tronqués ; aucun *chunking* ni fenêtre glissante n'est appliqué. Tous les résultats rapportés (y compris pour les modèles existants) sont produits avec ce même protocole et nos 9 seeds, et non repris des publications originales.

5.2 Modèles comparés

Nous comparons ModernCamemBERT-bio à deux encodeurs généralistes français pour mesurer l'apport de l'adaptation au domaine, à deux encodeurs biomédicaux français représentant les deux approches usuelles (pré-entraînement continu, pré-entraînement de zéro), et à un contrôle MLM qui isole l'apport spécifique du détour CLM :

- **CamemBERT** (Martin *et al.*, 2020), encodeur généraliste français de référence à contexte 512 tokens.
- **ModernCamemBERT** (Antoun *et al.*, 2025), encodeur généraliste français à contexte 8192 tokens dont nous partons pour le pré-entraînement continu.
- **CamemBERT-bio** (Touchent & de la Clergerie, 2024), obtenu par pré-entraînement continu de CamemBERT sur le corpus biomed-fr.
- **DrBERT** (Labrak *et al.*, 2023), pré-entraîné de zéro sur le corpus médical NACHOS plutôt que par adaptation d'un modèle existant.
- **MLM (contrôle)**, identique à ModernCamemBERT-bio mais sans le détour CLM : la phase 1 utilise un MLM standard à 30 % de masquage à la place de la phase causale.

5.3 Résultats

Le tableau 3 présente les résultats. ModernCamemBERT-bio atteint 61,6 % de F1 moyen et dépasse le contrôle MLM sur l'ensemble des 8 tâches (+2,8pp en moyenne ; détail des gains par tâche en annexe C). Le gain est concentré sur les tâches à documents longs (CANTEMIST +6,1pp, FrACCO-30 +4,9pp, DiaMed +4,0pp) et quasi-nul sur les deux tâches NER à entrée courte. Le gain par rapport à CamemBERT-bio atteint +23,3pp, et +11,5pp par rapport à DrBERT.

ModernCamemBERT-bio obtient les meilleures performances sur 6 des 8 tâches. CamemBERT-bio reste meilleur sur les deux tâches NER, mais l'écart est faible (1 à 3pp). La NER d'entités biomédicales se résout localement, en exploitant le contexte de quelques tokens autour de la mention, alors que la classification multi-étiquette de codes nécessite d'intégrer le document entier. Les encodeurs entraînés sur des séquences de 512 tokens spécialisent leurs représentations pour ces relations locales, ce qui

TABLE 3 – F1 moyen sur 8 tâches biomédicales et cliniques françaises (9 seeds, contexte 8192 pour les encodeurs longs, 512 sinon). En **gras**, le meilleur résultat par tâche. Abréviations : *ModernCBERT* = ModernCamemBERT, *CBERT-bio* = CamemBERT-bio, *CBERT* = CamemBERT.

		ModernCamemBERT-bio		Modèles existants			
Tâche	Type	Base	Large	ModernCBERT	CBERT-bio	CBERT	DrBERT
Classification / multi-étiquettes (documents longs)							
DiaMed	CLS	**67,4**	64,8	56,4	47,7	40,6	57,0
FrACCO-30	Multi	74,8	**80,7**	70,1	41,9	40,8	53,0
FrACCO-100	Multi	60,1	**65,4**	55,3	20,1	19,4	35,6
CANTEMIST	Multi	71,0	**74,4**	63,3	12,8	11,9	37,9
DisTEMIST	Multi	25,5	**30,4**	20,2	9,6	9,5	21,4
MedDialog	Multi	63,6	**64,5**	60,6	38,6	37,4	63,6
NER (entrées courtes ou intermédiaires)							
EMEA	NER	68,6	70,3	68,0	**70,8**	69,5	69,6
Medline	NER	61,9	63,1	59,7	**65,2**	62,7	62,8
Moyenne		61,6	**64,2**	56,7	38,3	36,5	50,1

les avantage sur la NER ; à l'inverse, un contexte de 8192 tokens est nécessaire pour les tâches à document long. Ce compromis se reflète dans nos résultats.

5.4 Passage à l'échelle (Large)

La recette est-elle robuste lorsqu'on augmente la taille du modèle, et l'avantage du détour CLM se maintient-il ? Nous entraînons également une version Large (350M) avec la même recette en deux phases (CLM puis phase MLM finale), à un budget de 50B+5B tokens (cf. §4.2). La version Large améliore la moyenne de $+2{,}6$pp par rapport à la version Base (64,2% contre 61,6% dans le tableau 3), avec les gains les plus importants sur FrACCO-30 ($+5{,}9$pp), FrACCO-100 ($+5{,}3$pp) et CANTEMIST ($+3{,}4$pp).

Pour isoler l'effet du détour CLM aux deux échelles, le tableau 4 compare nos modèles finaux à un contrôle MLM entraîné sur les mêmes données et au même budget, sans détour. L'avantage du détour CLM sur le contrôle MLM se maintient à l'échelle Large ($+1{,}2$pp contre $+2{,}8$pp en Base). La recette produit donc un gain aux deux échelles.

6 Discussion et conclusion

ModernCamemBERT-bio atteint l'état de l'art sur les 8 tâches biomédicales et cliniques françaises évaluées, avec les gains les plus importants sur les tâches cliniques à documents longs où les encodeurs à 512 tokens sont structurellement pénalisés. L'essentiel du gain par rapport à CamemBERT-bio s'explique par le contexte long : le passage à ModernCamemBERT généraliste apporte déjà $+18{,}4$pp, et l'adaptation biomédicale sur biomed-fr-v2 ajoute $+4{,}9$pp.

TABLE 4 – Comparaison du détour CLM (MC-bio) au contrôle MLM, qui utilise les mêmes données et le même budget sans phase CLM, pour les deux tailles. En **gras**, le meilleur des deux par taille (CLM vs MLM-ctrl). F1 moyen, 9 seeds, 8 tâches. L'écart moyen est de +2,8pp pour Base (IC 95% [+1,0, +4,6], paired t-test sur les 8 tâches, $p < 0{,}01$) et de +1,2pp pour Large (IC 95% [+0,1, +2,3], $p < 0{,}05$).

	Base (149M)		Large (350M)	
Tâche	MC-bio	MLM-ctrl.	MC-bio	MLM-ctrl.
DiaMed	**67,4**	63,4	**64,8**	61,2
FrACCO-30	**74,8**	69,9	**80,7**	79,4
FrACCO-100	**60,1**	56,8	**65,4**	63,3
CANTEMIST	**71,0**	64,9	**74,4**	72,6
DisTEMIST	**25,5**	23,5	**30,4**	29,1
MedDialog	**63,6**	62,5	**64,5**	**64,5**
EMEA	**68,6**	68,5	70,3	**70,4**
Medline	**61,9**	61,4	63,1	**63,5**
Moyenne	**61,6**	58,9	**64,2**	63,0

Curation de données biomédicales. Le choix des signaux de qualité n'est pas neutre : edu et cont apportent un gain (+0,67pp en combiné), tandis que writ et term dégradent les performances en éliminant des documents cliniques informels mais informatifs. La composition du corpus a un impact comparable : exclure `drug_information`, redondant avec EMEA, apporte +0,67pp.

Détour CLM. Le détour par un entraînement causal apporte +2,8pp en Base par rapport au MLM standard, avec les mêmes données et le même budget. Le gain est concentré sur les tâches à documents longs et se réduit à +1,2pp à l'échelle Large, ce qui prolonge au pré-entraînement continu de domaine les observations de Gisserot-Boukhlef *et al.* (2025) sur le pré-entraînement de zéro en anglais. Une analyse mécaniste du détour est proposée séparément (Touchent & de la Clergerie, 2026a).

Empreinte environnementale. En appliquant la méthodologie de Lacoste *et al.* (2019) reprise par Touchent & de la Clergerie (2024), le pré-entraînement de la version Base émet 0,46 kg CO_2eq, soit moins que CamemBERT-bio (0,80 kg) et 57 fois moins que DrBERT (26,11 kg). La version Large, à budget plus élevé, émet 9,25 kg, soit comparable à AliBERT (8,16 kg) et 2,8 fois moins que DrBERT. Le détail du calcul figure en annexe A.

Limites. Notre évaluation porte sur 8 tâches publiques, dont deux tâches NER à entrée courte sur lesquelles CamemBERT-bio reste meilleur. Une évaluation sur des documents cliniques hospitaliers réels reste à conduire. Le corpus biomed-fr-v2 n'est pas publié dans sa forme complète en raison de contraintes de licence sur certaines sources ; nous publions le modèle et la description du pipeline de curation.

Déploiement. Nous publions les modèles sous licence libre pour le traitement automatique du langage clinique en français. Les grands modèles génératifs ont montré des performances remar-

quables sur les tâches biomédicales, mais leur utilisation passe souvent par des serveurs distants peu compatibles avec les contraintes de confidentialité ou de puissance de calcul des établissements de santé. Un encodeur bidirectionnel spécialisé, suffisamment compact pour tourner sur le matériel d'un établissement de santé et avec un contexte long pour absorber un compte rendu d'hospitalisation entier sans troncature, reste un objet utile en complément des grands modèles génératifs.

Remerciements

Ce travail a bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2025-AD011014393R2 attribuée par GENCI.

Références

ANTOUN W., SAGOT B. & SEDDAH D. (2025). ModernBERT or DeBERTaV3 ? Examining architecture and data influence on transformer encoder models performance. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, p. 3061–3074, Mumbai, India : The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. DOI : 10.18653/v1/2025.ijcnlp-long.164.

BANNOUR N., SERVAN C., NÉVÉOL A. & TANNIER X. (2024). A benchmark evaluation of clinical named entity recognition in French. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 14–21, Torino, Italy.

BERHE A., DRAZNIEKS G., MARTENOT V., MASDEU V., DAVY L. & ZUCKER J.-D. (2023). AliBERT : A pre-trained language model for French biomedical text. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, p. 223–236, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.bionlp-1.19.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 8440–8451, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.747.

DAO T., FU D. Y., ERMON S., RUDRA A. & RÉ C. (2022). FlashAttention : Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, p. 16344–16359.

ELECTRICITY MAPS (2025). France – carbon intensity (december 2025). https://app.electricitymaps.com/zone/FR.

GISSEROT-BOUKHLEF H., BOIZARD N., FAYSSE M., ALVES D. M., MALHERBE E., MARTINS A. F. T., HUDELOT C. & COLOMBO P. (2025). Should we still pretrain encoders with masked language modeling ? arXiv preprint arXiv :2507.00994. DOI : 10.48550/arXiv.2507.00994.

GRATTAFIORI A., DUBEY A., JAUHRI A., PANDEY A., KADIAN A., AL-DAHLE A., LETMAN A., MATHUR A., SCHELTEN A., VAUGHAN A. *et al.* (2024). The Llama 3 herd of models. *arXiv preprint arXiv :2407.21783*.

LABRAK Y., BAZOGE A., DUFOUR R., ROUVIER M., MORIN E., DAILLE B. & GOURRAUD P.-A. (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 (Volume 1 : Long Papers)*, p. 16207–16221, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.896.

LABRAK Y., BAZOGE A., KHETTARI O. E., ROUVIER M., BEAUFILS P. C. D., GRABAR N., DAILLE B., QUINIOU S., MORIN E., GOURRAUD P.-A. & DUFOUR R. (2024). DrBenchmark : A large language understanding evaluation benchmark for French biomedical domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 5376–5390, Torino, Italia : ELRA and ICCL.

LACOSTE A., LUCCIONI A., SCHMIDT V. & DANDRES T. (2019). Quantifying the carbon emissions of machine learning. arXiv preprint arXiv :1910.09700.

LIU X., SEGONNE V., MANNION A., SCHWAB D., GOEURIOT L. & PORTET F. (2024). MedDialog-FR : A French version of the MedDialog corpus for multi-label classification and response generation related to women's intimate health. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, p. 173–183, Torino, Italia : ELRA and ICCL.

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 preprint arXiv :1907.11692*.

MAGNINI B., ALTUNA B., LAVELLI A., SPERANZA M. & ZANOLI R. (2020). The E3C project : Collection and annotation of a multilingual corpus of clinical cases. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, p. 190–196, Bologna, Italy : CEUR Workshop Proceedings.

MARTIN L., MULLER B., SUÁREZ P. J. O., 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 Association for Computational Linguistics*, p. 7203–7219, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

MIRANDA-ESCALADA A., FARRÉ-MADUELL E. & KRALLINGER M. (2020). Named entity recognition, concept normalization and clinical coding : Overview of the CANTEMIST track for cancer text mining in Spanish, corpus, guidelines, methods and results. In *IberLEF 2020, CEUR Workshop Proceedings*, p. 303–323 : CEUR-WS.org. DOI : 10.5281/zenodo.3773228.

MIRANDA-ESCALADA A., GASCÓ L., LIMA-LÓPEZ S., FARRÉ-MADUELL E., ESTRADA D., NENTIDIS A., KRITHARA A., KATSIMPRAS G., PALIOURAS G. & KRALLINGER M. (2022). Overview of DisTEMIST at BioASQ : Automatic detection and normalization of diseases from clinical texts : Results, methods, evaluation and multilingual resources. In *Proceedings of the Working Notes of CLEF 2022*, p. 179–203 : CEUR-WS.org.

NÉVÉOL A., GROUIN C., LEIXA J., ROSSET S. & ZWEIGENBAUM P. (2014). The QUAERO French medical corpus : A ressource for medical entity recognition and normalization. In *Proceedings of BioTxtM 2014*, p. 24–30.

PENEDO G., KYDLÍČEK H., ALLAL L. B., LOZHKOV A., MITCHELL M., RAFFEL C., VON WERRA L. & WOLF T. (2024). The FineWeb Datasets : Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, p. 30811–30849. DOI : 10.52202/079017-0970.

PIGNAT J., VUCETIC M., GAUDET-BLAVIGNAC C., ZAGHIR J., STETTLER A., AMREIN F., BONJOUR J., GOLDMAN J.-P., MICHIELIN O., LOVIS C. & BJELOGRLIC M. (2025). FRACCO :

A gold-standard annotated corpus of oncological entities with ICD-O-3.1 normalisation. *arXiv preprint arXiv :2510.13873*. DOI : 10.48550/arXiv.2510.13873.

RAE J. W., BORGEAUD S., CAI T., MILLICAN K., HOFFMANN J., SONG F., ASLANIDES J., HENDERSON S., RING R., YOUNG S. *et al.* (2021). Scaling language models : Methods, analysis & insights from training Gopher. *arXiv preprint arXiv :2112.11446*. DOI : 10.48550/arXiv.2112.11446.

SOUNACK T., DAVIS J., DURIEUX B., CHAFFIN A., POLLARD T. J., LEHMAN E., JOHNSON A. E. W., MCDERMOTT M., NAUMANN T. & LINDVALL C. (2025). BioClinical ModernBERT : A state-of-the-art long-context encoder for biomedical and clinical NLP. arXiv preprint arXiv :2506.10896. DOI : 10.48550/arXiv.2506.10896.

SU J., AHMED M., LU Y., PAN S., BO W. & LIU Y. (2024). RoFormer : Enhanced transformer with rotary position embedding. *Neurocomputing*, **568**, 127063. DOI : 10.1016/j.neucom.2023.127063.

THE MOSAIC ML TEAM (2021). composer. https://github.com/mosaicml/composer.

TOUCHENT R. & DE LA CLERGERIE É. (2024). CamemBERT-bio : Leveraging continual pretraining for cost-effective models on French biomedical data. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 2692–2701, Torino, Italia : ELRA and ICCL.

TOUCHENT R. & DE LA CLERGERIE É. (2026a). A causal language modeling detour improves encoder continued pretraining.

TOUCHENT R. & DE LA CLERGERIE É. (2026b). OntoBook : Ontology-grounded synthetic textbooks for medical encoder pretraining. In G. SÉRASSET, K. GKIRTZOU, M. COCHEZ & J.-C. KALO, Éds., *Proceedings of Knowledge Graphs and Large Language Models Workshop 2026 (KG-LLM) @ LREC 2026*, p. 11–19, Palma de Mallorca, Spain : ELRA Language Resources Association.

TOUCHENT R., GODEY N. & DE LA CLERGERIE É. (2025). Biomed-Enriched : A biomedical dataset enriched with LLMs for pretraining and extracting rare and hidden content. arXiv preprint arXiv :2506.20331. DOI : 10.48550/arXiv.2506.20331.

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

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

A Empreinte carbone : détail du calcul

Nous appliquons le calculateur d'impact ML de Lacoste *et al.* (2019). L'énergie consommée est calculée comme GPU-h × TDP × PUE et l'empreinte comme énergie × intensité carbone du réseau. ModernCamemBERT-bio Base et Large ont été pré-entraînés en décembre 2025 sur Jean Zay (4×H100 SXM5, TDP de 700 W, PUE de 1,2), avec une intensité carbone de 34 g CO_2eq/kWh relevée pour la France en décembre 2025 (Electricity Maps, 2025).

TABLE 5 – Empreinte carbone du pré-entraînement (PUE 1,2, intensité carbone 34 g CO_2eq/kWh). Les chiffres des trois premiers modèles sont repris de Touchent & de la Clergerie (2024).

Modèle	Durée	Matériel	GPU-h	kg CO_2eq
DrBERT (Labrak *et al.*, 2023)	20 h	128×V100	2 560	26,11
AliBERT (Berhe *et al.*, 2023)	20 h	48×A100	960	8,16
CamemBERT-bio (Touchent & de la Clergerie, 2024)	39 h	2×V100	78	0,80
ModernCamemBERT-bio Base (149M)	4 h	4×H100	16	0,46
ModernCamemBERT-bio Large (350M)	81 h	4×H100	324	9,25

B Statistiques de longueur des jeux d'évaluation

Le tableau 6 donne, pour chaque jeu de test, la distribution de longueur en tokens ModernCamemBERT (sans tokens spéciaux), à l'échelle du document. Toutes les tâches de classification et multi-étiquettes opèrent sur des documents longs : DiaMed, FrACCO, CANTEMIST, DisTEMIST et MedDialog ont leurs documents principalement au-dessus de 512 tokens. Seules les deux tâches NER (EMEA sur des notices de médicaments à longueur intermédiaire, et Medline sur des titres d'articles) restent dans des régimes courts ou modérés.

TABLE 6 – Distribution des longueurs en tokens ModernCamemBERT des documents des jeux de test. % >512 indique la proportion de documents qui ne tiennent pas dans le contexte d'un encodeur 512 tokens. CANTEMIST et DisTEMIST sont des corpus de comptes rendus oncologiques de profil similaire à FrACCO.

Tâche	N	moy.	méd.	p95	max	% >512
DiaMed	154	523	496	888	1654	46,1 %
MedDialog	297	333	298	609	1236	11,4 %
EMEA (NER)	15	1 062	1 067	1 257	1 309	100,0 %
Medline (NER)	833	21	19	39	78	0,0 %
FrACCO	130	∼1 050	∼1 030	∼2 110	∼2 800	100 %

C Gains du détour CLM par tâche

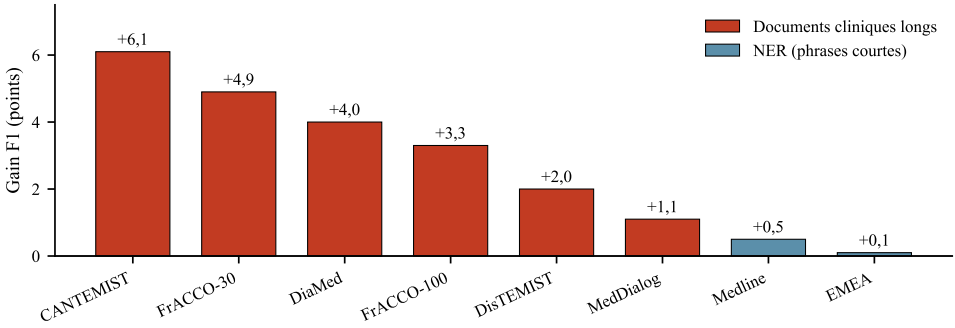

FIGURE 2 – Gain du détour CLM sur le contrôle MLM, par tâche (Base 149M, 9 seeds).

D Coefficients de suréchantillonnage

TABLE 7 – Coefficients de suréchantillonnage par bucket de qualité d'article.

Ratio de paragraphes edu≥7 ET cont≥7	Coefficient
0%	exclu
1–25%	4,3×
26–50%	8,5×
51–75%	12,8×
76–99%	21,3×
100%	34,1×

E Fidélité de l'extraction

E.1 Distribution du recouvrement caractère

Le tableau 8 donne la distribution complète du recouvrement caractère (après normalisation typographique : minuscules, suppression des diacritiques et de la ponctuation, compression des espaces) entre les paragraphes extraits et leur document source, sur 300 paragraphes tirés aléatoirement de la partition ISTEX.

TABLE 8 – Recouvrement caractère entre paragraphe extrait et document source (300 paragraphes ISTEX).

Seuil de recouvrement	% de paragraphes
$\geq 99\%$ (extraction presque verbatim)	84,1
$\geq 95\%$	92,7
$\geq 90\%$	94,4
$\geq 80\%$	96,6
$\geq 50\%$ (aucun cas en dessous)	100,0

E.2 Exemple 1 : correction d'océrisation

Document source (OCR) :

```
... vise ~ remonter l&apos;organe prolab6 (correction de
la pt6se) et h le soutenir dans sa position id6ale plus
qu&apos;/~ le fixer (fixation). Ce sout~nement est assur6 par le
renforcement du p6rin6e grace aux tissus naturels que l&apos;on
r6pare (rapphie) ou que l&apos;on consolide avec du mat6riel
local (plastie, ligamentopexie)...
```

Paragraphe extrait :

> ... la chirurgie vise à remonter l'organe prolapsé (correction de la ptose) et à le soutenir dans sa position idéale plus qu'à le fixer (fixation). Ce soutien est assuré par le renforcement du périnée grâce aux tissus naturels que l'on répare (raphie) ou que l'on consolide avec du matériel local (plastie, ligamentopexie)...

Les mots présents dans le paragraphe extrait et absents du source au sens lexical (*périnée, ptose, matériel, idéale, répare, assuré*, etc.) correspondent à des corrections d'océrisation où des chiffres ont été substitués à des lettres accentuées (*p6rin6e* → *périnée*, *pt6se* → *ptose*, *mat6riel* → *matériel*). Tous les termes médicaux et la structure du paragraphe sont préservés.

E.3 Exemple 2 : extraction verbatim

Document source et paragraphe extrait (identiques) :

> La dépression en cancérologie renvoie à la dimension de la souffrance psychique et de la détresse psychologique. Paradoxalement, elle reste encore trop souvent sous-estimée et banalisée car considérée comme réactionnelle au cancer. Pourtant, son retentissement sur la qualité de vie des patients et son incidence tant sur le plan médicoéconomique que sur la prise en charge oncologique sont loin d'être négligeables...

Sur ce paragraphe de 1 143 caractères, le recouvrement est de 100% et aucun mot supplémentaire n'est introduit.