L'utilisation de corpus textuels pour l'entraînement des modèles de langage

Rian Touchent, Eric de la Clergerie

Inria

CamemBERT-bi o

Une adaptation biomédical
publique de CamemBERT

Le cas présenté concerne un homme âgé de 61 ans (71 kg, 172 cm, soit un indice de masse corporelle de 23,9 kg/m²) admissible à une transplantation pulmonaire en raison d'une insuffisance respiratoire chronique terminale sur emphysème post-tabagique, sous oxygénothérapie continue (1 L/min) et ventilation non invasive nocturne. Il présente, comme principaux antécédents, une dyslipidémie, une hypertension artérielle et un tabagisme sevré estimé à 21 paquets-années (facteurs de risque cardiovasculaires).

+

Maladie, Symptôme, Âge, Poids, Taille, Substance, Traitement

↓

Le cas présenté concerne un homme âgé de 61 ans ÂGE (71 kg TAILLE , 172 cm TAILLE , soit un indice de masse corporelle de 23,9 kg/m²) admissible à une transplantation pulmonaire TRAITEMENT en raison d'une insuffisance respiratoire chronique terminale MALADIE sur emphysème post-tabagique MALADIE , sous oxygénothérapie continue TRAITEMENT (1 L/min) et ventilation non invasive nocturne TRAITEMENT . Il présente, comme principaux antécédents, une dyslipidémie MALADIE , une hypertension artérielle MALADIE et un tabagisme sevré estimé à 21 paquets-années (facteurs de risque cardiovasculaires).

Reconnaissance d'entités nommées

Models	**DEFT**			
	Precision	*Recall*	*F-measure*	*CO_2 eq (g.)*
CamemBERT	0.73 [0.71-0.75]	0.75 [0.73-0.77]	0.74 [0.73-0.76]	7.7
FlauBERT	0.73 [0.72-0.76]	0.75 [0.73-0.77]	0.74 [0.73-0.76]	8
mBERT	0.72 [0.70-0.74]	0.74 [0.72-0.76]	0.73 [0.71-0.75]	8.3
frALBERT	0.68 [0.66-0.69]	0.67 [0.65-0.69]	0.67 [0.66-0.69]	4.5
CamemBERT-bio	**0.75** [0.73-0.77]	**0.77** [0.75-0.78]	**0.76** [0.74-0.78]	8.7
DrBERT	0.71 [0.68-0.73]	0.72 [0.70-0.74]	0.71 [0.69-0.73]	5.5
Baseline	0.38	0.32	0.35	-
Copara et al. (2020)	-	-	0.73	-

Table 3: Performance of nested entity extraction on the DEFT test set.

A Benchmark Evaluation of Clinical Named Entity Recognition in French
Bannour et al. (2024)

Désambiguïsation lexicale

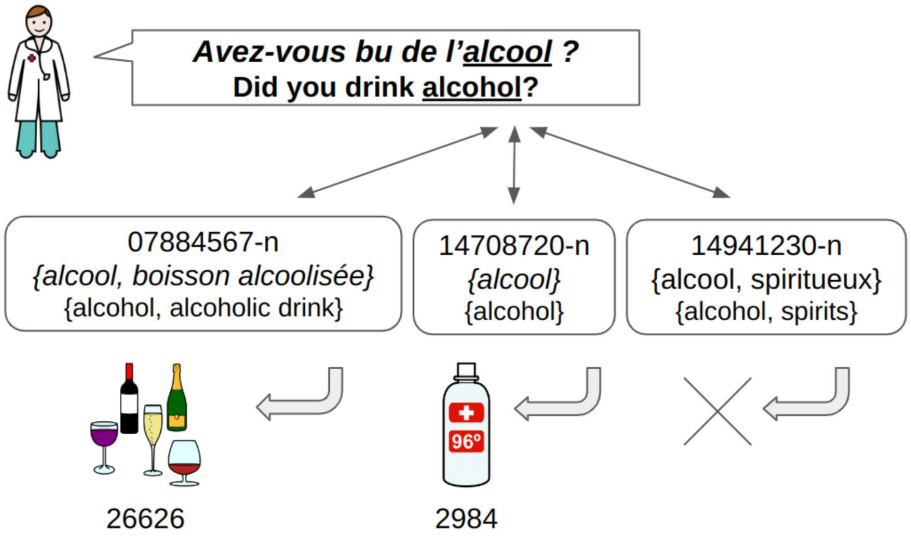


Figure 1: Pictographs for the word to be disambiguated: "*alcool*" (alcohol). Ids are indicated for the WOLF synsets and the Arasaac pictographs.

Word Sense Disambiguation for Automatic Translation of Medical Dialogues
Norré et al. (2023)

Classification de texte

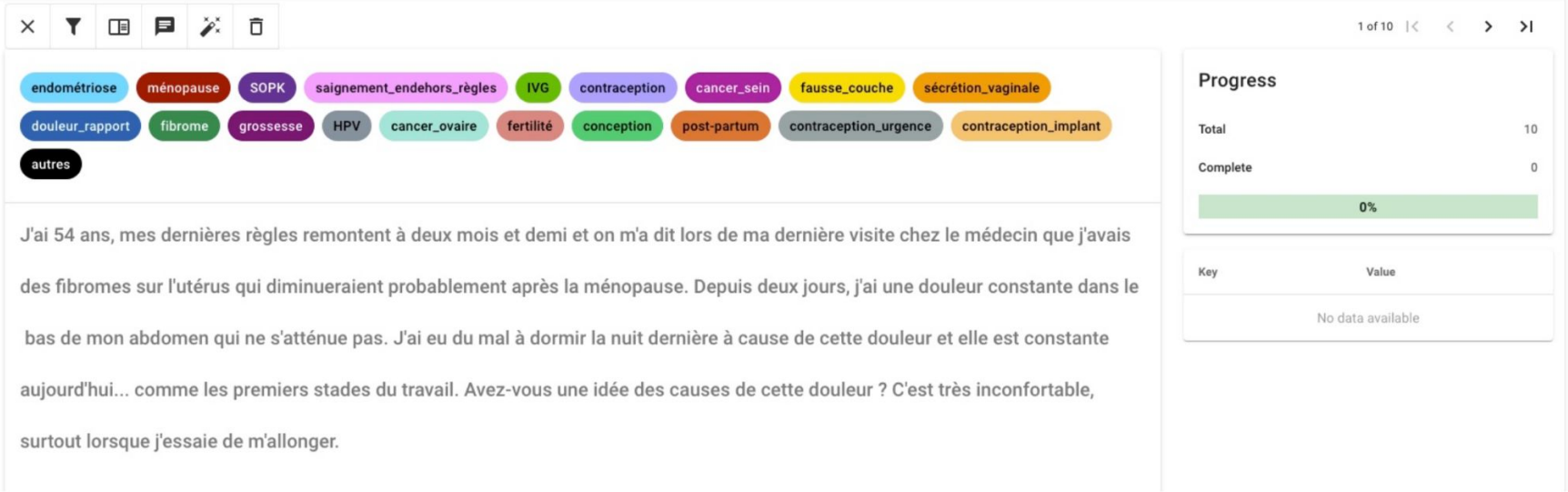


Figure 2: Multi-label Annotation Interface with *doccano*

MedDialog-FR: a French Version of the MedDialog Corpus for Multi-label Classification and Response Generation related to Women's Intimate Health

Liu et al. (2024)

Classification de texte

Model	macro			weighted
	P	**R**	**F1**	**F1**
FlauBERT-base	0.36	0.43	**0.38**	0.59
FlauBERT-base-weighted	0.41	0.36	0.37	0.54
CamemBERT-base	0.23	0.33	0.26	0.58
CamemBERT-base-weighted	0.38	0.29	0.32	0.53
CamemBERT-bio-base	0.33	0.40	0.35	0.59
CamemBERT-bio-base-weighted	0.45	0.44	<u>0.42</u>	0.63
DrBERT-4gb	0.45	0.29	0.33	0.50
DrBERT-4gb-weighted	0.40	0.31	0.31	0.46

Table 5: Model performance with on the MedDialog-FR-women test set containing 22 labels

MedDialog-FR: a French Version of the MedDialog Corpus for Multi-label Classification and Response Generation related to Women's Intimate Health

Liu et al. (2024)

Extraction et interprétation des chiffres médicaux

Input medical note:

"Heterotaxie avec isomerisme gauche. Écho cardiaque (14/08): gradient VD-VG AP de 50-60mmHg. en attente de Chx $\rightarrow$ dérivation cavo-pulmonaire. Suivi par Dr. F. saturation habituelle $80 - 85\%$ *Polysplénie Malrotation intestinale opéré."*

Expected output:

Value	Attributes	Unit	Critical
$14/08$	Divers (date)		No
$50 - 60$	gradient VD-VG AP	**mmHg**	No
$80 - 85$	saturation en oxygène	%	No

Are Medium-Sized Transformers Models still Relevant for Medical Records Processing?
Lompo & Le (2024)

Extraction et interprétation des chiffres médicaux

“Results: Our findings demonstrate **significant performance** improvements, **with CamemBERT-bio** achieving an F1 score of 0.89, [...] **only 0.06 units lower than GPT-4** “

“Impact Statement:**The application of CamemBERT-bio** to numerical values classification **represents a promising avenue**, especially **regarding limited clinical data availability**. “

Are Medium-Sized Transformers Models still Relevant for Medical Records Processing?
Lompo & Le (2024)

Détection de la négation

3. À noter que il [n'] est inséré sur le mur en postérieur [que] par un pont filiforme.
4. L' artère marginale [n'] est [pas] décelable [que] en aval du pontage.
5. L'[absence d'] anomalie au niveau de l'aorte thoracique ascendante.
6. [Pas] d'anomalie du pancréas, et des surrénales.

Automatic analysis of negation cues and scopes for medical texts in French using language models
Sadoune et al. (2024)

Détection de la négation

Scen	Models	TRAIN	TEST	Cue detection			Scope detection		
				P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
1	Neg-Radio-CamemBERT-bio-base	RADIO	ESSAI+CAS	92.0	93.7	92.9	66.7	80.6	73.0
2	CamemBERT-base	ESSAI+CAS	RADIO	98.6	99.3	98.9	84.7	80.1	82.3
	CamemBERT-bio-base			**98.8**	**99.4**	**99.0**	**85.5**	**80.3**	**82.8**
	DrBERT			98.3	99.2	98.8	80.4	77.6	79.0
3	CamemBERT-base	RADIO+ESSAI	CAS	95.0	94.6	94.8	83.3	**85.8**	84.5
	CamemBERT-bio-base			**95.8**	**94.7**	**95.3**	**84.7**	85.3	**85.0**
	DrBERT			94.5	93.2	93.8	77.1	78.8	78.0
4	CamemBERT-base	ESSAI	CAS	94.1	**95.8**	**95.0**	85.8	**85.5**	85.6
	CamemBERT-bio-base			**94.5**	95.4	94.9	**87.4**	85.4	**86.4**
	DrBERT			93.6	94.2	93.9	77.5	80.3	78.8
5	CamemBERT-base	NLMFC	(10%) NLMFC	**97.76 ± 0.54**	98.64 ± 0.38	98.19 ± 0.31	90.45 ± 1.24	91.31 ± 1.30	90.88 ± 1.13
	CamemBERT-bio-base			97.71 ± 0.79	**98.75 ± 0.40**	**98.22 ± 0.39**	**90.69 ± 0.92**	**91.80 ± 1.40**	**91.24 ± 0.97**
	DrBERT			97.41 ± 0.82	98.40 ± 0.61	97.90 ± 0.51	87.55 ± 1.62	88.70 ± 1.96	88.11 ± 1.47

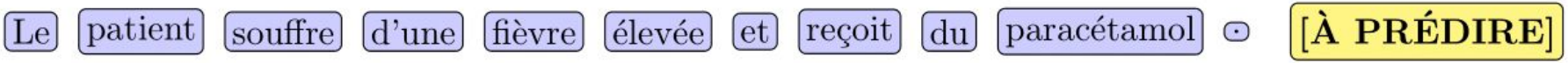
Table 6

Performance metrics (Precision, Recall, and F1-score) obtained with the four scenarios described in Section 3.4. The best models are shown in bold. Scenario 1 to 4 evaluate tokenization on an external dataset, so that we did not use cross-validation but present the results of the model fine-tuned on the complete training set. Thus the absence of indication of variance in these cases. For scenario 5, results are given as average±standard-deviation over the 10-fold cross-validation.

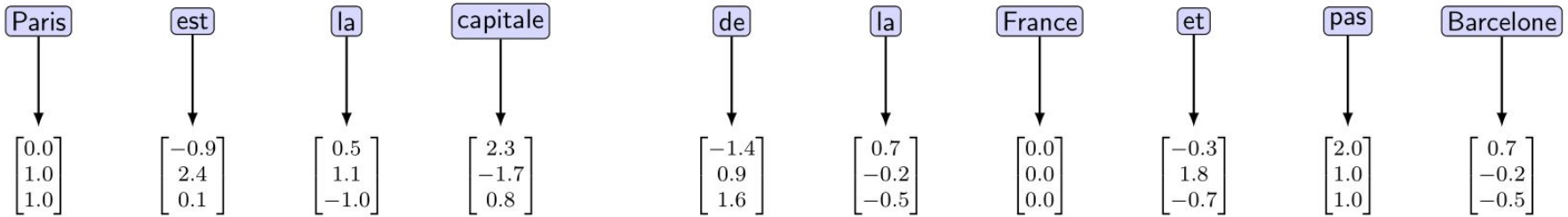
Automatic analysis of negation cues and scopes for medical texts in French using language models
Sadoune et al. (2024)

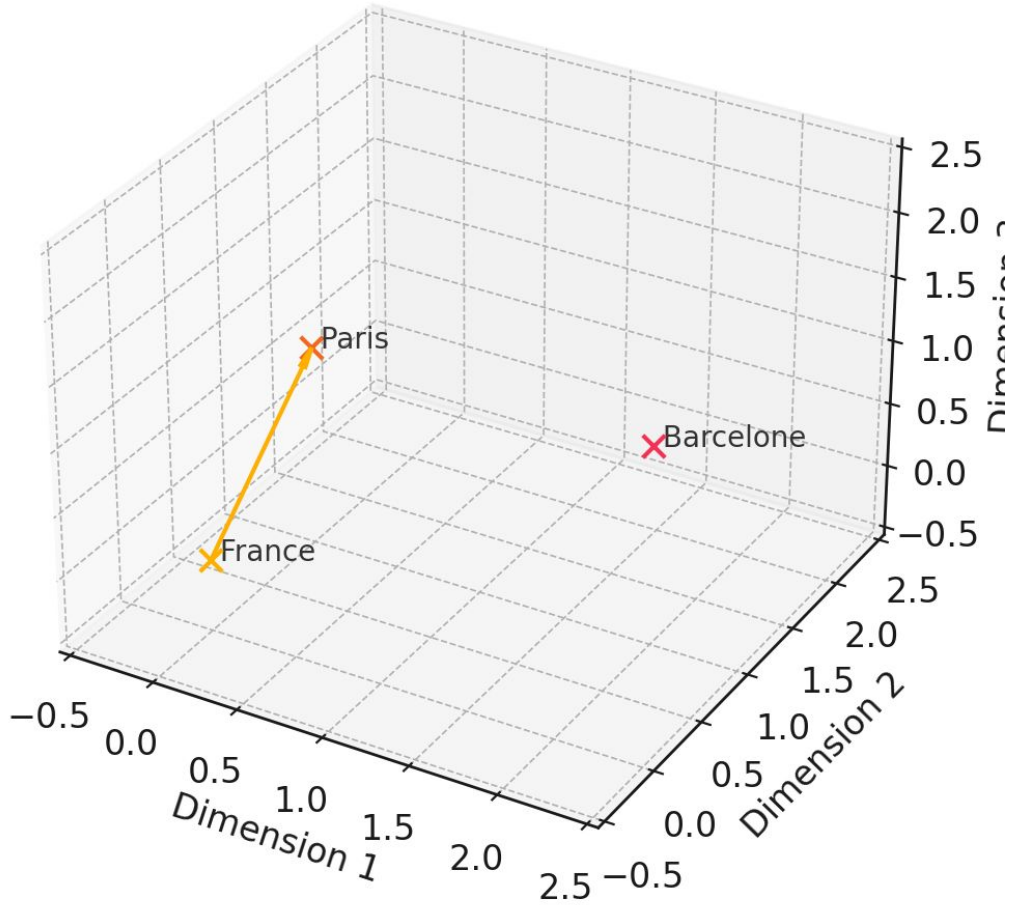
Fonctionnement de CamemBERT-bio

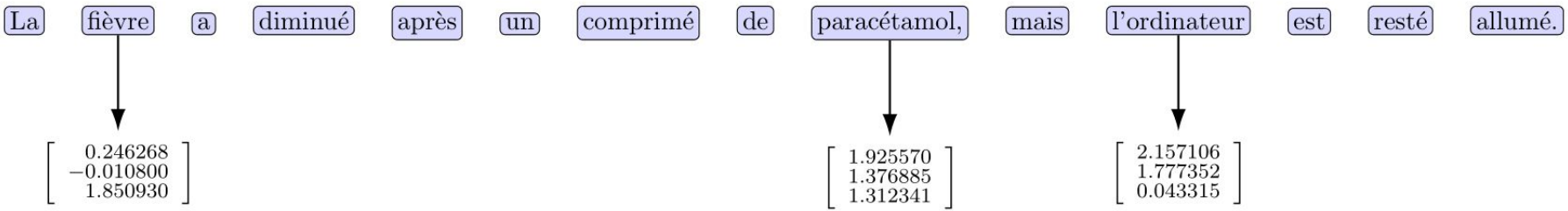
Décodeurs (ChatGPT)

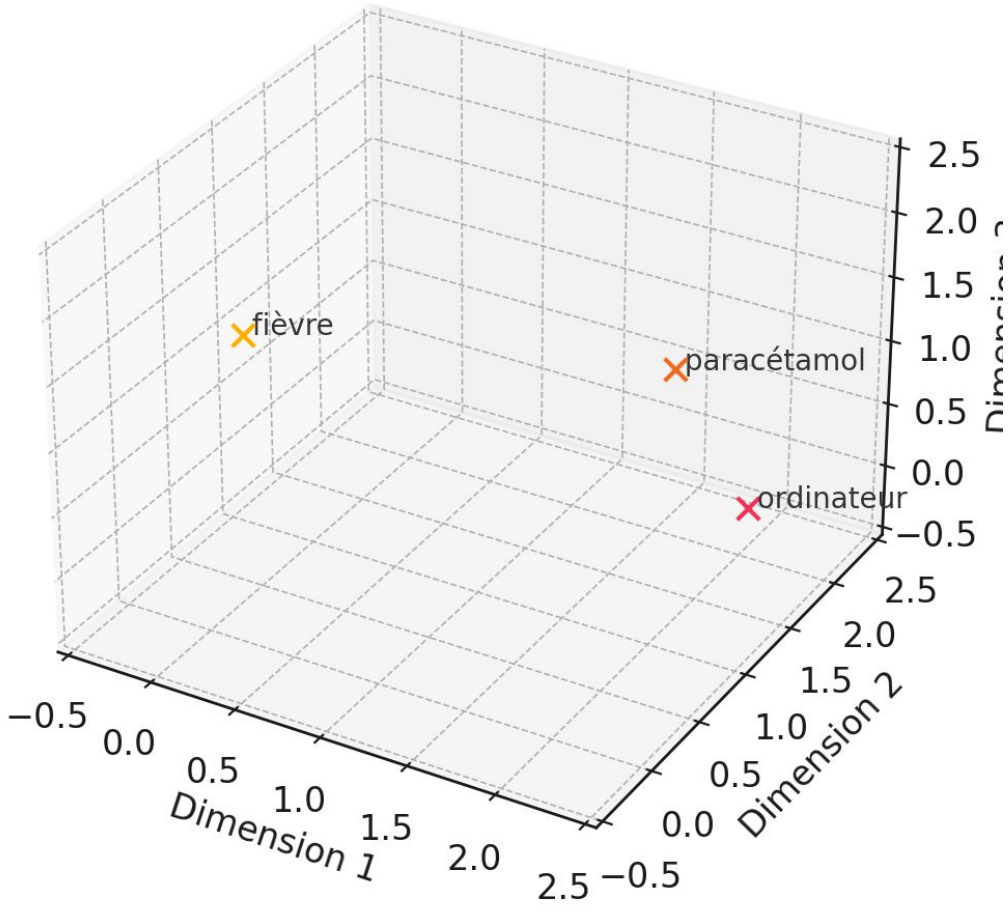


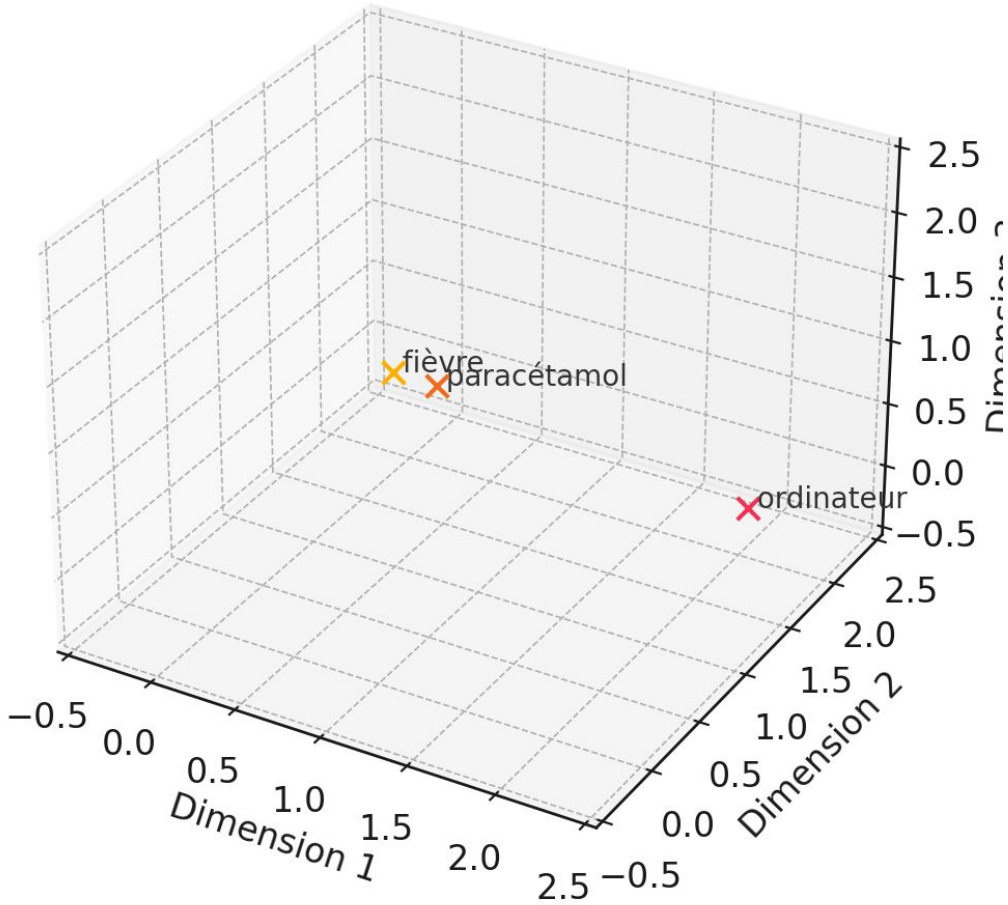
Encodeurs (CamemBERT-bio)




Paris
Barcelone
France
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
Dimension 1
Dimension 2


La
fièvre
a
diminué
après
un
comprimé
de
paracétamol,
mais
l'ordinateur
est
resté
allumé.
0.246268
−0.010800
1.850930
1.925570
1.376885
1.312341
2.157106
1.777352
0.043315


fièvre
paracétamol
ordinateur
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
Dimension 1
Dimension 2


fièvre
paracétamol
ordinateur
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
Dimension 1
Dimension 2

Construction de CamemBERT-bio

Le patient souffre d'une *fièvre élevée* et reçoit du *paracétamol* pour soulager la *douleur*. Une perfusion de sérum physiologique est également en place ...

$\longrightarrow$ **[À PRÉDIRE]**

Le patient présente une fièvre élevée (>39 °C) et une [**MASKED**] persistante.

Une dose orale de [**MASKED**]

a été administrée pour réduire la douleur, associée à une perfusion de [**MASKED**].

Constitution d'un corpus: *biomed-fr*

Corpus	Détails	Taille
ISTEX	Divers documents de la littérature scientifique indexés sur ISTEX	276 M
CLEAR	Notices de médicaments	73 M
E3C	Divers documents issus de journaux, de notices et de cas cliniques	64 M
Total		413 M

TABLE 1 – Composition du corpus biomed-fr (en millions de mots)

Autres sources considérées pour de futures versions :

- Articles biomédicaux libres sur HAL ou PudMed
- Wikipedia
- …

biomed-fr-small :

- 10% de biomed-fr sélectionnée aléatoirement

Utilisation de ISTEX

- **Corpus extrait :**
  108 183 documents en français, publiés dans des revues de biologie ou de médecine **depuis 1990**.

- **Qualité des articles avant 1990 :**
  Nombreuses **erreurs typographiques**.
  Articles souvent **scannés**, nécessitant une **reconnaissance optique de caractères (OCR)**.

- **Articles après 1990 :**
  - Moins d'erreurs typographiques.
  - Meilleure qualité globale.

- **Langue :**
  - Principalement en **français**.
  - Certains passages en **anglais**, mais en **quantité indéterminée**.

- **Impact sur l'entraînement :**
  - Présence d’anglais peu susceptible d’avoir un **impact significatif** sur le **pré-entraînement**.

Evaluation : NER

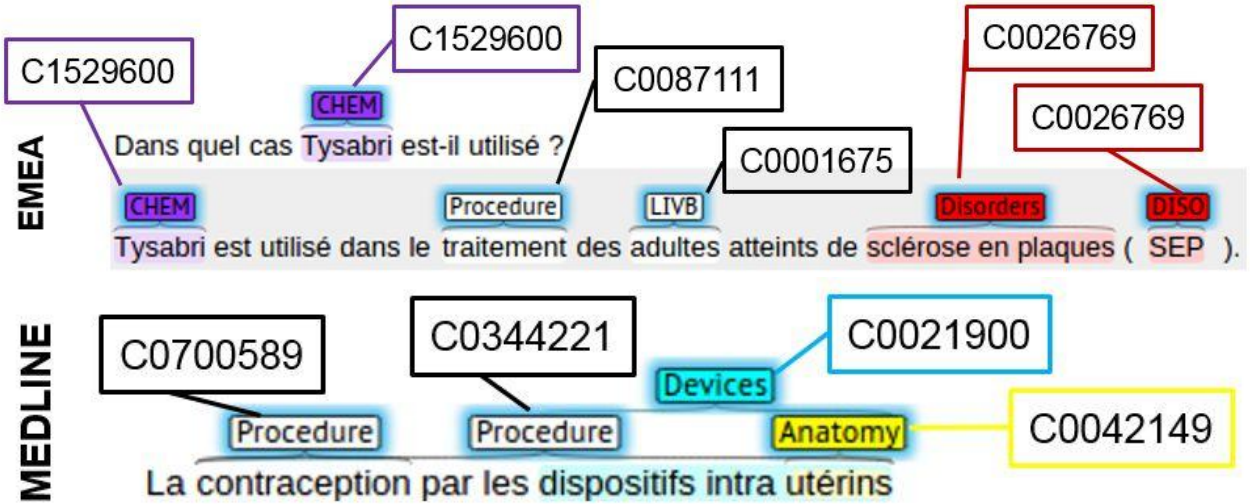
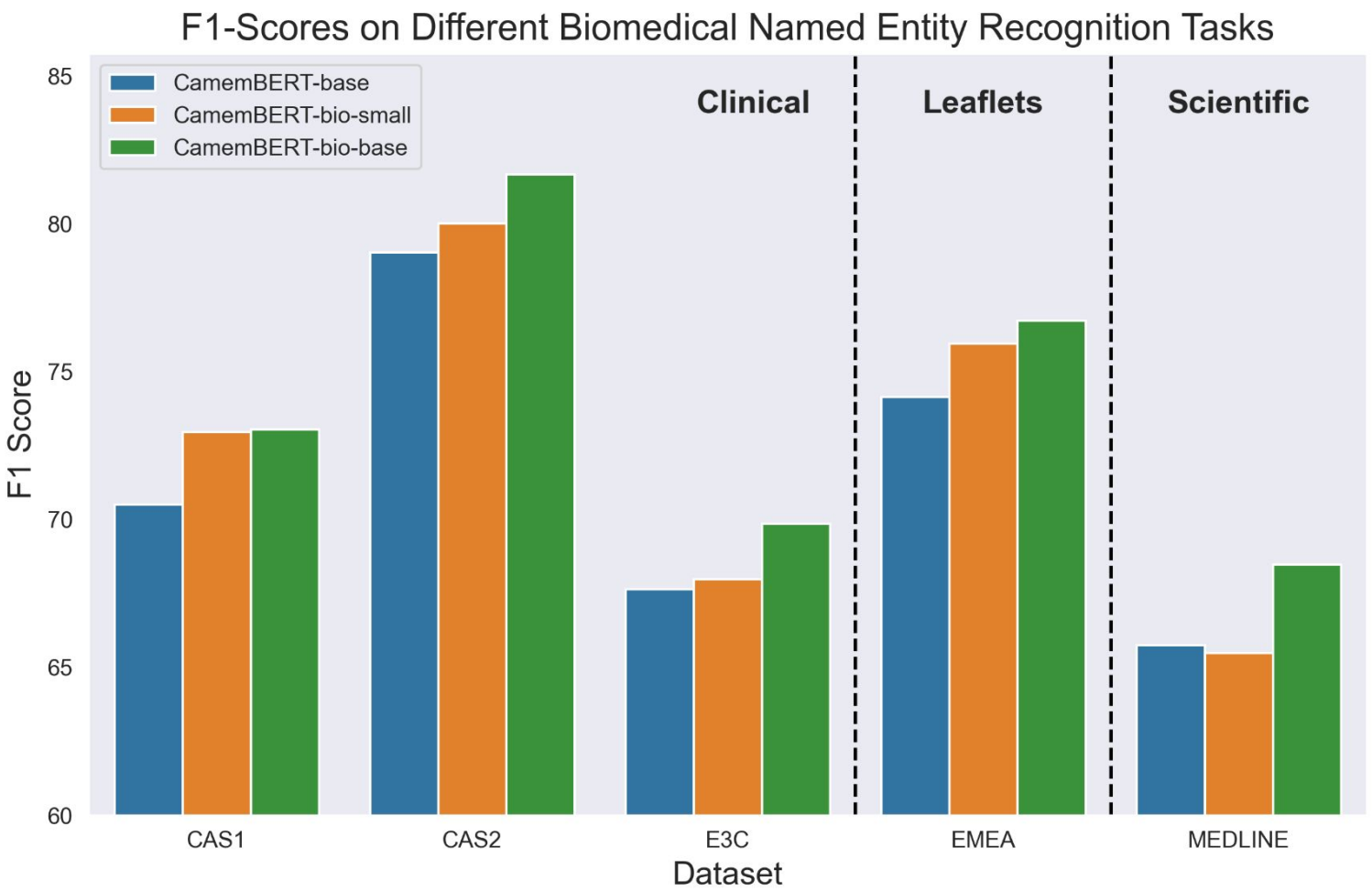
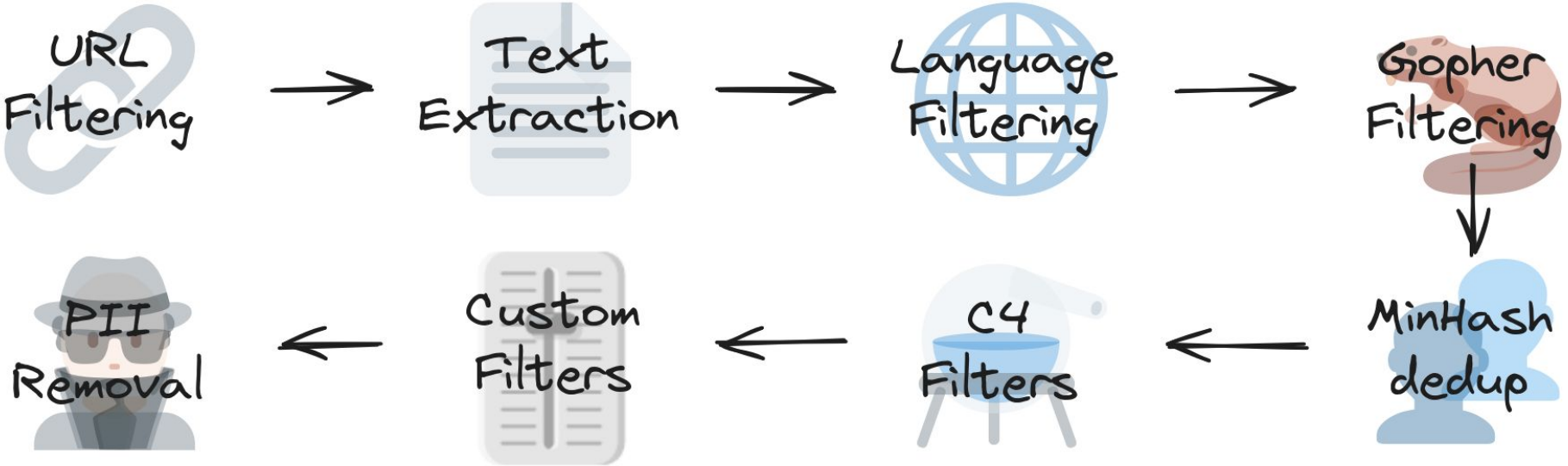


Fig.2 : Examples issues de EMEA et MEDLINE

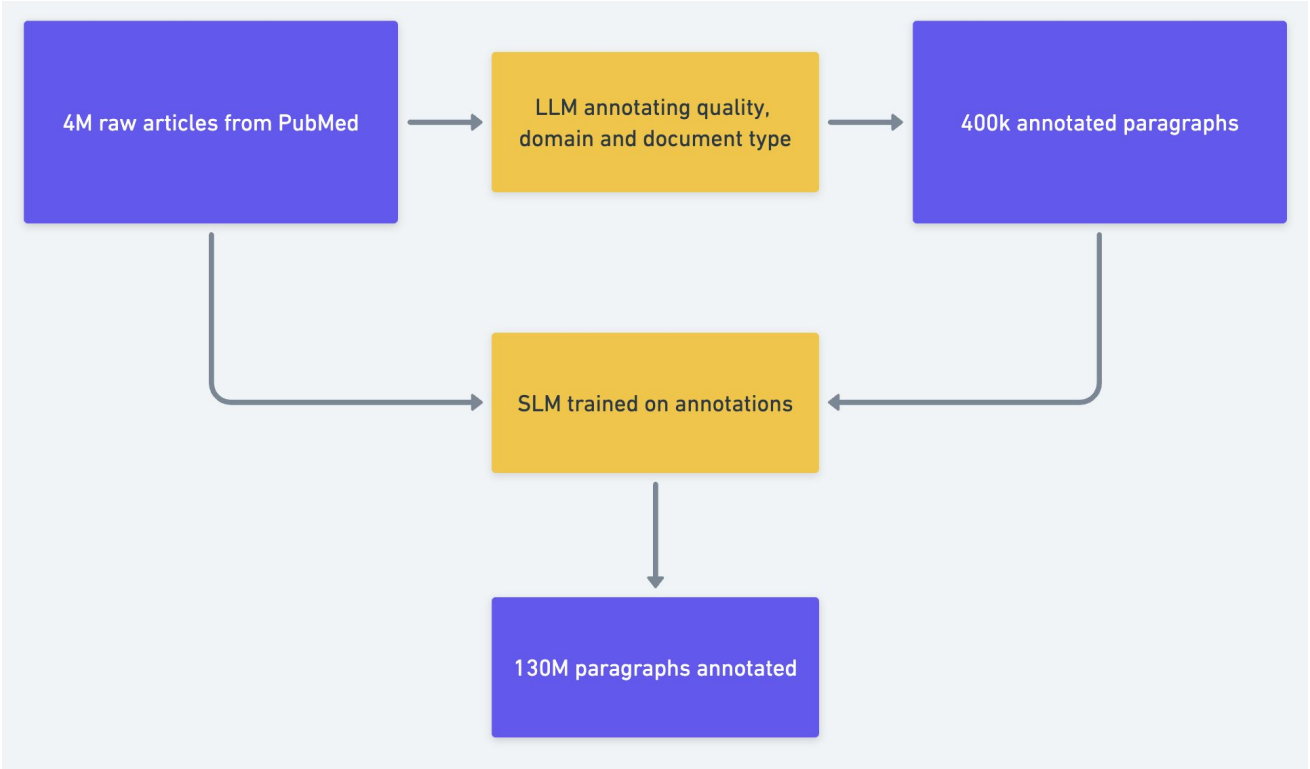

F1-Scores on Different Biomedical Named Entity Recognition Tasks
CamemBERT-base
CamemBERT-bio-small
CamemBERT-bio-base
Clinical
Leaflets
Scientific
F1 Score
85
80
75
70
65
60
CAS1
CAS2
E3C
EMEA
MEDLINE
Dataset

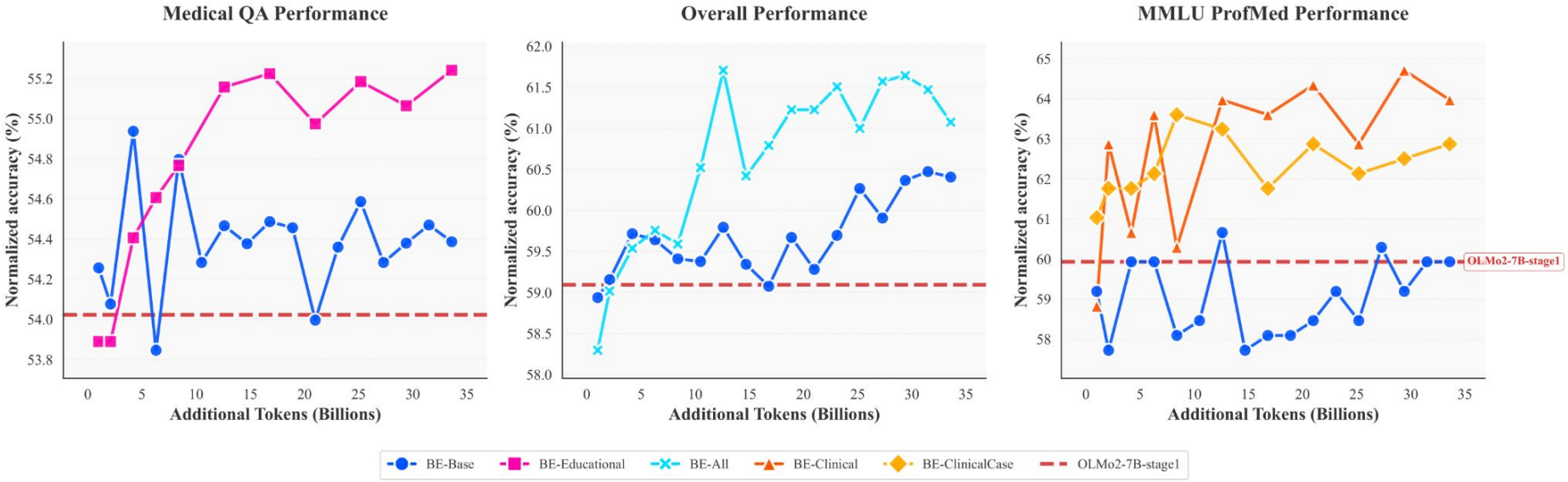
Progrès récent dans la construction de dataset



The FineWeb pipeline

Biomed-Enriched: Using LLMs to Create Enriched Biomedical Text Datasets for Pre-training




Medical QA Performance
Overall Performance
MMLU ProfMed Performance
Normalized accuracy (%)
Additional Tokens (Billions)
OLMo2-7B-stage1
BE-Base
BE-Educational
BE-All
BE-Clinical
BE-ClinicalCase
OLMo2-7B-stage1

Language	Articles	Paragraphs	Clinical Case Paragraphs	% of Paragraphs as Clinical Cases
en	4 113 275	131 579 445	2 113 185	1.61%
es	4339	181 779	1235	0.68%
zh-cn	3649	59 719	0	0.0%
fr	3410	173 325	2586	1.49%
de	2976	248 608	51	0.02%
it	2708	274 819	521	0.19%
pt	934	85 242	4540	5.33%
ko	636	25 535	0	0.0%
ru	222	10 553	0	0.0%
id	189	91 865	15	0.02%

Continual-pretraining vs From-scratch

Pretraining	Wiki + Books → PubMed		PubMed (half time)	PubMed
Vocab	Wiki + Books	PubMed	PubMed	PubMed
BC5-chem	92.85	**93.41**	93.05	93.33
BC5-disease	84.70	85.43	85.02	**85.62**
NCBI-disease	**89.13**	87.60	87.77	87.82
BC2GM	83.82	84.03	84.11	**84.52**
JNLPBA	78.55	79.01	78.98	**79.10**
EBM PICO	73.18	**73.80**	73.74	73.38
ChemProt	76.14	77.05	76.69	**77.24**
DDI	80.88	81.96	81.21	**82.36**
GAD	82.36	82.47	82.8	**83.96**
BIOSSES	89.52	89.93	92.12	**92.30**
HoC	81.54	**83.14**	82.13	82.32
PubMedQA	**60.24**	54.84	55.28	55.84
BioASQ	84.14	79.00	79.43	**87.56**
BLURB score	80.34	80.03	80.23	**81.16**

Evaluation de l'impact du corpus d'entraînement sur les performances sur BLURB

- L'entraînement from-scratch semble en général donner de meilleures performances
- Un entraînement from-scratch est bien plus coûteux en ressource
- Les deux approches ont été exploré en France (CamemBERT-bio, DrBERT, CamemBERT-EDS)

Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing (Gu et al., 2021)

DrBERT argumente pour le from-scratch

	MUSCA-DET T1			MUSCA-DET T2			ESSAI POS			CAS POS			FrenchMedMCQA		QUAERO-EMEA			QUAERO-MEDLINE		
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*Hamming*	*EMR*	*P*	*R*	*F1*	*P*	*R*	*F1*
CamemBERT OSCAR 138 GB	89.04	88.59	88.54	89.87	87.12	88.20	81.57	81.01	81.10	96.37	94.53	95.22	36.24	**16.55**	90.57	91.06	90.71	76.58	78.67	77.41
CamemBERT OSCAR 4 GB	86.09	85.45	85.43	92.68	90.34	91.27	84.01	83.51	83.69	98.15	95.34	96.42	35.75	15.37	90.75	91.16	90.83	78.55	79.33	78.76
CamemBERT CCNET 4 GB	91.12	89.91	90.33	93.10	90.42	91.38	85.60	85.63	85.42	**98.19**	**96.75**	**97.33**	34.71	14.41	90.31	90.59	90.33	78.06	78.11	77.61
PubMedBERT	93.04	91.45	91.99	84.41	80.60	81.97	88.43	87.93	87.78	97.40	94.86	95.90	33.98	14.14	86.89	87.33	86.79	77.33	77.28	77.09
ClinicalBERT	91.79	89.44	90.36	85.43	81.23	82.95	89.09	88.78	88.24	97.94	95.88	96.73	32.78	14.19	84.91	85.47	84.79	75.56	74.85	75.05
BioBERT 1.1	91.82	89.82	90.46	85.52	80.14	81.91	86.76	84.90	85.18	98.10	96.39	97.12	36.19	15.43	84.55	85.03	84.29	72.62	73.30	72.68
DrBERT NACHOS$_{large}$	92.10	90.27	91.04	**94.97**	90.41	92.24	**90.96**	**89.19**	**89.75**	97.37	94.49	95.65	36.66	15.32	**91.93**	**92.52**	**92.09**	77.85	78.54	77.88
DrBERT NACHOS$_{small}$	93.35	90.62	91.77	91.31	86.60	88.57	90.12	88.37	88.76	97.04	94.88	95.70	**37.37**	13.34	91.54	92.00	91.66	77.91	79.34	78.18
ChuBERT NBDW$_{small}$	**94.88**	90.79	92.23	94.77	90.27	92.17	88.53	87.73	87.71	97.00	94.65	95.61	35.16	14.79	88.11	88.78	88.15	75.05	76.57	74.94
ChuBERT NBDW$_{mixed}$	94.39	**91.93**	**92.73**	94.22	90.02	91.71	86.36	85.50	85.73	97.77	95.30	96.35	34.58	12.21	90.36	90.94	90.52	78.61	79.32	78.63
CamemBERT NACHOS$_{small}$	81.44	81.39	80.96	79.74	78.08	78.70	80.59	79.88	80.04	95.64	91.57	92.46	32.87	13.76	67.56	77.48	71.10	55.45	62.34	57.43
PubMedBERT NACHOS$_{small}$	92.51	91.49	91.53	94.95	**92.55**	**93.62**	84.73	83.80	83.85	97.82	96.12	96.81	35.88	15.21	90.97	91.27	91.03	**82.03**	**81.71**	**81.73**
CamemBERT NBDW$_{small}$	82.35	81.59	81.57	78.14	76.38	77.12	79.44	79.79	79.25	95.98	92.11	93.18	27.73	11.89	53.44	73.11	61.75	48.71	61.33	53.05

Table 7: Performance on public biomedical downstream tasks. Best model in bold and second is underlined.

DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains (Labrak et al., 2023)

Des différences dans la méthodologie d'évaluation

- **Pour DrBERT, l'évaluation est réalisée au niveau du token:**

  -> Classification classique des différents labels dont "O"

- **Pour CamemBERT-bio, l'évaluation est réalisée au niveau de l'entité:**

  -> seqeval avec IOB2 strict

Phrase à predire	Le	par	acet	am	ol	tra	ite	la	fièvre
Token classification	O	CHEM	CHEM	CHEM	CHEM	O	O	O	DISO
seqeval	O	B-CHEM	I-CHEM	I-CHEM	I-CHEM	O	O	O	B-DISO

Différence entre seqeval et la classification de token

Models	MEDLINE			
	Precision	*Recall*	*F-measure*	*CO_2 eq (g.)*
CamemBERT	0.64 [0.62-0.66]	0.66 [0.64-0.67]	0.65 [0.63-0.66]	1.9
FlauBERT	0.67 [0.65-0.68]	0.69 [0.67-0.71]	0.68 [0.66-0.69]	2.2
mBERT	0.63 [0.61-0.65]	0.67 [0.65-0.69]	0.65 [0.63-0.67]	3
frALBERT	0.53 [0.51-0.54]	0.52 [0.49-0.54]	0.52 [0.50-0.54]	1.1
CamemBERT-bio	0.66 [0.65-0.68]	0.70 [0.68-0.72]	0.68 [0.66-0.70]	2.2
DrBERT	0.63 [0.61-0.65]	0.65 [0.63-0.67]	0.64 [0.62-0.66]	2
Baseline	**0.73**	0.30	0.42	-
Van Mulligen et al. (2016)	0.68	**0.72**	**0.70**	-

Table 5: Performance of nested entity extraction on the MEDLINE test set.

Models	EMEA			
	Precision	*Recall*	*F-measure*	*CO_2 eq (g.)*
CamemBERT	0.66 [0.62-0.70]	0.65 [0.56-0.73]	0.65 [0.59-0.72]	3.9
FlauBERT	0.69 [0.67-0.72]	0.66 [0.59-0.73]	0.68 [0.63-0.71]	4.4
mBERT	0.67 [0.64-0.72]	0.67 [0.61-0.73]	0.67 [0.63-0.72]	4
frALBERT	0.62 [0.57-0.67]	0.65 [0.61-0.70]	0.63 [0.59-0.68]	2.4
CamemBERT-bio	0.70 [0.66-0.74]	0.68 [0.61-0.75]	0.69 [0.63-0.74]	5.2
DrBERT	0.69 [0.66-0.72]	0.64 [0.58-0.71]	0.66 [0.62-0.71]	3
Baseline	**0.73**	0.43	0.55	-
Van Mulligen et al. (2016)	0.72	**0.79**	**0.75**	-

Table 6: Performance of nested entity extraction on the EMEA test set.

DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains (Labrak et al., 2023)
A Benchmark Evaluation of Clinical Named Entity Recognition in French (Bannour et., 2024)

Le continual-pretraining offre également des performances similaires sur des données privées

Model	APMed (F1-score)	QUAERO (F1-score) EMEA	QUAERO (F1-score) MEDLINE	Total
EDS-fine-tuned	.902 (±0.003)*	.729 (±0.008) *	.597 (±0.007) *	**.655 (±0.007)**
EDS-from-scratch	**.908 (±0.005)** *	.693 (±0.012) *	**.601 (±0.01)** *	.642 (±0.007) *
CamemBERT-base	.866 (±0.007)	**.737 (±0.006)**	.584 (±0.004)	.651 (±0.004)
Best QUAERO [26] model		.749	.698	

Evaluation de différents modèles entraînés sur les données de l'EDS de l'APHP

Learning structures of the French clinical language: development and validation of word embedding models using 21 million clinical reports from electronic health records (Dura et al., 2022)

Impact environnemental

	Training time (hours)	Hardware type	Total GPU-hours	Estimation of carbon emitted (kg CO2 eq.)
DrBERT	20h	128xV100	2560	26.11
AliBERT	20h	48xA100	960	8.16
CamemBERT-bio	39h	2xV100	78	0.8

Table 7: Carbon emitted estimation based on hardware and training time. We used a rate of 34g CO2eq. per kWh, reflecting the average over the last 12 months in France starting from September 2022. This time frame and location coincide with when and where all experiments were conducted.

-> Compte-tenu des différences de performances et des ressources impliquées, on estime préférable le continual-pretraining pour le domaine biomédical

AliBERT: A Pretrained Language Model for French Biomedical Text (Berhe et al., 2023)

Adaptation de LLMs au biomédical

MediTron : continual-pretraining sur PubMed

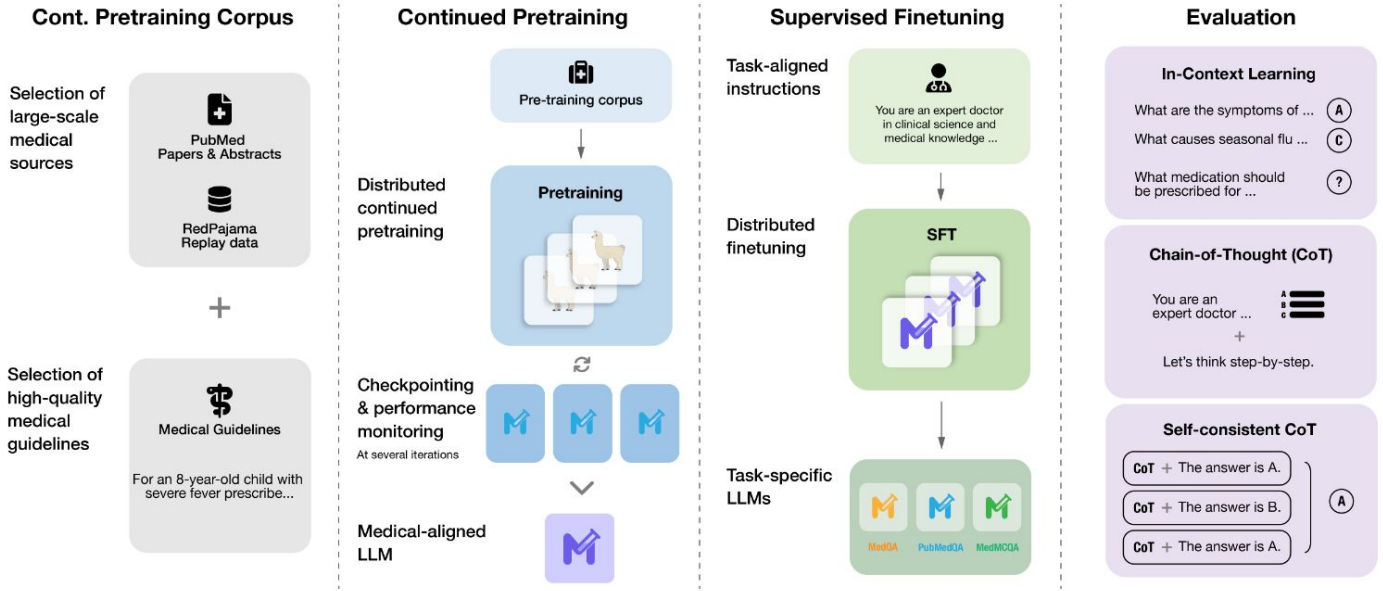


Figure 2: **MEDITRON.** The complete pipeline for continued pretraining, supervised finetuning, and evaluation of MEDITRON-7B and MEDITRON-70B.

MEDITRON-70B: Scaling Medical Pretraining forLarge Language Models (Chen et al., 2023)

MediTron : continual-pretraining de LlaMa sur PubMed

Model	Accuracy (↑)					
	MMLU-Medical	PubMedQA	MedMCQA	MedQA	MedQA-4-Option	Avg
Top Token Selection						
Mistral-7B*	55.8	17.8	40.2	32.4	41.1	37.5
Zephyr-7B-β*	63.3	46.0	43.0	42.8	48.5	48.7
PMC-Llama-7B	59.7	59.2	57.6	42.4	49.2	53.6
Llama-2-7B	56.3	61.8	54.4	44.0	49.6	53.2
MEDITRON-7B	55.6	74.4	59.2	47.9	52.0	57.5
Clinical-Camel-70B*	65.7	67.0	46.7	50.8	56.8	57.4
Med42-70B*	74.5	61.2	59.2	59.1	63.9	63.6
Llama-2-70B	74.7	78.0	62.7	59.2	61.3	67.2
MEDITRON-70B	73.6	80.0	65.1	60.7	65.4	69.0
Chain-of-thought						
Llama-2-70B	76.7	79.8	62.1	60.8	63.9	68.7
MEDITRON-70B	74.9	81.0	63.2	61.5	67.8	69.7
Self-consistency Chain-of-thought						
Llama-2-70B	**77.9**	80.0	62.6	61.5	63.8	69.2
MEDITRON-70B	77.6	**81.6**	**66.0**	**64.4**	**70.2**	**72.0**

MEDITRON-70B: Scaling Medical Pretraining forLarge Language Models (Chen et al., 2023)

BioMistral : continual-pretraining de Mistral sur PubMed

	MMLU										
	Clinical KG	Medical Genetics	Anatomy	Pro Medicine	College Biology	College Medicine	MedQA	MedQA 5 opts	PubMedQA	MedMCQA	Avg.
BioMistral 7B	60.9 ±1.5	61.7 ±2.1	49.6 ±1.2	55.1 ±1.3	56.9 ±1.0	55.5 ±1.7	44.4 ±0.2	37.4 ±0.4	37.6 ±1.5	43.9 ±0.3	50.3
Mistral 7B Instruct	57.0 ±0.8	56.7 ±0.5	46.9 ±0.3	51.0 ±1.1	58.6 ±0.9	50.1 ±1.0	42.3 ±0.3	34.5 ±0.5	**72.2** ±0.5	42.8 ±0.5	51.2
BioMistral 7B Ensemble	62.8 ±0.5	62.7 ±1.7	46.9 ±0.3	57.0 ±0.6	60.6 ±0.9	56.3 ±0.3	44.7 ±0.4	37.1 ±0.6	68.0 ±0.4	44.8 ±0.3	54.1
BioMistral 7B DARE	61.3 ±0.4	61.0 ±2.8	**49.9** ±0.9	55.3 ±0.7	**64.4** ±0.9	53.9 ±1.4	**47.0** ±0.5	38.8 ±0.7	70.0 ±0.7	44.9 ±0.2	54.6
BioMistral 7B TIES	62.3 ±0.5	61.3 ±1.9	48.1 ±2.2	55.8 ±0.8	57.2 ±0.7	56.5 ±1.5	44.0 ±0.4	37.7 ±0.4	44.3 ±0.8	44.0 ±0.3	51.1
BioMistral 7B SLERP	**63.1** ±1.6	**63.3** ±0.9	**49.9** ±1.9	57.4 ±0.3	63.4 ±0.9	**57.8** ±0.9	46.6 ±0.2	**38.9** ±0.4	68.1 ±1.4	**45.7** ±0.7	**55.4**
MedAlpaca 7B	49.1 ±1.3	49.0 ±5.7	48.4 ±1.9	**63.8** ±0.8	47.2 ±0.6	43.5 ±1.8	35.4 ±0.3	30.4 ±0.6	56.0 ±0.9	31.2 ±0.2	45.4
PMC-LLaMA 7B	25.3 ±1.5	26.0 ±3.7	31.9 ±1.8	16.9 ±0.5	28.0 ±2.4	24.9 ±1.2	27.6 ±0.8	21.1 ±0.8	53.3 ±0.6	23.5 ±0.3	27.8
MediTron-7B	37.9 ±1.5	47.0 ±3.7	39.3 ±1.6	34.2 ±1.0	42.6 ±1.4	30.4 ±0.7	34.8 ±0.6	26.3 ±0.5	55.9 ±1.0	33.6 ±0.2	38.2
BioMedGPT-LM-7B	50.1 ±1.0	52.0 ±0.8	46.2 ±1.8	47.3 ±1.7	47.9 ±2.5	45.5 ±0.7	39.3 ±1.2	34.9 ±0.4	58.6 ±0.3	34.9 ±0.5	45.7
GPT-3.5 Turbo 1106	74.71 ±0.3	74.00 ±2.2	65.92 ±0.6	72.79 ±1.6	72.91 ±1.7	64.73 ±2.9	57.71 ±0.3	50.82 ±0.7	72.66 ±1.0	53.79 ±0.2	66.0

Table 2: Performance of 3-shot in-context learning. The scores represent accuracy (↑) and are averaged across 3 random seeds. BioMistral 7B Ensemble, DARE, TIES, and SLERP are model merging strategies that combine BioMistral 7B and Mistral 7B Instruct. Best model in bold, and second-best underlined.

Limites du continual-pretraining

- Pour obtenir la même augmentation de F1-scores qu'avec les modèles BERT, il faut considérablement plus de ressources pour les LLMs

Meditron a eu besoin de 128xA100 pendant 2 semaines pour le modèle 70B

- Cela semble vain de vouloir encoder toutes les connaissances biomédicales dans le modèle