Adaptation des connaissances médicales pour les grands modèles de langue : Stratégies et analyse comparative

Ikram Belmadani[1] Benoit Favre[1] Richard Dufour[2] Frédéric Béchet[1] Carlos Ramisch[1]

(1) Aix-Marseille Université, CNRS, LIS UMR 7020, 13000, Marseille, France

(2) Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

ikram.belmadani@lis-lab.fr, benoit.favre@lis-lab.fr, richard.dufour@univ-nantes.fr, carlos.ramisch@lis-lab.fr, frederic.bechet@lis-lab.fr

RÉSUMÉ

Cet article présente une étude sur l'adaptation des grands modèles de langue (LLMs) à des domaines spécialisés disposant de données limitées. Bien que certaines recherches remettent en question le pré-entraînement adaptatif (DAPT) dans le contexte médical en anglais, nous montrons que l'adaptation au domaine peut être efficace sous certaines conditions. En prenant comme exemple l'adaptation au domaine médical en français, nous comparons de manière systématique le pré-entraînement continu (CPT), l'affinage supervisé (SFT) et une approche combinée (CPT suivi de SFT). Nos résultats indiquent que l'adaptation d'un modèle généraliste à de nouvelles données dans le domaine médical offre des améliorations notables (taux de réussite de 87%), tandis que l'adaptation supplémentaire de modèles déjà familiarisés avec ce domaine procure des bénéfices limités. Bien que CPT+SFT offre les meilleures performances globales, SFT-seul présente des résultats solides et requiert moins de ressources matérielles.

ABSTRACT

Medical Knowledge Adaptation for Large Language Models : Strategies and Comparative Analysis

This paper presents an investigation into the adaptation of Large Language Models (LLMs) to specialized domains with limited data. Although recent studies have questioned the effectiveness of domain-adaptive pre-training (DAPT) in English medical contexts, we show that domain adaptation can be effective under certain conditions. Using French medical data as a case study, we systematically compare continual pre-training (CPT), supervised fine-tuning (SFT), and a combined approach (CPT followed by SFT). Our results indicate that extending a general-purpose model with new domain data yields substantial improvements (87% success rate), whereas further adaptation of models already exposed to similar knowledge offers limited gains. Although CPT+SFT demonstrates the highest overall performance, direct SFT provides strong results and is more computationally efficient.

MOTS-CLÉS : Grands modèles de langue (LLMs), Pré-entraînement continu (CPT), Affinage supervisé (SFT), Médical, Français.

KEYWORDS: Large Language Models (LLMs), Continual Pre-training (CPT), Supervised Fine-tuning (SFT), Medical, French.

1 Introduction

Les récents progrès des grands modèles de langue (LLMs) ont intensifié le débat sur leur adaptation à des domaines et langues spécialisés. Bien que diverses stratégies existent, choisir la meilleure approche demeure complexe lorsque l'on vise un domaine précis dans une langue à ressources limitées. Cela est notamment le cas dans le domaine médical, où la plupart des travaux se concentrent sur des modèles construits principalement à partir de données anglophones (p. ex. BioMistral (Labrak *et al.*, 2024) ou OpenBioLLM (Pal & Sankarasubbu, 2024)).

Le développement de modèles spécialisés hors anglais se heurte à la rareté des données, à l'importance du coût de calcul et aux biais potentiels introduits par l'évaluation sur des jeux de données traduits. En médecine, ces défis sont cruciaux, car la fiabilité et la précision sont primordiales, alors même que les ressources francophones restent limitées.

Les travaux de Jeong *et al.* (2024) questionnent l'intérêt du pré-entraînement adaptatif (DAPT) en contexte médical, montrant que les modèles biomédicaux n'apportent pas toujours d'améliorations intéressantes par rapport aux versions génériques. Cette remise en cause soulève la question de la stratégie la plus efficace, surtout lorsque les connaissances du domaine sont peu présentes dans le modèle initial.

Pour répondre à ces enjeux, nous étudions diverses méthodes d'adaptation de LLMs biomédicaux en français, comparant le pré-entraînement continu (CPT), l'affinage supervisé (SFT) et une approche hybride (CPT suivi de SFT). Notre point de départ est le modèle généraliste Mistral-7B (Jiang *et al.*, 2023), utilisé à la fois dans sa version d'origine et sa variante pré-entraînée sur des textes médicaux en anglais.

Notre protocole d'évaluation inclut des jeux de données anglais traduits en français (PubMedQA (Jin *et al.*, 2019) et MedMCQA (Pal *et al.*, 2022)) ainsi que des questions natives en français, évaluées via des tâches de questions à choix multiples et à réponses ouvertes.

Les apports de travail peuvent se résumer ainsi :

1. La mise à disposition de ressources biomédicales en traitement automatique des langues (modèles et données) pour le français, publiées sous licence Apache 2.0 [1].
2. La proposition d'un cadre d'évaluation commun pour comparer différentes stratégies d'adaptation.
3. Une analyse comparative qui propose des recommandations pratiques d'adaptation de modèles selon la disponibilité des données (brutes ou annotées) et des ressources de calcul.

2 État de l'art

L'adaptation des LLMs au domaine médical connaît un essor important pour traiter diverses tâches en santé. Deux stratégies principales, et potentiellement complémentaires, dominent : le pré-entraînement continu (CPT), qui prolonge l'entraînement du modèle sur des corpus médicaux non annotés pour enrichir ses connaissances spécialisées, et l'affinage supervisé (SFT), qui adapte le modèle de façon supervisée via des paires instruction-réponse ciblant des tâches précises. L'approche CPT a montré son efficacité dans des travaux tels que MediTron (Chen *et al.*, 2023), BioMistral (Labrak *et al.*,

1. https://github.com/ikram28/medllm-strategies

2024) et PMC-LLaMA (Wu *et al.*, 2024), mais Jeong *et al.* (2024) remettent en cause ces avancées en soulignant que les gains rapportés ne sont pas toujours significatifs. Parallèlement, des approches SFT comme ChatDoctor (Li *et al.*, 2023) et MedAlpaca (Han *et al.*, 2023) obtiennent des résultats prometteurs, bien que limités aux données en anglais.

L'adaptation à d'autres langues demeure complexe, faute de ressources spécialisées disponibles. Plusieurs travaux récents adoptent une perspective multilingue, notamment Medical mT5 (García-Ferrero *et al.*, 2024), BiMediX (Pieri *et al.*, 2024), Apollo (Wang *et al.*, 2024) et MMedLM (Qiu *et al.*, 2024). Toutefois, leur évaluation repose souvent sur des jeux de données traduits, avec peu de tests en langue native, ce qui limite l'analyse de leurs performances réelles.

L'évaluation des LLMs médicaux pose des défis particuliers : la plupart des ensembles de référence (PubMedQA (Jin *et al.*, 2019), MedQA (Jin *et al.*, 2019), MedMCQA (Pal *et al.*, 2022)) sont conçus pour l'anglais, et les traductions peuvent atténuer les nuances culturelles et terminologiques. De plus, nombre de benchmarks se concentrent sur des QCM, laissant planer des doutes quant à l'évaluation globale des capacités du modèle.

Enfin, il n'existe pas encore de cadre unifié permettant de comparer objectivement les stratégies d'adaptation. Les études existantes se limitent souvent à une seule méthode ou comparent des modèles qui diffèrent en taille ou en jeu de données, rendant la comparaison directe difficile. L'impact du coût de calcul sur les gains de performance reste également mal défini, surtout dans des environnements à ressources limitées. Dans ce travail, nous proposons une évaluation contrôlée des différentes approches, en utilisant la même architecture de base et un protocole commun, afin d'offrir des recommandations claires pour le développement de LLMs médicaux dans des langues peu dotées.

3 Expériences

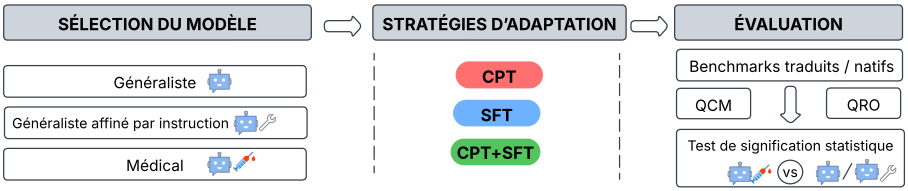


FIGURE 1 – Schéma d'évaluation des stratégies d'adaptation, illustrant les trois composantes principales : sélection du modèle, stratégies d'adaptation et méthodologie d'évaluation (incluant questions à choix multiples (QCM) et questions à réponses ouvertes (QRO).

Nous proposons un cadre d'évaluation pour l'adaptation de modèles de langue dans des contextes à faibles ressources, comme illustré dans la Figure 1. À partir de plusieurs modèles de base, nous comparons différents chemins d'adaptation pour déterminer l'impact du choix initial et de la stratégie sur les performances et le coût de calcul.

Notre étude s'articule autour de deux questions de recherche :

QR1 : Le choix du modèle de base (généraliste *vs.* déjà adapté au domaine médical en anglais) influence-t-il significativement le succès de l'adaptation ?

QR2 : Quelle stratégie d'adaptation offre le meilleur compromis entre performance et coût de calcul ?

Pour y répondre, nous présentons : (i) les modèles de base et les approches d'adaptation en Section 3.1, (ii) les données d'entraînement en Section 3.2, (iii) les procédures d'entraînement en Section 3.3, et (iv) le protocole d'évaluation en Section 3.4.

3.1 Modèles de base et approches d'adaptation

Nous étudions des stratégies d'adaptation pour des modèles biomédicaux en français, en nous appuyant sur l'architecture Mistral-7B. Trois modèles de base servent de points de départ :

— **Mistral-7B-v0.1** (Jiang *et al.*, 2023) : modèle à 7 milliards de paramètres entraîné sur un domaine général.

— **Mistral-7B-instruct-v0.1** (Jiang *et al.*, 2023) : version de Mistral-7B-v0.1 affinée via des paires instruction-réponse.

— **BioMistral-7B** (Labrak *et al.*, 2024) : variante de Mistral-7B-instruct-v0.1 adaptée à l'anglais médical (pré-entraînement supplémentaire sur PubMed Central Open Access).

Nous avons retenu Mistral pour sa bonne couverture du français parmi les modèles open source et parce qu'il est utilisé dans des études comparables en anglais (Labrak *et al.*, 2024).

Trois stratégies d'adaptation sont évaluées :

— **Pré-entraînement continu (CPT)** : entraînement supplémentaire sur un corpus médical.

— **Affinage supervisé (SFT)** : adaptation via des paires instruction-réponse.

— **CPT+SFT** : application séquentielle du pré-entraînement continu suivie de l'affinage supervisé.

Ces approches sont appliquées à chaque modèle initial choisi, organisé en familles de modèles comme illustré dans le Tableau (b) de la Figure 2.

3.2 Données d'entraînement

Nous utilisons deux ensembles de données distincts selon la stratégie d'adaptation.

Stratégie CPT Nous exploitons le corpus NACHOS (opeN crAwled frenCh Healthcare cOrpuS), un ensemble de textes médicaux en français totalisant 7,4 Go de données (plus d'un milliard de mots) provenant de 24 sites web médicaux (Labrak *et al.*, 2023). Les détails sur la collecte et les caractéristiques du corpus figurent en Annexe B.

Stratégie SFT Nous avons construit un ensemble de 30K paires question-réponse médicales en français, réparties en trois catégories : (1) 10K questions-réponses natives (examens de pharmacie et concours médicaux), (2) 10K questions-réponses traduites d'ensembles anglophones (examens médicaux américains, flashcards médicales), et (3) 10K questions-réponses générées à partir de textes médicaux en français via un modèle de langue. Les données incluent des QCM à une ou plusieurs

réponses correctes et des questions à réponses ouvertes (QRO) avec ou sans contexte. Les sources et la composition sont détaillées en Annexe C.

3.3 Procédure d'entraînement

Nos procédures comparent deux approches afin de mesurer le compromis entre coût de calcul et flexibilité des modèles.

Stratégie CPT Nous suivons la méthode de *batching* optimisée de BioMistral (Labrak *et al.*, 2024), qui regroupe les séquences en fonction de la tokénisation et évite le *padding* superflu. L'entraînement couvre 2,8 époques ; nous utilisons l'optimiseur AdamW (Loshchilov & Hutter, 2019) avec un taux d'apprentissage de 2×10^{-5} et un ordonnanceur cosinus, sans phase d'échauffement. Le *weight decay* est fixé à 0,01, la taille de batch à 16 et l'accumulation de gradients à 2. Les entraînements ont été effectués sur 32 GPU NVIDIA A100 80GB ou H100 80GB.

Stratégie SFT Nous adoptons DoRA (Weight-Decomposed Low-Rank Adaptation) (Liu *et al.*, 2024), une amélioration de LoRA (Hu *et al.*, 2021) qui décompose les poids pré-entraînés en composantes de magnitude et de direction. Cette méthode réduit le nombre de paramètres à entraîner tout en préservant une forte capacité d'adaptation. Des expériences préliminaires, réalisées sur un sous-ensemble des données SFT, ont comparé DoRA à LoRA, VeRA (Kopiczko *et al.*, 2024) et à l'affinage complet, montrant que DoRA offrait les meilleures performances, probablement en évitant le surapprentissage observé avec l'affinage complet. Le SFT a été mené sur 10 époques. Les détails des hyperparamètres sont fournis en Annexe F.

3.4 Protocole d'évaluation

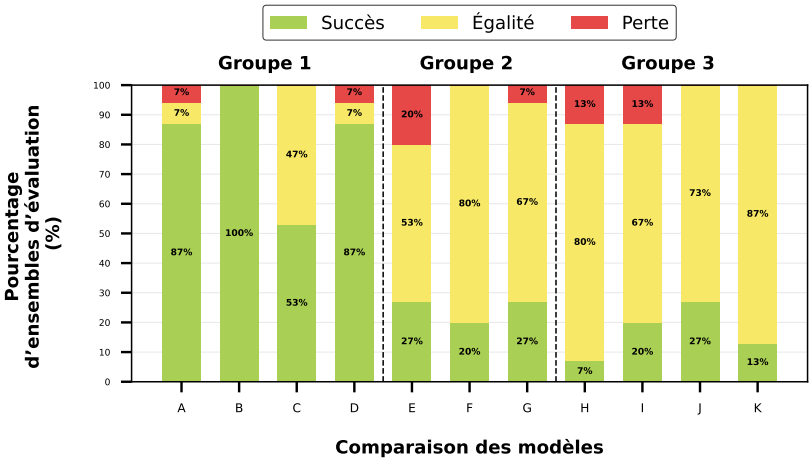
Jeu de données d'évaluation

- **QCM traduits** : Nous traduisons vers le français, via GPT-3.5-turbo, plusieurs benchmarks médicaux anglophones : MedQA (Jin *et al.*, 2021), MedMCQA (Pal *et al.*, 2022), PubMedQA (Jin *et al.*, 2019) et les catégories médicales de MMLU (Hendrycks *et al.*, 2020).
- **QCM natifs** : Nous utilisons FrBMedQA (Kaddari & Bouchentouf, 2022), issu d'articles biomédicaux Wikipédia en français couvrant huit groupes sémantiques UMLS (chimiques/médicaments, anatomie, physiologie, pathologies, phénomènes, procédures, gènes/séquences moléculaires, dispositifs). Les questions, à l'origine sous forme de tests à trous, ont été converties en QCM via GPT-4o-mini (détails en Annexe D). Nous ajoutons FrMedMCQA, extrait depuis S-Editions [2] (une plateforme offrant du matériel d'étude pour les étudiants en médecine en France), qui couvre divers domaines médicaux (oncologie, cardiologie, dermatologie, etc.).
- **QRO** : Nous traduisons l'ensemble K-QA (Manes *et al.*, 2024) (201 questions avec réponses de médecins) grâce à GPT-4o-mini. Nous ajoutons deux ensembles natifs depuis S-Editions : FrClinicalQA (questions de cas cliniques) et FrMedQA (questions générales), tous deux validés par un processus de nettoyage et vérification manuelle.

Un résumé des sources et de la taille de chaque ensemble figure en Table 6.

2. https://s-editions.fr/

(a)



(b)

Groupe	ID	Modèle Médical	Modèle de base	Stratégie
Groupe 1 (Mistral)	A	MedMistral-CPT-7B	Mistral-7B-v0.1	CPT
	B	MedMistral-CPT-SFT-7B	Mistral-7B-v0.1	CPT+SFT
	C	MedMistral-CPT-SFT-7B	MedMistral-CPT-7B	CPT+SFT
	D	MedMistral-SFT-7B	Mistral-7B-v0.1	SFT
Groupe 2 (Mistral-Instruct)	E	MedMistralInstruct-CPT-7B	Mistral-7B-Instruct-v0.1	CPT
	F	MedMistralInstruct-CPT-SFT-7B	MedMistralInstruct-CPT-7B	CPT+SFT
	G	MedMistralInstruct-CPT-SFT-7B	Mistral-7B-Instruct-v0.1	CPT+SFT
Groupe 3 (BioMistral)	H	BioMistral-CPT-7B	BioMistral-7B	CPT
	I	BioMistral-CPT-SFT-7B	BioMistral-7B	CPT+SFT
	J	BioMistral-CPT-SFT-7B	BioMistral-CPT-7B	CPT+SFT
	K	BioMistral-SFT-7B	BioMistral-7B	SFT

FIGURE 2 – **Évaluation des stratégies d'adaptation de modèles.** (a) Analyse Succès/Égalité/Perte sur les ensembles d'évaluation, montrant la proportion de jeux de données pour lesquels chaque modèle médical adapté affiche une amélioration significative (Succès), aucune différence significative (Égalité) ou une dégradation significative (Perte) par rapport à son modèle de base. Les comparaisons (A à K) renvoient à (b). (b) Présentation des modèles médicaux, de leurs bases et des stratégies d'adaptation.

Stratégie de *prompting* Nous réalisons une évaluation en *zero-shot* pour simuler des scénarios réels et en raison des contraintes des ensembles de données, la plupart (sauf les QCM traduits) ayant de petits ensembles de test (Table 6). Nous utilisons un décodage *greedy* pour générer les réponses. Pour les QCM, conformément à Liang *et al.* (2022), Beeching *et al.* (2023) et Chen *et al.* (2023), nous filtrons le vocabulaire pour ne conserver que les lettres de réponse possibles, empêchant ainsi la génération de jetons non pertinents ou d'hallucinations. Les *prompts* sont détaillés en Annexe E.

Analyse de la significativité statistique Pour vérifier si les améliorations observées sont significatives, nous appliquons une méthode de *bootstrap* par centiles, similaire à Jeong *et al.* (2024). Nous réalisons des ré-échantillonnages (avec remise) de la même taille que l'ensemble de test original. Pour chaque échantillon, nous calculons l'écart de performances entre modèles (exactitude (*exact-match accuracy*) pour les QCM ou F1 BERTScore (Zhang *et al.*, 2020) pour les QRO). L'opération est répétée 10 000 fois afin d'établir une distribution des écarts. Une amélioration est jugée significative si l'intervalle de confiance à 95% ne contient pas zéro. Contrairement à Jeong *et al.* (2024), nous

Modèle	Stratégie	Moyenne	PubMedQA	MedQA 4 Options	MedQA 5 Options	MedMCQA	MMLU					
							Clinical Knowledge	Medical Genetics	Anatomy	Pro. Medicine	College Biology	College Medicine
Mistral-7B-v0.1	Modèle de base	0,87	4,00	0,08	0,24	0,19	1,51	0,00	0,00	0,00	2,08	0,58
MedMistral-CPT-7B	CPT	36,00	41,00	28,83	23,10	33,09	44,15	43,00	37,78	27,21	46,53	35,26
MedMistral-CPT-SFT-7B	CPT+SFT	**47,83**	**64,00**	**42,66**	**35,19**	**37,99**	**49,43**	**56,00**	42,96	**53,68**	**47,22**	49,13
MedMistral-SFT-7B	SFT	43,65	54,60	39,67	31,89	35,79	44,91	43,00	**45,19**	48,53	40,28	**52,60**
Mistral-7B-Instruct-v0.1	Modèle de base	39,96	54,40	29,14	24,90	31,87	46,42	44,00	37,78	**46,32**	40,28	44,51
MedMistralInstruct-CPT-7B	CPT	43,03	34,80	**37,08**	**32,29**	**38,42**	**50,57**	**59,00**	**42,96**	40,81	**49,31**	**45,09**
MedMistralInstruct-CPT-SFT-7B	CPT+SFT	**43,20**	**59,20**	36,29	31,50	36,39	42,26	53,00	42,22	40,07	46,53	44,51
BioMistral-7B	Modèle de base	**41,39**	**54,60**	32,44	26,08	31,68	**52,08**	43,00	40,74	**45,22**	**42,36**	**45,66**
BioMistral-CPT-7B	CPT	35,14	14,80	28,75	27,49	30,62	44,15	42,00	39,26	43,38	41,67	39,31
BioMistral-CPT-SFT-7B	CPT+SFT	34,53	36,60	35,59	**30,24**	**35,41**	32,83	37,00	34,07	29,41	38,89	35,26
BioMistral-SFT-7B	SFT	37,68	44,60	**36,68**	**30,24**	31,87	38,49	**44,00**	**41,48**	39,34	35,42	34,68

TABLE 1 – Performance en *zero-shot* sur des tâches de QCM traduites. Les scores sont rapportés en utilisant l'exactitude(%). Le modèle le plus performant au sein de chaque groupe est mis en évidence en **gras**, et le modèle global le plus performant est souligné.

appliquons la correction de Bonferroni pour tenir compte des multiples comparaisons, réduisant ainsi le risque de faux positifs. Cette correction ajuste le seuil de significativité à α/m, où α est fixé à 0,05 dans notre cas et m représente le nombre total de comparaisons. Enfin, pour évaluer la généralisation des modèles adaptés, nous analysons les performances par ensemble de test et calculons le taux de succès, c'est-à-dire la proportion de jeux de données où un modèle affiné surpasse sa version de base.

4 Résultats et discussions

4.1 Choix du modèle de base

Nous présentons les résultats de l'évaluation de trois groupes de modèles sur plusieurs tâches. Les tableaux 1, 2 et 3 montrent respectivement les performances sur les QCM traduits, les QCM natifs et les OEQ.

Analyse des performances des modèles

Groupe 1 (basé sur Mistral) : MedMistral-CPT-SFT-7B montre des améliorations significatives sur tous les indicateurs. L'exactitude sur les QCM traduits passe de 0,87 % à 47,83 %, et sur les QCM natifs de 3,97 % à 36,55 %. Le score F1 BERTScore en QRO progresse de 0,55 à 0,67. Ces gains sont statistiquement significatifs avec un taux de succès de 100 %.

Groupe 2 (basé sur Mistral-instruct) : Les résultats de ce groupe sont intermédiaires. MedMistralInstruct-CPT-SFT-7B améliore les scores de 39,96 % à 43,03 % (QCM traduits), de 27,13 % à 36,46 % (QCM natifs), et conserve un score F1 BERTScore de 0,67 en QRO. Toutefois, les tests statistiques révèlent une significativité faible (27 % de taux de succès).

Groupe 3 (basé sur BioMistral) : En partant d'un modèle médical anglais, on obtient des résultats contrastés. Les QCM traduits montrent une baisse de performance après adaptation. Les QCM natifs s'améliorent de 30,13 % à 35,52 % (BioMistral-CPT-SFT-7B), mais sans significativité statistique. En revanche, les QRO présentent une amélioration significative avec une hausse de 0,22 en F1 BERTScore.

Impact du choix du modèle de base La comparaison des meilleurs modèles de chaque groupe (Table 5) montre que MedMistral-CPT-SFT-7B (Groupe 1) surpasse significativement BioMistral-CPT-SFT-7B (Groupe 3) avec un taux de succès de 73 %. Cela suggère qu'il est plus efficace de partir

Modèle	Stratégie	Moyenne	FrBMedQA	FrMedMCQA
Mistral-7B-v0.1	Modèle de base	3,97	7,93	0,00
MedMistral-CPT-7B	CPT	33,48	50,56	16,39
MedMistral-CPT-SFT-7B	CPT+SFT	**36,55**	**50,70**	**22,40**
MedMistral-SFT-7B	SFT	29,88	48,28	11,47
Mistral-7B-Instruct-v0.1	Modèle de base	27,13	43,88	10,38
MedMistralInstruct-CPT-7B	CPT	**36,46**	**53,25**	19,67
MedMistralInstruct-CPT-SFT-7B	CPT+SFT	35,50	50,79	**20,21**
BioMistral-7B	Modèle de base	30,13	46,06	14,20
BioMistral-CPT-7B	CPT	32,83	**47,63**	18,03
BioMistral-CPT-SFT-7B	CPT+SFT	**35,52**	47,54	**23,49**
BioMistral-SFT-7B	SFT	27,93	46,57	9,28

TABLE 2 – Performances en *zero-shot* sur des QCM en français natif. Les scores correspondent à l'exactitude (%). Le meilleur modèle de chaque groupe est indiqué en **gras** et le meilleur modèle global est souligné.

d'un modèle généraliste que d'affiner à partir d'un modèle déjà spécialisé dans le domaine médical. Comparé à MedMistralInstruct-CPT-SFT-7B (Groupe 2), les résultats sont mitigés (40 % de succès, 60 % d'égalités), indiquant que, bien que partir d'un modèle généraliste puisse être avantageux par rapport à une variante affinée par instruction, les avantages sont moins prononcés. Ces tendances s'expliquent par plusieurs facteurs. Les gains limités du Groupe 3 montrent qu'un modèle ayant déjà appris des connaissances médicales en anglais tire peu de bénéfices d'une adaptation supplémentaire sur des données médicales françaises, et peut même en être affecté négativement. Le Groupe 2, quant à lui, montre que l'affinage par instruction peut améliorer certaines performances, mais qu'il peut aussi restreindre la flexibilité du modèle pour de nouvelles adaptations. Étant donné que le Groupe 1 (modèles basés sur Mistral) constitue le meilleur point de départ, nous concentrons notre analyse des stratégies d'adaptation sur ce groupe, où les améliorations sont significatives.

Modèle	Stratégie	Moyenne	FrClinicalQA	FrMedQA	K-QA
Mistral-7B-v0.1	Modèle de base	0,55	0,35	0,61	0,70
MedMistral-CPT-7B	CPT	0,55	0,56	0,59	0,51
MedMistral-CPT-SFT-7B	CPT+SFT	**0,67**	**0,65**	**0,64**	0,72
MedMistral-SFT-7B	SFT	0,52	0,20	0,62	**0,73**
Mistral-7B-Instruct-v0.1	Modèle de base	**0,67**	**0,65**	**0,67**	0,70
MedMistralInstruct-CPT-7B	CPT	0,60	0,51	0,65	0,63
MedMistralInstruct-CPT-SFT-7B	CPT+SFT	**0,67**	**0,65**	0,64	**0,71**
BioMistral-7B	Modèle de base	0,44	0,18	0,52	0,63
BioMistral-CPT-7B	CPT	0,47	0,45	0,60	0,35
BioMistral-CPT-SFT-7B	CPT+SFT	**0,66**	**0,64**	**0,63**	**0,71**
BioMistral-SFT-7B	SFT	0,57	0,39	**0,63**	0,70

TABLE 3 – Performances en *zero-shot* sur des QRO. Les scores correspondent au F1 BERTScore. Le meilleur modèle de chaque groupe est indiqué en **gras** et le meilleur modèle global est souligné.

4.2 Comparaison des stratégies d'adaptation

CPT : L'entraînement continu seul (MedMistral-CPT-7B) améliore les QCM traduits (+11,83 %) et natifs (+29,51 %), sans impact sur les QRO.

CPT+SFT : L'ajout du SFT renforce ces gains, atteignant 47,83 % (QCM traduits), 36,55 % (QCM natifs) et 0,67 en F1 BERTScore sur les QRO.

SFT : L'affinage supervisé direct (MedMistral-SFT-7B) donne de bons résultats avec 43,65 % pour les QCM traduits et 29,88 % pour les QCM natifs, bien qu'il présente une légère dégradation pour les QRO, qui passent de 0,55 à 0,52.

Analyse statistique des stratégies Les tests statistiques (Table 4) confirment la supériorité de CPT+SFT sur CPT-seul (67 % succès, 33 % égalités), mais montrent un avantage plus modéré sur SFT (40 % succès, 54 % égalités, 6 % pertes). CPT+SFT offre les meilleures performances, mais le SFT direct apparaît comme une alternative intéressante, notamment en termes d'efficacité computationnelle (voir Section 4.3).

4.3 Efficacité computationnelle et impact environnemental

Notre analyse des stratégies d'adaptation prend en compte non seulement les performances, mais aussi les coûts computationnels et l'impact environnemental. Nous évaluons ces facteurs à l'aide de trois métriques : le temps d'entraînement, les émissions de carbone (kgCO2e) et les coûts monétaires. L'approche CPT, bien qu'efficace, exige des ressources considérables. D'après la documentation technique du cluster Jean Zay [3], l'entraînement sur 7,4 Go de données médicales nécessite 32 GPU (NVIDIA H100 ou A100), générant environ 9 à 10 kgCO2e par adaptation. En revanche, le SFT traite seulement 36 Mo de données et utilise 1 à 2 GPU, réduisant les émissions à 2,5–2,6 kgCO2e. L'approche combinée CPT+SFT cumule les coûts des deux étapes, aboutissant à des émissions totales de 10 à 12 kgCO2e. Les coûts monétaires suivent une tendance similaire : l'entraînement CPT varie de 590 USD à 1 073 USD selon le type de GPU, tandis que le SFT direct ne coûte que 43 à 45 USD. Ces résultats soulignent que le SFT direct est une voie d'adaptation nettement plus économe en ressources, nécessitant seulement 25 % des ressources computationnelles et des émissions de carbone des approches CPT ou combinées. Des détails supplémentaires sur le temps d'entraînement, les émissions de carbone et les coûts monétaires sont fournis dans l'Annexe H.

4.4 Discussion générale

Nos résultats présentent un parallèle intéressant avec des travaux récents Jeong *et al.* (2024) qui remettent en question l'efficacité de l'adaptation au domaine médical. Alors que ces auteurs ont constaté que les modèles généraux en anglais possèdent déjà de solides compétences médicales, rendant un entraînement médical supplémentaire redondant en raison de leur exposition à des données médicales (PubMed) lors du pré-entraînement, nos résultats avec BioMistral (Groupe 3) montrent des gains tout aussi limités avec une adaptation médicale supplémentaire. Cependant, les résultats du Groupe 1 montrent que, en partant d'un modèle général et en l'adaptant avec des données médicales dans la langue cible (ici le français), potentiellement non incluses dans le pré-entraînement de Mistral, l'adaptation au domaine peut offrir des bénéfices significatifs (taux de succès de 87 % par rapport à la référence). Cela suggère que l'efficacité de l'adaptation au domaine dépend à la fois du point de départ et du fait que le modèle de base ait déjà été exposé à des données spécifiques au domaine lors du pré-entraînement, répondant directement à la question de recherche QR1 sur l'impact du choix du modèle de base sur le succès de l'adaptation.

De plus, bien que l'approche combinée CPT+SFT obtient les meilleures performances dans ce contexte, notre analyse montre que le SFT direct constitue une alternative assez convaincante lorsque

3. http://www.idris.fr/jean-zay/calcul-empreinte-carbone.html

Stratégie	Succès	Égalité	Perte
CPT+SFT vs. CPT	**0,67**	0,33	0
CPT+SFT vs. SFT	**0,4**	0,54	0,06

TABLE 4 – Comparaison statistique des stratégies d'adaptation. Les taux de Succès/Égalité/Perte comparent CPT+SFT à CPT-seul et à SFT-seul dans le Groupe 1.

Groupes	Succès	Égalité	Perte
Groupe 1 vs. Groupe 3	**0,73**	0,27	0
Groupe 1 vs. Groupe 2	**0,4**	0,6	0

TABLE 5 – Comparaison statistique entre les groupes pour déterminer le meilleur point de départ pour l'adaptation. Les taux de Succès/Égalité/Perte comparent les meilleurs modèles du Groupe 1 à ceux des Groupes 2 et 3.

les ressources computationnelles sont limitées. Avec seulement 25 % du coût computationnel et des émissions de carbone de CPT ou CPT+SFT, le SFT direct offre des améliorations substantielles pour une fraction des besoins en ressources. Cela met en évidence un compromis important entre les gains de performance et l'efficience, suggérant que, dans des scénarios aux ressources contraintes, le SFT direct peut être une stratégie d'adaptation pratique et efficiente, répondant ainsi à la question de recherche QR2 sur l'efficacité des stratégies d'adaptation.

Enfin, notre méthodologie d'évaluation met en lumière des considérations importantes sur l'évaluation des modèles adaptés au domaine. Les performances divergentes entre les tâches à choix multiples et les tâches ouvertes (dans le Groupe 3) soulèvent des questions cruciales sur la méthodologie d'évaluation, suggérant que les métriques actuelles, telles que BERTScore, pourraient ne pas distinguer efficacement les améliorations dans les connaissances factuelles par rapport aux capacités de génération de langage.

5 Conclusion

Cette étude examine différentes stratégies pour développer des modèles en langue médicale française, en fournissant des informations sur l'adaptation au domaine dans des contextes à faibles ressources. Nos résultats suggèrent que l'efficacité de l'adaptation au domaine dépend fortement de l'exposition préalable du modèle de base aux connaissances spécialisées.

Notre analyse révèle que, malgré les performances supérieures de la combinaison CPT+SFT sur l'ensemble des tâches, un SFT direct constitue une alternative solide, offrant de bons résultats avec des besoins computationnels nettement inférieurs. Cette conclusion présente un intérêt particulier pour une adaptation à un domaine cible plus économe en ressources.

Ces observations enrichissent notre compréhension de l'adaptation interlingue et proposent des recommandations pratiques pour développer des modèles de langue spécialisés en contextes contraints. Les travaux futurs devraient s'orienter vers la conception de stratégies d'adaptation plus efficaces et l'élaboration de méthodologies d'évaluation plus fiables pour mesurer les capacités propres à chaque domaine.

6 Limitations

Notre évaluation des stratégies d'adaptation comporte plusieurs limites. Premièrement, en raison de la rareté des ensembles de test en français natif dans le domaine médical, nous nous appuyons largement sur des jeux de données traduits. Bien que nous incluions des tests en français natif, une évaluation plus exhaustive nécessiterait davantage de jeux de données natifs couvrant diverses spécialités médicales.

Deuxièmement, l'évaluation des performances sur les QRO repose sur BERTScore, qui ne reflète pas nécessairement la précision médicale des réponses générées. Le développement de métriques spécialisées pour évaluer la génération en langage médical, notamment dans des langues autres que l'anglais, reste un défi majeur.

Troisièmement, si nous mettons en évidence l'efficience du SFT (par rapport au CPT) en termes de ressources informatiques, notre analyse ne prend pas en compte l'effort humain nécessaire à la création de jeux de données d'affinage supervisé de haute qualité. Cette considération est d'autant plus cruciale dans des environnements à faibles ressources, où la constitution de données d'instructions spécialisées peut être onéreuse.

Quatrièmement, nos expériences portent sur la famille de modèles Mistral-7B et pourraient aboutir à des conclusions différentes s'il s'agissait de modèles de tailles différentes ou entraînés sur un mélange de données très différent. De plus, les ensembles de données utilisés sont majoritairement axés sur la question-réponse, ce qui n'est pas représentatif de toute la variété d'usages potentiels des LLMs dans le domaine médical. Des tâches d'évaluation plus diversifiées pourraient nuancer nos conclusions.

Enfin, nos observations sur l'efficacité des stratégies d'adaptation sont spécifiques à la langue française et au domaine médical. Leur généralisation à d'autres domaines ou langues — notamment celles présentant des contraintes de ressources ou des caractéristiques linguistiques différentes — demande des investigations supplémentaires.

7 Remerciements

Ces travaux ont bénéficié des ressources de calcul et de stockage de l'IDRIS grâce aux allocations AD011015256 et A0161014871 attribuées par GENCI, sur les partitions V100, A100 et H100 du supercalculateur Jean Zay. Ce travail a également été soutenu financièrement par l'Agence Nationale de la Recherche (ANR) dans le cadre du projet MALADES (ANR-23-IAS1-0005).

Références

BEECHING E., FOURRIER C., HABIB N., HAN S., LAMBERT N., RAJANI N., SANSEVIERO O., TUNSTALL L. & WOLF T. (2023). Open llm leaderboard hugging face. *Récupérée mai*, **24**, 2024.

CHEN Z., CANO A. H., ROMANOU A., BONNET A., MATOBA K., SALVI F., PAGLIARDINI M., FAN S., KÖPF A., MOHTASHAMI A., SALLINEN A., SAKHAEIRAD A., SWAMY V., KRAWCZUK I., BAYAZIT D., MARMET A., MONTARIOL S., HARTLEY M.-A., JAGGI M. & BOSSELUT A. (2023). Meditron-70b : Scaling medical pretraining for large language models. arXiv : 2311.16079.

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

GARCÍA-FERRERO I., AGERRI R., ATUTXA A., CABRIO E., DE LA IGLESIA I., LAVELLI A., MAGNINI B., MOLINET B., RAMIREZ-ROMERO J., RIGAU G. *et al.* (2024). Medical mt5 : An open-source multilingual text-to-text llm for the medical domain. In *LREC-COLING 2024-2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*.

HAN T., ADAMS L. C., PAPAIOANNOU J.-M., GRUNDMANN P., OBERHAUSER T., LÖSER A., TRUHN D. & BRESSEM K. K. (2023). Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv e-prints*, p. arXiv–2304.

HENDRYCKS D., BURNS C., BASART S., ZOU A., MAZEIKA M., SONG D. & STEINHARDT J. (2020). Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

HU E. J., WALLIS P., ALLEN-ZHU Z., LI Y., WANG S., WANG L., CHEN W. *et al.* (2021). Lora : Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

HURST A., LERER A., GOUCHER A. P., PERELMAN A., RAMESH A., CLARK A., OSTROW A., WELIHINDA A., HAYES A., RADFORD A. *et al.* (2024). Gpt-4o system card. arXiv : 2410.21276.

JEONG D., GARG S., LIPTON Z. C. & OBERST M. (2024). Medical adaptation of large language and vision-language models : Are we making progress ? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 12143–12170.

JIANG A. Q., SABLAYROLLES A., MENSCH A., BAMFORD C., CHAPLOT D. S., DE LAS CASAS D., BRESSAND F., LENGYEL G., LAMPLE G., SAULNIER L., LAVAUD L. R., LACHAUX M.-A., STOCK P., SCAO T. L., LAVRIL T., WANG T., LACROIX T. & SAYED W. E. (2023). Mistral 7b. arXiv : 2310.06825.

JIN D., PAN E., OUFATTOLE N., WENG W.-H., FANG H. & SZOLOVITS P. (2021). What disease does this patient have ? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, **11**(14), 6421.

JIN Q., DHINGRA B., LIU Z., COHEN W. & LU X. (2019). Pubmedqa : A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 2567–2577.

KADDARI Z. & BOUCHENTOUF T. (2022). Frbmedqa : the first french biomedical question answering dataset. *IAES International Journal of Artificial Intelligence*, **11**(4), 1588.

KIM S., SUK J., LONGPRE S., LIN B. Y., SHIN J., WELLECK S., NEUBIG G., LEE M., LEE K. & SEO M. (2024). Prometheus 2 : An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 4334–4353.

KOPICZKO D. J., BLANKEVOORT T. & ASANO Y. M. (2024). Vera : Vector-based random matrix adaptation. arXiv : 2310.11454.

LABRAK Y., BAZOGE A., DUFOUR R., DAILLE B., GOURRAUD P.-A., MORIN E. & ROUVIER M. (2022). FrenchMedMCQA : A French Multiple-Choice Question Answering Dataset for Medical domain. In *LOUHI 2022*, Abou Dhabi, United Arab Emirates. HAL : hal-03824241.

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 *61th Annual Meeting of the Association for Computational Linguistics (ACL'23)*.

LABRAK Y., BAZOGE A., MORIN E., GOURRAUD P.-A., ROUVIER M. & DUFOUR R. (2024). Biomistral : A collection of open-source pretrained large language models for medical domains. In *62th Annual Meeting of the Association for Computational Linguistics (ACL'24)*.

LI Y., LI Z., ZHANG K., DAN R., JIANG S. & ZHANG Y. (2023). Chatdoctor : A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, **15**(6), e40895.

LIANG P., BOMMASANI R., LEE T., TSIPRAS D., SOYLU D., YASUNAGA M., ZHANG Y., NARAYANAN D., WU Y., KUMAR A. *et al.* (2022). Holistic evaluation of language models. *Transactions on Machine Learning Research*.

LIU S.-Y., WANG C.-Y., YIN H., MOLCHANOV P., WANG Y.-C. F., CHENG K.-T. & CHEN M.-H. (2024). Dora : weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, p. 32100–32121.

LOSHCHILOV I. & HUTTER F. (2019). Decoupled weight decay regularization. arXiv : 1711.0510.

MANES I., RONN N., COHEN D., BER R. I., HOROWITZ-KUGLER Z. & STANOVSKY G. (2024). K-qa : A real-world medical q&a benchmark. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, p. 277–294.

PAL A. & SANKARASUBBU M. (2024). Openbiollms : Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B. Hugging Face model repository.

PAL A., UMAPATHI L. K. & SANKARASUBBU M. (2022). Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, p. 248–260 : PMLR.

PIERI S., MULLAPPILLY S. S., KHAN F. S., ANWER R. M., KHAN S., BALDWIN T. & CHOLAKKAL H. (2024). Bimedix : Bilingual medical mixture of experts llm. In *Findings of the Association for Computational Linguistics : EMNLP 2024*, p. 16984–17002 : Association for Computational Linguistics. DOI : 10.18653/v1/2024.findings-emnlp.989.

QIU P., WU C., ZHANG X., LIN W., WANG H., ZHANG Y., WANG Y. & XIE W. (2024). Towards building multilingual language model for medicine. *Nature Communications*, **15**(1), 8384.

WANG X., CHEN N., CHEN J., HU Y., WANG Y., WU X., GAO A., WAN X., LI H. & WANG B. (2024). Apollo : An lightweight multilingual medical llm towards democratizing medical ai to 6b people. *CoRR*.

WU C., LIN W., ZHANG X., ZHANG Y., XIE W. & WANG Y. (2024). Pmc-llama : toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, **31**(9), 1833–1843.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2020). Bertscore : Evaluating text generation with bert. arXiv : 1904.09675.

A Ensembles de données d'évaluation

Type de question	Ensemble de données	Contexte	Taille
QCM traduits	MedQA (4 & 5 Options)	✗	1,273
	MedMCQA	✗	4,183
	PubMedQA	✓	500
	MMLU : Anatomy	✗	135
	MMLU : Clinical Knowledge	✗	265
	MMLU : College Biology	✗	144
	MMLU : College Medicine	✗	173
	MMLU : Professional Medicine	✗	272
	MMLU : Medical Genetics	✗	100
QCM natifs	FrBMedQA	✗	2,156
	FrMedMCQA	✗	183
QRO traduites	K-QA	✗	201
QRO natives	FrClinicalQA	✓	262
	FrMedQA	✗	81

TABLE 6 – Ensembles de données d'évaluation, classés par type de question, langue source (traduit ou natif), disponibilité du contexte et taille de l'ensemble de test.

B Description du corpus NACHOS

Le corpus NACHOS est un ensemble de données médicales en français, disponible en source ouverte, construit grâce à un *web crawling* et à la collecte de textes. Il comporte 7,4 Go de données, soit plus d'un milliard de mots (1 088 867 950 mots), issus de 24 sites web francophones de qualité (Labrak *et al.*, 2023).

Note : Les détails complets sur la compilation et le traitement du corpus figurent dans la publication originale (Labrak *et al.*, 2023).

B.1 Composition du corpus

Le corpus NACHOS couvre un large éventail de textes médicaux, incluant :

— Descriptions de maladies et de pathologies
— Informations sur les traitements et les médicaments
— Conseils généraux liés à la santé
— Comptes rendus officiels de réunions scientifiques
— Cas cliniques anonymisés
— Littérature scientifique
— Thèses
— Paires de traductions en français
— Cours universitaires de santé

B.2 Sources de données

Le corpus intègre des données provenant de multiples sources, avec les contributions les plus importantes issues de :

— HAL (638 508 261 mots)

— Haute Autorité de Santé (HAS) (113 394 539 mots)
— Notices de médicaments (74 770 229 mots)
— Extraction de sites web médicaux (60 561 495 mots)
— ANSES SAISINE (51 372 932 mots)
— Base de données publique des médicaments (BDPM) (48 302 695 mots)

B.3 Préparation du corpus

Les chercheurs ont appliqué plusieurs étapes de prétraitement :

1. Collecte de textes via *web scraping*, sources textuelles brutes et reconnaissance optique de caractères (OCR).
2. Segmentation des phrases selon des méthodes heuristiques.
3. Filtrage agressif pour éliminer les phrases trop courtes ou de faible qualité.
4. Classification de la langue à l'aide d'un classifieur personnalisé entraîné sur des corpus multilingues.

C Jeu de données d'entraînement SFT

Le jeu de données utilisé pour l'affinage supervisé (SFT) comprend 30 000 paires question-réponse provenant de trois catégories distinctes : contenu médical français natif, contenu médical anglais traduit en français et questions générées à partir de textes médicaux en français.

C.1 Contenu français natif

Nous avons sélectionné aléatoirement 10 000 paires question-réponse à partir de deux sources principales. La première est FrenchMedMCQA (Labrak *et al.*, 2022), un ensemble de 3 105 questions issues d'examens de spécialisation en pharmacie en France. Ces questions incluent des formats à réponse unique ou multiple, reflétant les conditions et standards réels des examens.

La seconde source comprend deux jeux de données complémentaires hébergés sur Hugging Face [4] : `mlabonne/medical-mqca-fr` [5] et `mlabonne/medical-cases-fr` [6]. Ces ressources proposent des questions à choix multiples et des études de cas cliniques provenant de bases d'examens médicaux français, couvrant un large éventail de spécialités (addictologie, gériatrie, neurologie, psychiatrie, etc.).

C.2 Contenu traduit

Nous avons extrait 10 000 paires question-réponse supplémentaires de ressources médicales en anglais, puis les avons traduites en français à l'aide de `jsontt` [7], un outil libre en ligne de commande

4. https://huggingface.co/
5. https://huggingface.co/datasets/mlabonne/medical-mqca-fr
6. https://huggingface.co/datasets/mlabonne/medical-cases-fr
7. https://github.com/mololab/json-translator

faisant appel à plusieurs services de traduction. Les sources incluent l'ensemble d'entraînement de MedQA (Jin *et al.*, 2021), composé de questions à choix multiples issues d'examens médicaux américains, ainsi que les *Medical Meadow Medical Flashcards* compilées par MedAplaca (Han *et al.*, 2023), couvrant des sujets médicaux fondamentaux (anatomie, physiologie, pathologie, pharmacologie, etc.).

C.3 Contenu généré

Les 10 000 paires finales ont été produites selon un processus en deux étapes.

Dans un premier temps, nous avons utilisé `Mistral-7b-instruct-v0.2` (Jiang *et al.*, 2023) pour générer des paires question-réponse à partir de contextes tirés de la sous-parégalité française du corpus Antidote (García-Ferrero *et al.*, 2024). Nous avons demandé au modèle de créer des questions-réponses à partir de chaque contexte médical en suivant le prompt présenté en Figure 3. Afin de garantir un format de sorégalité en JSON, nous avons eu recours à la bibliothèque `Outlines` [8], qui contraint la génération à respecter un schéma JSON prédéfini.

Prompt de génération

Vous êtes médecin et votre tâche consiste à fournir une paire de question-réponse en français à partir du contexte suivant :
Contexte : {{context}}
N'oubliez pas de répondre en français !

FIGURE 3 – Modèle d'instruction utilisé pour générer des paires question-réponse en français à partir d'un contexte donné.

Dans un second temps, la qualité des paires générées a été évaluée par trois grands modèles de langue : `Prometheus-7B-v2.0` (Kim *et al.*, 2024), `Meta-Llama-3-70B-Instruct` (Dubey *et al.*, 2024) et `GPT-4o` (Hurst *et al.*, 2024). Chacun a attribué une note sur une échelle de 1 à 5, basée sur la pertinence, l'exactitude et l'exhaustivité, conformément au modèle de la Figure 4. Seules les paires recevant la note de 4 ou 5 par tous les modèles évaluateurs ont été retenues pour l'entraînement, garantissant un jeu de données de haute qualité.

8. https://github.com/dottxt-ai/outlines

Prompt d'évaluation

You are a medical evaluator tasked with assessing question-answer pairs within a given context. Provide a score from 1 to 5 based on the provided score criteria.
[SCORE] : *(score from 1 to 5)*
Do not include any other opening, closing, or explanations.
Score criteria :

- ***Score 1 :*** *The question-answer pair is completely irrelevant or incorrect given the context. The answer has major factual errors.*
- ***Score 2 :*** *The question is somewhat relevant but the answer has significant inaccuracies or lacks important details from the context.*
- ***Score 3 :*** *The question is relevant and the answer is mostly accurate but contains some minor factual errors or omissions.*
- ***Score 4 :*** *The question is clear and relevant, and the answer is accurate based on the context with only very minor omissions.*
- ***Score 5 :*** *The question is clear, relevant, and the answer is completely accurate and comprehensive based on the given context.*

Remember, your score should consider both the relevance of the context to the medical domain and the accuracy of the question-answer pair.
Context : {{context}}
Question : {{question}}
Answer : {{answer}}
[SCORE] :

FIGURE 4 – Modèle d'instruction utilisé pour évaluer la pertinence et l'exactitude des paires question-réponse dans le domaine médical.

D Prompt-système pour la reformulation de FrBMedQA

Prompt-système

You are a medical question generation assistant. Given the follosuccèsg passage and a question based on it, transform the question into a valid multiple-choice question (MCQ). The MCQ should :

- *Focus on the placeholder by asking specifically about the information that corresponds to it.*
- *The question should not contain* `@placeholder`
- *The choices in the MCQ should be taken directly from the* `entiégalités_list` *and be formatted as options A, B, C, etc.*
- *The question should be phrased in a formal, clear, and precise manner, as a medical expert would phrase it.*
- *The MCQ should not contain any reference to the passage, such as "according to the passage" or "as stated in the passage". The question should be able to stand alone and should not explicitly refer to the passage.*
- *Provide one correct answer, which should correspond to the letter in the MCQ options.*
- *The MCQ should be written in French.*
- *Return the MCQ in JSON format*

FIGURE 5 – Prompt-système fourni à GPT-4o-mini pour générer des questions à choix multiples en français à partir de passages et de questions donnés.

E Prompt utilisé pour l'évaluation en *zero-shot* sur les tâches QCM et QRO

Nous avons utilisé un prompt standard (Figure 6) pour toutes les évaluations de questions à choix multiples (QCM), à l'exception de FrMedMCQA qui propose des questions avec plusieurs réponses correctes et requiert un prompt spécifique (Figure 7).

Prompt

Nous vous présentons une question scientifique, (un contexte) et (quatre/cinq) choix de réponse. Votre tâche est de trouver la réponse correcte en vous basant sur des faits scientifiques, vos connaissances et votre raisonnement (le contexte fourni). Générez uniquement l'une des lettres suivantes : A, B, C, D, (E). Chaque question n'a qu'une seule réponse. Les justifications ne sont pas permises.
Répondez à cette question (en vous basant sur le contexte) :
Contexte : *{{context}}*
Question : *{{question}}*
Choix :
{% for letter, option in options.items() %}
{{letter}} : {{option}}
{% endfor %}
Réponse :

FIGURE 6 – Modèle d'instruction pour les évaluations *zero-shot*.

Le prompt de base a été modifié dynamiquement selon les caractéristiques spécifiques de chaque corpus :

— Si le corpus comprenait un contexte, le terme *(un contexte)* était remplacé par le texte du contexte réel. Pour les corpus sans contexte, cette parégalité était supprimée.
— Le nombre de choix de réponse variait en fonction du corpus, avec *(quatre/cinq)* remplacé par "quatre" ou "cinq" selon le cas.
— La lettre (E) était incluse pour les corpus à cinq options et omise dans le cas contraire.
— *(en vous basant sur le contexte)* : cette mention était incluse pour les corpus fournissant un contexte et supprimée sinon.

Prompt

Nous vous présentons une question scientifique suivie de plusieurs choix de réponse. Votre tâche est de sélectionner la ou les lettres correspondant aux réponses correctes, en vous basant sur des faits scientifiques, vos connaissances et votre raisonnement. Générez uniquement les lettres correspondant aux réponses correctes (par exemple : A C D). Chaque question peut avoir une ou plusieurs réponses correctes. Les justifications ne sont pas permises.
Question : *{{question}}*
Choix :
{% for letter, option in options.items() %}
{{letter}} : {{option}}
{% endfor %}
Réponse :

FIGURE 7 – Modèle d'instruction pour l'évaluation *zero-shot* utilisé dans FrMedMCQA.

Pour les évaluations de questions à réponses ouvertes (QRO), nous avons adopté un prompt standard

(Figure 8) pour tous les jeux de données QRO, à l'exception de FrClinicalQA. Dans ce dernier, les questions sont liées à un cas clinique et se succèdent, le gabarit (Figure 9) inclut donc le cas clinique, les questions précédentes et la question courante. Cette structure maintient la continuité contextuelle entre les questions d'un même cas clinique.

Prompt

Veuillez lire l'instruction médicale ci-dessous et fournir une réponse adaptée à la situation décrite. Votre tâche est de répondre correctement en vous basant sur des faits scientifiques et vos connaissances. Répondez uniquement à la question posée de manière brève et concise. Faites des phrases courtes contenant la réponse, évitez les informations non essenégalitélles et concentrez-vous sur les éléments cruciaux pour une réponse efficace et pertinente.
Instruction : *{{question}}*
Réponse :

FIGURE 8 – Modèle d'instruction pour l'évaluation *zero-shot* utilisé dans les jeux de données QRO.

Prompt

Vous allez lire un cas clinique suivi de plusieurs questions liées. Votre tâche est de répondre correctement à la dernière question en utilisant uniquement le contexte clinique fourni et les questions précédentes. N'incluez pas d'informations non pertinentes ou de réponses aux questions précédentes. Répondez de manière brève et concise à la question posée, en vous basant sur le cas clinique et les questions précédentes comme contexte.
Cas Clinique : *{{clinical_case}}*
{{previous_questions}}
Répondez uniquement à la question suivante, en utilisant le cas clinique et les questions précédentes comme contexte, sans inclure de réponses précédentes.
Question : *{{question}}*
Réponse :

FIGURE 9 – Modèle d'instruction pour l'évaluation *zero-shot* utilisé dans le jeu de données FrClinicalQA.

F Les hyperparamètres du SFT

Paramètre	Valeur
Rang	16
LoRA Aplha	16
LoRA Dropout	0.05
Taux d'apprentissage	2e-05
Taille du batch d'entraînement	4
Taille du batch d'évaluation	8
Seed	42
Nombre de GPUs	1 NVIDIA H100 80GB Or 2 NVIDIA L40 48GB
Étapes d'accumulation du gradient	2
Optimizeur	AdamW
Ordonnanceur	Cosinus
Nombre d'époques	10
Modules cibles	QKVOGUD

TABLE 7 – Hyperparamètres pour l'entraînement SFT

G Résultats de l'analyse de la significativité statistique

Les résultats de l'analyse de la significativité statistique sont présentés pour différents contextes d'évaluation : les Tables 9, 10 et 11 indiquent respectivement les taux de victoire, d'égalité et de défaite (*succès/égalité/perte*) pour les QCM traduits, les QCM natifs et les questions à réponses ouvertes (QRO). Chaque taux exprime la proportion d'ensembles de données pour lesquels un modèle présente une amélioration significative (*succès*), aucune différence significative (*égalité*) ou une dégradation significative (*perte*) par rapport à son modèle de base.

H Ressources de calcul et impact environnemental

Le Tableau 8 présente les ressources de calcul nécessaires pour chaque stratégie d'adaptation ainsi que leur impact environnemental. Nous y indiquons le temps d'entraînement, le nombre de GPU requis, les émissions de carbone (*kgCO2e*) et les coûts associés en dollars américains (USD). Les émissions de carbone ont été estimées à partir de la consommation d'énergie des différents types de GPU et de la durée d'entraînement.

Modèle	Stratégie	Taille des données (GB)	Type de GPU	Memoire par GPU (GB)	Nombre de GPUs	Temps d'entraînement (heures)	Émissions (KgCO2e)	Coût (USD)
MedMistral-CPT-7B	CPT	7.4	NVIDIA H100	80	32	12	9.86	643.64
MedMistral-CPT-SFT-7B	CPT+SFT	7.4+ 0.036	NVIDIA H100/A100	80	32 + 1	12 + 75	9.86 + 1.92	643.64 + 63.22
MedMistral-SFT-7B	SFT	0.036	NVIDIA A40	48	2	53	2.62	44.57
MedMistralInstruct-CPT-7B	CPT	7.4	NVIDIA A100	80	32	40	32.89	1072.74
MedMistralInstruct-CPT-SFT-7B	CPT+SFT	7.4+ 0.036	NVIDIA A100/H100	80	32 + 1	40 + 42	32.89 + 1.07	1072.74 + 70.48
BioMistral-CPT-7B	CPT	7.4	NVIDIA H100	80	32	11	9.04	589.75
BioMistral-CPT-SFT-7B	CPT+SFT	7.4+ 0.036	NVIDIA H100	80	32 + 1	11 + 42	9.04 + 1.07	589.75 + 70.48
BioMistral-SFT-7B	SFT	0.036	NVIDIA L40	48	2	52	2.57	43.53

TABLE 8 – Exigences en ressources et impact environnemental pour différentes stratégies d'adaptation. Les temps d'entraînement et les coûts sont indiqués pour chaque Stratégie d'adaptation et chaque modèle de base. Les émissions de carbone sont estimées en fonction de la consommation d'énergie des GPU durant l'entraînement.

			Moyenne			PubMedQA			MedQA 4 Options			MedQA 5 Options			MedMCQA			MMLU																	
																		Clinical Knowledge			Medical Genetics			Anatomy			Pro. Medicine			College Biology			College Medicine		
Modèle	Modèle de base	Stratégie	Succès	Egalité	Perte	Succès	Egalité	Perte	Succès	Egalité	Perte	Succès	Egalité	Perte	Succès	Egalité	Perte	Succès	Egalité	Perte	Succès	Egalité	Perte	Succès	Egalité	Perte	Succès	Egalité	Perte	Succès	Egalité	Perte	Succès	Egalité	Perte
MedMistral-CPT-7B	Mistral-7B-v0.1	CPT	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0
MedMistral-CPT-SFT-7B	Mistral-7B-v0.1	CPT+SFT	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0
MedMistral-CPT-SFT-7B	MedMistral-CPT-7B	CPT+SFT	**0.5**	0.5	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	0	1	0	0	1	0	0	1	0	**1**	0	0	0	1	0	0	1	0
Med-Mistral-7B-chat	Mistral-7B-v0.1	SFT	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0
MedMistralInstruct-CPT-7B	Mistral-7B-Instruct-v0.1	CPT	**0.3**	0.6	0.1	0	0	**1**	**1**	0	0	**1**	0	0	**1**	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0
MedMistralInstruct-CPT-SFT-7B	MedMistralInstruct-CPT-7B	CPT+SFT	**0.1**	0.9	0	**1**	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0
MedMistralInstruct-CPT-SFT-7B	Mistral-7B-Instruct-v0.1	CPT+SFT	**0.3**	0.7	0	0	1	0	**1**	0	0	**1**	0	0	**1**	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0
BioMistral-CPT-7B	BioMistral-7B	CPT	0	0.9	**0.1**	0	0	**1**	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0
BioMistral-CPT-SFT-7B	BioMistral-7B	CPT+SFT	0	0.8	**0.2**	0	0	**1**	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0
BioMistral-CPT-SFT-7B	BioMistral-CPT-7B	CPT+SFT	**0.2**	0.8	0	**1**	0	0	0	1	0	0	1	0	**1**	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0
BioMistral-SFT-7B	BioMistral-7B	SFT	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0

TABLE 9 – Les taux de Succès/Egalité/Perte en *zero-shot* pour toutes les comparaisons médicales sur les QCM traduits. Pour chaque modèle médical, le taux de Succès est mis en **gras** s'il est supérieur au taux de perte par rapport à son modèle de base, et inversement.

			Moyenne			FrBMedQA			FrMedMCQA		
Modèle	Modèle de base	Stratégie	Succès	Egalité	Perte	Succès	Egalité	Perte	Succès	Egalité	Perte
MedMistral-CPT-7B	**Mistral-7B-v0.1**	CPT	**1**	0	0	**1**	0	0	**1**	0	0
MedMistral-CPT-SFT-7B	**Mistral-7B-v0.1**	CPT+SFT	**1**	0	0	**1**	0	0	**1**	0	0
MedMistral-CPT-SFT-7B	**MedMistral-CPT-7B**	CPT+SFT	0	1	0	0	1	0	0	1	0
Med-Mistral-7B-chat	**Mistral-7B-v0.1**	SFT	**1**	0	0	**1**	0	0	**1**	0	0
MedMistralInstruct-CPT-7B	**Mistral-7B-Instruct-v0.1**	CPT	**0.5**	0.5	0	**1**	0	0	0	1	0
MedMistralInstruct-CPT-SFT-7B	**MedMistralInstruct-CPT-7B**	CPT+SFT	0	1	0	0	1	0	0	1	0
MedMistralInstruct-CPT-SFT-7B	**Mistral-7B-Instruct-v0.1**	CPT+SFT	**0.5**	0.5	0	**1**	0	0	0	1	0
BioMistral-CPT-7B	**BioMistral-7B**	CPT	0	1	0	0	1	0	0	1	0
BioMistral-CPT-SFT-7B	**BioMistral-7B**	CPT+SFT	0	1	0	0	1	0	0	1	0
BioMistral-CPT-SFT-7B	**BioMistral-CPT-7B**	CPT+SFT	0	1	0	0	1	0	0	1	0
BioMistral-SFT-7B	**BioMistral-7B**	SFT	0	1	0	0	1	0	0	1	0

TABLE 10 – Les taux de Succès/Égalité/Perte en 0-shot pour toutes les comparaisons médicales sur 2 ensembles de QCM en français natif. Pour chaque modèle médical, le taux de Succès est mis en **gras** lorsqu'il l'emporte plus souvent qu'il ne perd face à son modèle de base, et inversement.

			Moyenne			FrClinicalQA			FrMedQA			K-QA		
Modèle	**Modèle de base**	**Stratégie**	**Succès**	**Egalité**	**Perte**	**Succès**	**Egalité**	**Perte**	**Succès**	**Egalité**	**Perte**	**Succès**	**Egalité**	**Perte**
MedMistral-CPT-7B	**Mistral-7B-v0.1**	CPT	0.33	0.33	0.33	**1**	0	0	0	1	0	0	0	**1**
MedMistral-CPT-SFT-7B	**Mistral-7B-v0.1**	CPT+SFT	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0
MedMistral-CPT-SFT-7B	**MedMistral-CPT-7B**	CPT+SFT	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0
Med-Mistral-7B-chat	**Mistral-7B-v0.1**	SFT	0.33	0.33	0.33	0	0	**1**	0	1	0	**1**	0	0
MedMistralInstruct-CPT-7B	**Mistral-7B-Instruct-v0.1**	CPT	0	0.33	**0.67**	0	0	**1**	0	**1**	0	0	0	**1**
MedMistralInstruct-CPT-SFT-7B	**MedMistralInstruct-CPT-7B**	CPT+SFT	**0.67**	0.33	0	**1**	0	0	0	**1**	0	1	0	0
MedMistralInstruct-CPT-SFT-7B	**Mistral-7B-Instruct-v0.1**	CPT+SFT	0	0.67	**0.33**	0	1	0	0	0	**1**	0	1	0
BioMistral-CPT-7B	**BioMistral-7B**	CPT	0.33	0.33	0.33	**1**	0	0	0	1	0	0	0	**1**
BioMistral-CPT-SFT-7B	**BioMistral-7B**	CPT+SFT	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0
BioMistral-CPT-SFT-7B	**BioMistral-CPT-7B**	CPT+SFT	**0.67**	0.33	0	**1**	0	0	0	1	0	**1**	0	0
BioMistral-SFT-7B	**BioMistral-7B**	SFT	**0.67**	0.33	0	**1**	0	0	0	1	0	**1**	0	0

TABLE 11 – Les taux de Succès/Égalité/Perte en 0-shot pour toutes les comparaisons médicales sur 3 ensembles de QRO. Pour chaque Modèle médical, le taux de Succès est mis en **gras** lorsqu'il l'emporte plus souvent qu'il ne perd face à son modèle de base, et inversement.