# Amélioration des modèles basés sur Transformers pour le traitement des nombres dans les documents médicaux

par

Boammani Aser Lompo

MÉMOIRE PAR ARTICLES PRÉSENTÉ À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE COMME EXIGENCE PARTIELLE À L'OBTENTION DE LA MAÎTRISE AVEC MÉMOIRE EN TECHNOLOGIE DE LA SANTÉ
M. Sc. A.

MONTRÉAL, LE 12 MAI 2025

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

**PRÉSENTATION DU JURY**

CETTE THÈSE A ÉTÉ ÉVALUÉE

PAR UN JURY COMPOSÉ DE:

Mme. Rita Noumeir, directrice de mémoire
Département de génie électrique, École de technologie supérieure

M. Philippe Jouvet, codirecteur
Unité des soins intensifs pédiatriques, CHU Sainte-Justine

Mme. Sylvie Ratté, présidente du jury
Département de génie logiciel et TI, École de technologie supérieure

M. Mohamad Forouzanfar, membre du jury
Département de génie des systèmes, École de technologie supérieure

IL A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 6 MAI 2025

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

## REMERCIEMENTS

Je tiens à exprimer ma profonde gratitude à toutes les personnes qui ont joué un rôle important dans la réalisation de mon projet de maîtrise. Mes remerciements les plus sincères vont à mes estimés directeurs, Mme Rita Noumeir et M. Philippe Jouvet, pour leur soutien constant, leurs conseils et leurs précieuses idées tout au long de ce voyage à la fois difficile et gratifiant.

Je suis extrêmement reconnaissant au Dr Thanh-Dung Le pour son assistance et son soutien dévoués pendant mes études de maîtrise, ce qui a grandement enrichi la profondeur de ma recherche. J'adresse également mes remerciements à l'ensemble du groupe LATIS pour avoir encouragé un environnement de coopération et d'échange de connaissances. Cette équipe exceptionnelle possède non seulement une richesse de savoirs, mais est également composée d'individus bienveillants et généreux de leurs idées.

Enfin, je souhaite exprimer ma plus profonde gratitude à ma famille et à mes amis. Leurs encouragements constants, leur compréhension et leur soutien continu ont été mes piliers de force. Les défis rencontrés étaient significatifs, mais leur soutien n'a pas seulement été nécessaire, il a été également profondément apprécié. Cette expérience a été un voyage difficile pour moi, et j'ai énormément appris et grandi tout au long du processus.

Je tiens également à remercier l'Hôpital CHU Sainte-Justine pour avoir fourni les notes médicales utilisées dans cette étude, ainsi que le Dr Jérôme Rambaud et le Dr Guillaume Sans qui les ont annotées. J'exprime aussi ma gratitude au physiothérapeute Kevin Albert pour avoir réalisé la revue de littérature sur les références médicales.

Ce travail a été soutenu en partie par le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG), en partie par l'Institut de valorisation des données de l'Université de Montréal (IVADO), et en partie par le Fonds de recherche du Québec - Santé (FRQS).

Merci à tous ceux qui ont joué un rôle, aussi petit soit-il, dans la réalisation de cet effort académique.

**Amélioration des modèles basés sur Transformers pour le traitement des nombres dans les documents médicaux**

Boammani Aser Lompo

## RÉSUMÉ

Amélioration des modèles basés sur les Transformers pour le traitement des nombres dans les documents médicaux

Les documents médicaux créés par les professionnels de santé lors de l'admission des patients regorgent de détails essentiels pour le diagnostic. Cependant, leur potentiel n'est pas pleinement exploité en raison d'obstacles tels que le langage médical complexe, la compréhension insuffisante des données numériques médicales par les modèles du Language à la pointe de la technologie, et les limitations imposées par de petits jeux de données annotés.

Ces recherches portent sur la classification des valeurs numériques extraites de documents médicaux en sept catégories physiologiques distinctes, en s'appuyant sur CamemBERT-bio, un modèle Transformer. Bien que certaines études antérieures aient suggéré que les modèles Transformers pouvaient être moins performants que les approches classiques en Traitement Automatique du Langage (TAL) pour ce type de tâche, des travaux plus récents montrent qu'une augmentation significative du nombre de paramètres permet aux modèles de langage d'acquérir de nouvelles compétences, améliorant ainsi leurs performances. Cependant, ces modèles de très grande envergure, appelés LLMs, nécessitent des ressources computationnelles considérables.

Dans cette étude, nous nous concentrons sur des modèles de langage de taille intermédiaire basés sur l'architecture Transformer, comme CamemBERT-bio. Pour en optimiser les performances, notre approche repose sur deux phases.

Tout d'abord, nous introduisons deux innovations principales : l'intégration d'embeddings de mots-clés dans le modèle et l'adoption d'une stratégie agnostique aux nombres en excluant toutes les données numériques du texte. La mise en œuvre de techniques d'embedding de mots-clés affine les mécanismes d'attention, tandis que l'utilisation d'un jeu de données « aveugle aux nombres » vise à renforcer l'apprentissage centré sur le contexte. Un autre élément clé de notre recherche est de déterminer la criticité des données numériques extraites. Pour ce faire, nous avons utilisé une approche simple consistant à vérifier si la valeur se situe dans les plages standards établies. **Résultats** : Nos résultats sont encourageants, montrant des améliorations substantielles de l'efficacité de CamemBERT-bio, surpassant les méthodes conventionnelles avec un score F1 de 0,89. Cela représente une augmentation de plus de 20 % par rapport au score F1 de 0,73 des approches traditionnelles, avec un écart de seulement 6 % par rapport au score F1 de GPT-4, un modèle à la pointe de la technologie et plusieurs centaines de fois plus grand que le nôtre.

Dans la deuxième phase, en s'appuyant sur les résultats antérieurs qui révèlent les limitations potentielles des modèles basés sur les transformateurs, nous examinons deux stratégies : le

fine-tuning de CamemBERT-bio sur un petit jeu de données médicales avec l'intégration de l'Embedding de Label pour l'Attention de Soi (LESA), et la combinaison de LESA avec des techniques d'amélioration supplémentaires telles que Xval. Étant donné que CamemBERT-bio est déjà pré-entraîné sur un grand jeu de données médicales, la première approche vise à mettre à jour son encodeur avec la nouvelle technique d'embeddings de labels, tandis que la deuxième approche cherche à développer plusieurs représentations des nombres (contextuelles et basées sur la magnitude) pour obtenir des embeddings numériques plus robustes. **Résultats** : Comme prévu, le fine-tuning du CamemBERT-bio standard sur notre petit jeu de données médicales n'a pas amélioré les scores F1. Cependant, des améliorations significatives ont été observées avec CamemBERT-bio + LESA, entraînant une augmentation de plus de 15 %. Des améliorations comparables ont été observées en combinant LESA avec Xval, dépassant les méthodes classiques et réduisant encore l'écart de performance avec GPT-4, le modèle de référence à la pointe de la technologie.

En résumé, notre travail a introduit diverses méthodes pour traiter les données numériques, qui sont également applicables à d'autres modalités. Nous illustrons comment ces nouvelles approches peuvent aider les modèles basés sur les transformers à fournir des performances robustes sur les tâches de classification, même lorsqu'ils traitent de petits jeux de données.

**Mots-clés:** Traitement automatique du langage clinique, Classification des valeurs numériques, Entraînement des modèles linguistiques, Embedding de labels pour l'attention de soi, Xval, Unité de soins intensifs pédiatriques, Patients en état critique

# Improvement of Transformer-Based Models for Processing Numbers in Medical Documents

Boammani Aser Lompo

## ABSTRACT

Medical records created by healthcare professionals upon patient admission are rich in details critical for diagnosis. Yet, their potential is not fully realized due to obstacles such as complex medical language, inadequate comprehension of medical numerical data by state-of-the-art Language Models (LMs), and the limitations imposed by small annotated training datasets.

This research focuses on classifying numerical values extracted from medical documents into seven distinct physiological categories using CamemBERT-bio, a Transformer-based model. While earlier studies suggested that Transformer models might underperform compared to traditional Natural Language Processing (NLP) approaches for such tasks, more recent findings indicate that a significant increase in the number of parameters enables language models to develop new capabilities, thereby improving their performance. However, these large-scale models, known as LLMs, require substantial computational and memory resources.

In this study, we focus on medium-sized language models based on the Transformer architecture, such as CamemBERT-bio. To enhance its performance, our approach consists of two phases.

First, we introduce two main innovations : integrating keyword embeddings into the model and adopting a number-agnostic strategy by excluding all numerical data from the text. The implementation of label embedding techniques refines the attention mechanisms, while using a 'numerical-blind' dataset aims to bolster context-centric learning. Another key component of our research is determining the criticality of extracted numerical data. To achieve this, we utilized a straightforward approach that involves verifying if the value falls within established standard ranges. **Results** : Our findings are encouraging, showing substantial improvements in the effectiveness of CamemBERT-bio, surpassing conventional methods with an F1 score of 0.89. This represents an increase of over 20% compared to the F1 score of 0.73 achieved by traditional approaches, with only a 6% gap from the F1 score of GPT-4—a state-of-the-art model that is several hundred times larger than ours.

In the second phase, building on prior findings that reveal potential limitations of transformer-based models, we examine two strategies : fine-tuning CamemBERT-bio on a small medical dataset with the integration of Label Embedding for Self-Attention (LESA), and combining LESA with additional enhancement techniques such as Xval. Given that CamemBERT-bio is already pre-trained on a large medical dataset, the first approach aims to update its encoder with newly added label embeddings, while the second approach seeks to develop multiple representations of numbers (contextual and magnitude-based) for more robust number embeddings. **Results** : As anticipated, fine-tuning the standard CamemBERT-bio on our small medical dataset did not improve F1 scores. However, similar improvements were observed when combining LESA with

Xval, surpassing conventional methods and further narrowing the performance gap with GPT-4, the state-of-the-art model.

In summary, our work introduced various methods for handling numerical data, which are also applicable to other modalities. We illustrate how these novel approaches can support transformer-based models in delivering robust performance on classification tasks, even when dealing with small datasets.

**Keywords:** Clinical Natural Language Processing, Numerical values classification, Language models training, Label Embedding for Self Attention, Xval, Pediatric Intensive Care Unit, critically ill patients

## TABLE DES MATIÈRES

Page

## LISTE DES TABLEAUX

## LISTE DES FIGURES

Page

## LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

| | |
|---|---|
| CDSS | Clinical decision support system |
| CHUSJ | Centre Hospitalier Universitaire Sainte-Justine |
| ETS | École de Technologie Supérieure |
| LSTM | Long short-term memory |
| MSE | Mean square error |
| NER | Name Entity Recognition |
| TALN | Traitement Automatique du Langage Naturel |
| USIP | Unité de Soins Intensifs Pédiatriques |
| PLM | Pretrained Language Model |
| LLM | Large Language Model |
| RNN | Recurrent neural network |

# INTRODUCTION

Les applications de l'apprentissage automatique dans les contextes cliniques représentent un domaine de recherche en pleine évolution et dynamique. Ce domaine promet d'équiper les professionnels de santé avec des technologies avancées qui utilisent efficacement les ressources. Ces avancées visent à démocratiser l'accès à des soins de santé de haute qualité, indépendamment des contraintes temporelles et spatiales (Sutton *et al.*, 2020). Les progrès récents en Traitement Automatique du Langage Naturel (TALN) ont permis aux réseaux neuronaux d'analyser de vastes corpus de textes, facilitant l'extraction d'informations pertinentes et bénéfiques. Ces nouveaux outils technologiques ont le potentiel de contribuer de manière significative aux pratiques quotidiennes des professionnels de la santé, en aidant à l'exploration et à l'analyse des dossiers médicaux des patients.

Dans la majorité des unités de soins intensifs (USI), une quantité importante d'informations médicales est documentée quotidiennement, soit sous forme de notes textuelles, soit sous forme de données numériques générées par des machines. Les données générées par les machines, structurées sous forme de tableaux, peuvent être facilement intégrées dans divers algorithmes de soutien à la décision clinique. Cependant, comme le souligne Sutton *et al.* (2020), l'utilisation des notes textuelles est souvent moins efficace, principalement en raison de leur format non structuré et de l'utilisation fréquente de jargon médical, dense en informations et comprenant parfois des phrases incomplètes. En effet, la présence d'abréviations, de fautes d'orthographe et d'autres types d'erreurs d'entrée courants dans le domaine médical introduit une incertitude considérable, rendant inefficaces les approches systématiques comme les algorithmes au cas par cas. Un autre défi majeur dans le traitement des textes médicaux est la compréhension des données numériques. Les chiffres sont omniprésents dans les documents médicaux, apparaissant sous diverses formes telles que les mesures physiologiques, les séquences de codage ADN, les occurrences d'événements (par exemple, nombre de naissances, défaillances cardiaques), les durées (par exemple, durée de grossesse, durée de traitement), et les codes de protocoles

médicaux (par exemple, G1P2). Dans le cadre des Systèmes d'Aide à la Décision Clinique (SADC), il est crucial de reconnaître et d'interpréter la signification de ces chiffres dans leur contexte spécifique pour déterminer s'ils indiquent un problème potentiel. Par exemple, dans le texte "Patient âgé de 12 ans, rythme cardiaque de 130 bpm," il est essentiel d'identifier "12" comme l'âge du patient et "130" comme le rythme cardiaque, puis de comprendre qu'un rythme cardiaque de "130" est problématique pour un patient de 12 ans. De plus, les valeurs numériques dans les ensembles de données médicales couvrent généralement une large gamme. Par exemple, l'ensemble de données fourni par le Centre Hospitalier Universitaire Sainte-Justine (CHUSJ) pour notre étude comprend des valeurs numériques allant de $10^{-4}$ à $10^{5}$. Par conséquent, une gestion soigneuse des gradients est essentielle, car une seule valeur aberrante importante peut nuire à la performance globale du modèle.

Les méthodologies d'apprentissage automatique traditionnelles dans ce contexte ont principalement impliqué des modèles de représentation de mots non contextuels, complétés par des couches spécifiques à la tâche (Le, Noumeir, Rambaud, Sans & Jouvet, 2022a). Les avancées ultérieures ont introduit des Réseaux Neuronaux Récurrents (RNN) capables de représentations textuelles contextuelles. Bien que ces méthodes aient obtenu un succès notable dans plusieurs tâches de classification de textes Mascio *et al.* (2020), leur efficacité est en grande partie attribuée au fait que l'ensemble de données d'évaluation est spécifiquement personnalisé pour convenir à leurs capacités. Les avancées actuelles en TALN impliquent principalement des modèles basés sur les Transformers (Vaswani *et al.*, 2017) tels que BERT et GPT (Devlin, Chang, Lee & Toutanova, 2018; Brown *et al.*, 2020). Ces Grands Modèles de Langage Préentraînés (PLM) s'allègent en grande partie du prérequis de l'adaptation des données car ils nécessitent un prétraitement minimal. Cependant, dans des scénarios spécifiques comme la classification de textes médicaux avec des ensembles de données d'entraînement petits et déséquilibrés, ils peuvent ne pas surpasser les RNN (Mascio *et al.*, 2020; Chen, Stewart, Sun, Ng & Yan, 2019).

Ce travail contribue à un projet visant à utiliser des Grands Modèles de Langage Préentraînés (PLM) pour le diagnostic de l'insuffisance cardiaque. Le **National Heart, Lung, and Blood Institute (2023)** définit l'insuffisance cardiaque comme une condition où le cœur ne parvient pas à pomper suffisamment de sang pour répondre aux besoins de l'organisme. Cette insuffisance peut découler de l'incapacité du cœur à se remplir adéquatement de sang ou de sa capacité affaiblie à pomper efficacement. La manifestation de l'insuffisance cardiaque peut varier, affectant des paramètres physiologiques tels que la pression artérielle, les gradients ventriculaires et les fractions d'éjection. De plus, les facteurs génétiques et les conditions de santé antérieures du patient influencent la gravité de l'insuffisance cardiaque, nécessitant une approche personnalisée pour chaque cas. En termes simples, il est essentiel d'interpréter les informations médicales du patient dans le cadre de ses antécédents médicaux. Dans notre recherche, nous nous attaquons à la catégorisation des valeurs numériques obtenues à partir de notes médicales en une des sept catégories physiologiques prédéterminées en utilisant CamemBERT-bio (Touchent, Romary & de La Clergerie, 2023). Ces paramètres comprennent divers indicateurs de l'insuffisance cardiaque tels que la fraction d'éjection et de raccourcissement, la saturation en oxygène, le rythme cardiaque, le diamètre de l'artère pulmonaire, le gradient ventriculaire, et la taille du défaut septal auriculo-ventriculaire. Étant donné qu'une partie significative de nos patients sont des nouveau-nés, nous incluons également le score APGAR comme indicateur clé de la détresse vitale. Ensuite, nous évaluons si ces valeurs numériques indiquent une condition médicale critique.

Plus simplement, en entraînant CamemBERT-bio sur un petit ensemble de données déséquilibré, nous visons à obtenir le résultat suivant : **Note médicale d'entrée** :

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

**Sortie attendue** :

Tableau 0.1 Résultat souhaitée suite à la note médicale d'entrée

| Valeur | Attributs | Unité | Critique |
|---|---|---|---|
| 14/08 | Divers (date) | | Non |
| 50 – 60 | gradient VD-VG AP | **mmHg** | Non |
| 80 – 85 | saturation en oxygène | % | Non |

Enfin, nous explorons plusieurs axes d'évolution pour ce projet en nous intéressant aux différentes méthodes d'entraînement des modèles de langage pour le raisonnement. Cette approche constitue une prolongation naturelle des objectifs précédemment évoqués, car elle applique la représentation numérique développée à l'assistance au diagnostic de l'insuffisance cardiaque.

# CHAPITRE 1

# REVUE DE LITTÉRATURE

## 1.1 Classification de texte en TALN

La classification de texte dans le domaine de la santé a été le centre d'attention de divers travaux en apprentissage automatique (Ezen-Can, 2020; Mascio *et al.*, 2020; Le *et al.*, 2022a). Traditionnellement, ces approches emploient une procédure en deux phases. La phase initiale consiste à analyser les données brutes pour extraire des caractéristiques pertinentes, comme décrit dans Agarwal, Rahman, St-Charles, Prince & Kahou (2023) et Zhou, Duan, Liu & Shum (2020). L'efficacité des modèles d'apprentissage profond dans la classification de texte est largement reconnue comme étant dépendante de la qualité de ces caractéristiques extraites, communément appelées embeddings. La deuxième phase consiste au passage de ces embeddings par un classificateur de réseau neuronal pour la prédiction. Parmi les techniques d'extraction de caractéristiques les plus en vue, on trouve word2vec Mikolov, Chen, Corrado & Dean (2013), GloVe Pennington, Socher & Manning (2014), et ELMo Peters *et al.* (2018). Dans Cui *et al.* (2019), les auteurs ont abordé le défi de la classification de texte en utilisant une approche basée sur des règles qui emploie des expressions régulières. Cependant, le succès de cette méthode dépend fortement de la clarté du texte, en particulier de l'uniformité des abréviations et de l'absence de bruit textuel non représentatif pouvant fausser le générateur d'expressions régulières. Récemment, il y a eu un changement significatif dans la communauté TALN vers les modèles basés sur les transformateurs, avec BERT Devlin *et al.* (2018) étant un exemple notable. BioBert, introduit par Lee *et al.* (2019), représente une adaptation de BERT pré-entraîné sur un corpus biomédical conséquent, produisant des résultats à la pointe dans diverses tâches de classification en santé. En 2023, Touchent *et al.* (2023); Labrak *et al.* (2023) ont encore élargi ce concept avec l'introduction de CamemBERT-bio et DrBERT, deux variantes françaises de BioBERT, basées sur CamemBERT Martin *et al.* (2020) et pré-entraînées sur un corpus médical français.

## 1.2 Performances des récents LLMs

Au cours des dernières années, les meilleures performances en traitement automatique du langage naturel (TALN) ont été principalement atteintes grâce au déploiement de modèles de langage à grande échelle basés sur les Transformers (LLMs), tels que Llama-3, GPT-3.5, GPT-4 et PaLM. Le domaine du TALN médical ne fait pas exception, ces modèles ayant démontré des capacités remarquables dans la compréhension de l'état des patients, l'extraction de connaissances médicales et la prise de décision à un niveau professionnel Hadi *et al.* (2023). Toutefois, ces avancées s'accompagnent d'un coût computationnel considérable.

Malgré leurs résultats impressionnants, des études récentes ont mis en évidence certaines limites des LLMs. Par exemple, bien que l'augmentation de la taille des modèles ait permis des améliorations de performance, certaines recherches indiquent que ces modèles peinent encore à saisir des concepts sémantiques complexes Cheng *et al.* (2024). Par ailleurs, Shah (2024) a rapporté que GPT-3.5 était efficace pour extraire des informations essentielles à partir de notes médicales avec une grande rapidité. Toutefois, cette amélioration s'accompagnait de compromis en termes de précision, le modèle restant sujet à des erreurs liées aux abréviations et aux mauvaises interprétations. De plus, Ullah, Parwani, Baig & Singh (2024) a identifié des défis liés à la compréhension du contexte, à l'interprétabilité et aux biais présents dans les données d'entraînement en évaluant GPT-4 pour des applications médicales. Leurs résultats suggèrent que ces limites découlent d'un manque de véritable compréhension des concepts médicaux et d'une représentation insuffisante des dossiers cliniques réels dans les ensembles de formation, ce qui peut entraîner des erreurs et des disparités dans les diagnostics médicaux. En outre, l'utilisation de LLMs préentraînés à grande échelle, tels que GPT-4, soulève d'importantes préoccupations en matière de confidentialité et de sécurité des données des patients, ces modèles retenant intrinsèquement une quantité substantielle d'informations issues de leurs données d'entraînement.

Face à ces défis et à nos ressources computationnelles limitées, nous choisissons de ne pas nous appuyer sur des modèles très volumineux. Nous privilégions plutôt l'amélioration des

performances de modèles de langage de plus petite échelle, tels que BERT, qui offre une approche plus flexible pour les tâches de TALN médical.

## 1.3 Traitement des nombres en TALN

Considérant la capacité de la plupart des modèles d'embeddings de mots à capturer la numératie, comme noté par Wallace, Wang, Li, Singh & Gardner (2019), nous encadrons notre tâche—classifier les valeurs numériques des notes médicales—comme un problème de reconnaissance d'entités nommées (NER). Nous traiterons les valeurs numériques comme des mots qui doivent être catégorisés en une des sept entités ou paramètres physiologiques. Cette recherche confronte deux défis principaux : la nature unique du traitement des données numériques et la taille limitée de notre ensemble de données médicales. Des études antérieures Chen, Takamura, Kobayashi & Miyao (2023); Chen, Huang & Chen (2021); Zhang, Ramachandran, Tenney, Elazar & Roth (2020); Charton (2021) ont exploré la numératie en se concentrant sur la manière dont les nombres sont représentés textuellement, tels que par des chiffres ou des notations scientifiques. Ces études ont également introduit des tâches de pré-entraînement visant à améliorer la compréhension numérique des modèles Thawani, Pujara, Ilievski & Szekely (2021). Bien que ces méthodes aient réussi à impartir une connaissance générale de la numératie, elles restent insuffisantes dans le domaine médical en raison des règles complexes et hautement contextuelles qui régissent la sémantique des données médicales Sutton *et al.* (2020). Certaines études précédentes ont relevé le défi avec une approche agnostique aux nombres, remplaçant les nombres par des mots-clés de remplacement et en échelonnant leurs embeddings pour renforcer l'encodage centré sur le contexte (Golkar *et al.*, 2023; Loukas *et al.*, 2022).

À notre connaissance, il n'existe pas de recherches antérieures ciblant spécifiquement la classification des valeurs numériques dans le domaine médical. Bien qu'une étude Touchent *et al.* (2023) sur CamemBERT-bio ait abordé la reconnaissance des données numériques de manière générale, elle n'a pas approfondi les attributs spécifiques de ces valeurs numériques. En conséquence, nous anticipons que notre recherche apportera de nouvelles perspectives sur l'interprétation des valeurs numériques dans la documentation médicale.

## 1.4 Gérer les Données Limitées

Un autre défi majeur de nombreux projets dans le domaine médical est la contrainte de travailler avec de petits ensembles de données. Bien que les architectures basées sur les transformers affichent généralement de bonnes performances avec de grands ensembles de données, des recherches menées par Li *et al.* (2018); Ezen-Can (2020); Mascio *et al.* (2020) suggèrent que ces modèles peuvent ne pas toujours surpasser les approches traditionnelles (telles que GloVe et ELMo) lorsqu'ils sont appliqués à de petits corpus de notes cliniques. Compte tenu de cette limitation, plusieurs approches telles que la distillation des connaissances (KD) (Bucila, Caruana & Niculescu-Mizil, 2006; Hinton, Vinyals & Dean, 2015) et l'Embedding d'Étiquette pour l'Auto-Attention (LESA) (Si *et al.*, 2020), ont été développées pour améliorer les performances des LLM sur de petits ensembles de données. KD repose sur l'hypothèse qu'un plus petit nombre de paramètres diminue les risques de surapprentissage. Ainsi, cette approche consiste à entraîner un réseau neuronal plus petit à émuler les performances d'un plus grand, suivant un paradigme enseignant-étudiant. Ce concept est à la base du développement de DistilBERT (Sanh, Debut, Chaumond & Wolf, 2019) et de sa version française DistillCamemBERT (Delestre & Amar, 2022). LESA consiste à incorporer une description des différentes catégories de la tâche de classification directement dans la phase initiale d'extraction de caractéristiques. L'intention est d'aider l'extracteur de caractéristiques à se concentrer sur les caractéristiques pertinentes cruciales pour la tâche de classification dans la phase suivante. Cependant, il est important de noter que cette technique a été initialement conçue pour la classification de phrases. Notre travail propose une version étendue de cette approche adaptée à la classification de tokens. LESA est similaire à EmbedNum Nguyen, Nguyen, Ichise & Takeda (2018), qui classe les nombres en comparant leurs embeddings à ceux des catégories potentielles. Cependant, LESA diffère en utilisant la similarité exclusivement pour construire l'embedding du nombre plutôt que pour le processus de classification lui-même.

Une autre approche pour gérer les données limitées consiste à créer des représentations multiples pour les mots, en exploitant diverses méthodes pour obtenir une modélisation plus robuste et généralisable. Typiquement, cela implique de construire une fonction de perte objective comme

la somme de diverses pertes, chacune spécialisée dans différents aspects de la modalité étudiée. Cependant, cette approche soulève la question cruciale de pondérer correctement chaque perte pour éviter les biais. Dans Kendall, Gal & Cipolla (2018), les auteurs proposent une méthode automatique pour pondérer plusieurs pertes en optimisant la log-vraisemblance des prédictions du modèle. Bien que leur approche ait été appliquée aux images, nous avons trouvé leurs résultats pertinents et potentiellement bénéfiques pour notre cas.

## 1.5 Lien vers les articles

Dans les deux chapitres suivants, nous présenterons deux articles détaillant les résultats de notre étude sur la représentation et la classification des nombres dans les huit catégories physiologiques mentionnées précédemment. Cette approche innovante fournit aux professionnels de la santé des outils améliorés pour les soins aux patients.

Le premier article propose une méthodologie d'entraînement pour les LLM de taille moyenne tels que **CamemBERT-bio** pour la compréhension des nombres. Ce travail comble une lacune dans la littérature, car aucune étude antérieure n'avait mis l'accent sur l'entrainement de modèles du language de taille moyenne sur les nombres médicaux. Nous présentons les contributions suivantes :

- Nous introduisons une nouvelle méthodologie d'entraînement qui améliore considérablement les performances de CamemBERT-bio dans la classification des valeurs numériques à travers sept paramètres physiologiques prédéfinis. Notre approche intègre une version généralisée de la technique d'Embedding d'Étiquettes pour l'Auto-Attention (LESA) (Si *et al.*, 2020), adaptée à la classification de tokens.
- Nous développons un algorithme pour catégoriser les valeurs numériques comme critiques ou non critiques, en utilisant des plages standard établies pour la classification.
- Nous évaluons et comparons notre modèle personnalisé avec d'autres modèles traditionnels et avec GPT-4. Les résultats suggèrent que notre modèle surpasse les approches traditionnelles, en particulier les modèles basés sur LSTM, même lorsqu'il est appliqué à des ensembles de

données petits et déséquilibrés. De plus, notre modèle obtient des performances comparables à celles de GPT-4 malgré la différence substantielle de taille des paramètres.

Le deuxième article explore les limitations des modèles de langage de taille moyenne dans la compréhension des nombres. Il examine ensuite les solutions potentielles, décrites comme suit :

- Nous démontrons que **CamemBERT-bio + LESA**, tel qu'introduit dans Lompo, Le, Jouvet & Noumeir (2025), peut être encore amélioré par le préaffinage de la modélisation du langage. Ce préaffinage n'améliore pas CamemBERT-bio quand il est pris seul, ce qui indique que l'intégration de la technique d'Embedding d'Étiquettes pour l'Auto-Attention (LESA) nécessite un préentraînement pour ajuster les paramètres du modèle. Le préaffinage sur un petit ensemble de données médicales améliore significativement le score F1 pour **CamemBERT-bio + LESA**.
- Nous évaluons et comparons notre modèle sur mesure avec des modèles conventionnels y compris des LLMs tels que GPT-4. Les résultats montrent que notre modèle surpasse les approches standard et gère habilement des ensembles de données caractérisés par une taille limitée et un déséquilibre. De plus, notre modèle offre des performances comparables à celles de GPT-4 malgré la disparité significative de la taille des paramètres.

## 1.6 Conclusion

En conclusion, les travaux associés démontrent les progrès significatifs réalisés dans la classification de textes en santé, en particulier avec les modèles basés sur les transformeurs ajustés pour des tâches médicales. Bien que les méthodes traditionnelles d'extraction de features aient été efficaces, des approches récentes telles que BioBERT et CamemBERT-bio montrent des améliorations substantielles dans la capture des connaissances spécifiques au domaine. De plus, les grands modèles de langage (LLMs) tels que GPT-4 et PaLM ont réalisé des avancées significatives en NLP, y compris dans les applications médicales, en montrant de hautes performances dans des tâches comme la compréhension de l'état des patients et la récupération de connaissances. Cependant, les LLMs présentent des exigences computationnelles considérables

et ne sont pas sans leurs limitations, notamment dans la compréhension sémantique complexe et les spécificités du contexte médical. Des techniques telles que la distillation de connaissances et l'intégration des étiquettes pour l'auto-attention offrent des solutions prometteuses pour relever des défis tels que les tailles de jeux de données réduits, offrant ainsi une opportunité pour des modèles de taille moyenne de rivaliser avec leurs homologues plus grands, les LLMs.

# CHAPITRE 2

# ARE MEDIUM-SIZED TRANSFORMERS MODELS STILL RELEVANT FOR MEDICAL RECORDS PROCESSING ?

Boammani Aser Lompo[1] , Thanh-Dung Le[1] , Philippe Jouvet[2] , Rita Noumeir[1]

[1] Electrical Engineering Department, École de Technologie Supérieure,
1100 Notre-Dame Street West, Montreal, Quebec, Canada H3C 1K3
[2] CHU Sainte-Justine Research Center, CHU Sainte-Justine,
3175 Chemin de la Côte-Sainte-Catherine, Montreal, Quebec, Canada H3T 1C5

Article Submitted to the «IEEE Transactions on Artificial Intelligence» Journal
on February 27th 2025.

## 2.1 Abstract

As large language models (LLMs) become the standard in many NLP applications, we explore the potential of medium-sized pretrained transformer models as a viable alternative for medical record processing. Medical records generated by healthcare professionals during patient admissions remain underutilized due to challenges such as complex medical terminology, the limited ability of pretrained models to interpret numerical data, and the scarcity of annotated training datasets.

Objective : This study aims to classify numerical values extracted from medical records into seven distinct physiological categories using CamemBERT-bio. Previous research has suggested that transformer-based models may underperform compared to traditional NLP approaches in this context.

Methods : To enhance the performance of CamemBERT-bio, we propose two key innovations : (1) incorporating keyword embeddings to refine the model's attention mechanisms and (2) adopting a number-agnostic strategy by removing numerical values from the text to encourage context-driven learning. Additionally, we assess the criticality of extracted numerical data by verifying whether values fall within established standard ranges.

Results : Our findings demonstrate significant performance improvements, with CamemBERT-bio achieving an F1 score of 0.89—an increase of over 20% compared to the 0.73 F1 score of

traditional methods and only 0.06 units lower than GPT-4. These results were obtained despite the use of small and imbalanced training datasets.

Conclusions and Novelty : This study introduces a novel approach that combines keyword embeddings with a numerical-blinding technique, demonstrating that medium-sized transformer models can achieve performance levels comparable to LLMs while offering a more cost-effective solution. Our results highlight how these techniques enhance transformer-based models, enabling robust classification even in low-resource settings.

***Clinical and Translational Impact Statement*— The application of CamemBERT-bio to numerical values classification represents a promising avenue, especially regarding limited clinical data availability. By integrating keyword embeddings into the model and adopting a number-agnostic strategy by excluding all numerical data from the clinical text to improve CamemBERT-bio's performance, the framework has the potential to enhance accuracy in classifying numerical values from a small French clinical dataset. Our model is naturally extended to evaluate the criticality of these numerical values based on public medical benchmarks and is subsequently integrated into the hospital's Clinical Decision Support System (CDSS). Performance comparisons with state-of-the-art LLMs demonstrate that our model offers a viable, cost-effective alternative to large-scale LLMs while also enhancing data privacy.**

## 2.2 Introduction

Machine learning applications within clinical settings represent an evolving and dynamic field of research. This domain promises to equip healthcare professionals with advanced technologies that efficiently utilize resources. Such advancements aim to democratize access to high-quality healthcare irrespective of temporal and spatial constraints Sutton *et al.* (2020). Recent strides in Natural Language Processing (NLP) have enabled neural networks to analyze vast corpuses of text, facilitating the extraction of pertinent and beneficial information. These novel technological

tools have the potential to significantly contribute to the daily practices of healthcare professionals, aiding in the exploration and analysis of patient medical records.

In the majority of intensive care units (ICU), substantial medical information is documented daily, either in the form of textual notes or as numerical data generated by machines. The machine-generated data, structured in tabular format, can be readily incorporated into various clinical decision support algorithms. However, as pointed out by Sutton *et al.* (2020), the utilization of textual notes is often less efficient, primarily due to their unstructured format and the frequent use of medical jargon, which tends to be dense with information and sometimes includes incomplete sentences. Indeed, the presence of abbreviations, misspellings, and other types of input errors prevalent in the medical domain introduce considerable uncertainty, rendering systematic approaches like case-by-case algorithms ineffective.

Traditional machine learning methodologies in this context have primarily involved non-contextual word embedding models, complemented by task-specific layers Le *et al.* (2022a). Subsequent advancements introduced Recurrent Neural Networks (RNN) capable of contextual text representations. While these methods have achieved notable success in several text classification tasks Mascio *et al.* (2020), their effectiveness is largely attributed to the fact that the evaluation dataset is specifically customized to suit their capabilities. Current advancements in NLP predominantly involve Transformer-based models Vaswani *et al.* (2017) such as BERT and GPT Devlin *et al.* (2018); Brown *et al.* (2020). These Pretrained Large Language Models (PLMs) largely overcome data tailoring challenges as they necessitate minimal preprocessing. However, in specific scenarios like medical text classification with small and imbalanced training datasets, they may not outperform RNNs Mascio *et al.* (2020); Chen *et al.* (2019).

This paper contributes to a project that aims to utilize Pretrained Large Language Models (PLMs) for the diagnosis of cardiac failure. The National Heart, Lung, and Blood Institute (2023) defines cardiac failure as a condition where the heart fails to pump sufficient blood to meet the body's needs. This insufficiency can arise from the heart's inability to adequately fill with blood or weakened capacity to pump effectively. The manifestation of cardiac failure can vary, affecting

physiological parameters such as blood pressure, ventricular gradients, and ejection fractions. Additionally, genetic factors and the patient's prior health conditions influence the severity of cardiac failure, necessitating a tailored approach for each case. Simply put, it is essential to interpret the patient's medical information within the framework of their medical history. In our research, we tackle categorizing numerical values obtained from medical notes into one of seven predetermined physiological parameters using CamemBERT-bio Touchent *et al.* (2023). These parameters include diverse heart failure indicators such as ejection and shortening fraction, saturation in oxygen, heart rate, pulmonary artery diameter, ventricular gradient, and the size of the atrial/ventricular septal defect. Given that a significant amount of our patients are newborn children, we also include the APGAR score as key indicator or vital distress. Subsequently, we assess whether these numerical values indicate a critical medical condition.

### 2.2.1 Goal statement

By training CamemBERT-bio on a small and imbalanced dataset, we aim to achieve the following result :

**Input medical note** :

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

**Expected output** :

Tableau 2.1 Expected output from the input medical note

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

**Contribution**

In this study, we present the following contributions :

- We introduce a novel training methodology that substantially enhances the performance of CamemBERT-bio in the classification of numerical values across seven predefined physiological parameters. Our approach incorporates a generalized version of the Label-Embedding for Self-Attention (LESA) technique Si *et al.* (2020), tailored for token classification.
- We develop an algorithm to categorize numerical values as critical or non-critical, using established standard ranges for classification.
- We evaluate and compare our customized model with other traditional models and with GPT-4. The findings suggest that our model performs better than traditional approaches, particularly LSTM-based models, even when applied to small and imbalanced datasets. Additionally, our model achieves performance comparable to GPT-4 despite the substantial difference in parameter size.

This research aims to set a foundation for future methodologies in training BERT models effectively for similar tasks involving numerical values understanding. We anticipate that the findings from this initial study will facilitate further investigations into medical reasoning derived from medical notes.

### 2.3 The task and the dataset

The dataset used in this project was provided by the Pediatric Intensive Care Unit at CHU Sainte-Justine (CHUSJ) and consists of medical notes from patients under the age of 18. It contains 1,072 annotated samples, totaling approximately 30,000 tokens.

Following approval from the CHUSJ Research Ethics Board (protocol number 2020-2253), we focused on two types of medical notes : **admission notes** and **evaluation notes**. These note types were selected because they contain the physician's rationale for the patient's hospitalization, along with initial care instructions based on the patient's condition at admission.

From the full dataset, we selected 100 representative notes for manual annotation. Two independent physicians from CHUSJ, who were not the original authors, reviewed and annotated these notes by identifying pertinent numerical values. These values were then categorized into eight predefined classes.

To prepare the data for model training, we segmented each annotated note into shorter text units, each containing at least 12 tokens. Care was taken to ensure that each segment retained sufficient context for interpreting the numerical values.

To ensure that the segmentation provided enough context and preserved the clinical meaning, we followed a two-step approach :

- First, we retained the preexisting sections in each medical note. Most notes were already divided into sections such as *Historique du patient*, *Médication*, *Examens laboratoires*, and *Note à l'admission*. Since these sections were typically semantically self-contained, they served as our initial segmentation units.
- Then, for any of these sections exceeding 40 tokens in length, we applied an automatic splitting process. If a segment could be divided into two or three complete sentences of at least 12 tokens each, we performed the split. This approach leveraged the fact that most clinical sentences are contextually rich and understandable on their own. Manual verification was carried out for most of the resulting segments to ensure that their interpretation did not depend on other segments.

**Examples of segmented text :**

- *"14/08 : Bonne contractilité ventriculaire gauche qualitative. Simpson de 65%."*
- *"Brady ad 32 au Holter. Écho cœur N s/p 1 épisode de quasi-noyade 07/2015."*

Our downstream task is to systematically categorize tokens into eight distinct classes, each serving as a representative entity within the medical domain. These classes encompass :

- **Contractibilité** (contractibility). It regroups the ejection fraction and the shortening fractions, which are indicators of the contraction ability of the heart. It will be denoted **Cp**.

- **Fréquence cardiaque** (heart rate). It will be denoted **FC**.
- **Diamètre artère pulmonaire** (Pulmonary artery diameter). It will be denoted **D**.
- **Saturation en oxygène** (Saturation in oxygen). It will be denoted **SO2**.
- **APGAR** (APGAR score). It will be denoted **APGAR**.
- **Gradient** (gradient). The measurement of the gradient of blood pressure between the heart ventricles. We don't consider orientation here. It will be noted **G**.
- **CIA-CIV** : Taille de la Communication Inter Ventriculaire/Auriculaire (size of the atrial/ventricular septal defect). These parameters are utilized to assess the severity of the cardiac genetic malformation. This class will be denoted **CIA-CIV**.
- **Out of class** (Items not belonging to the aforementioned categories). It will be denoted **O**.

This selection of classes was carefully tailored through a collaborative effort with healthcare providers from CHUSJ. The intent behind this classification schema is to facilitate the identification of risks associated with heart failure based on the patient's medical history. Table 2.2 displays the distribution of the eight classes in the annotated dataset, indicating a notably small size and significant imbalance."

Tableau 2.2 Distribution of the eight classes of the annotated dataset

| **Class** | **Cp** | **FC** | **D** | **SO2** | **APGAR** | **G** | **CIA-CIV** | **O** |
|---|---|---|---|---|---|---|---|---|
| **Size** | 21 | 80 | 57 | 143 | 130 | 41 | 58 | 27387 |

The second component of our project involves determining the criticality of the numerical values. For this aspect, a physiotherapist from CHUSJ supplied us with the standard ranges applicable to most of the physiological parameters examined in our study.

### 2.3.1 Contractibility (Contractibilite)

- Ejection fraction (fraction d'éjection) : 50 – 70%
- Shortening fraction (fraction de raccourcissement) : 20 – 40%

First, contractility is one of the key indicators for diagnosing heart failure. Specifically, the ejection fraction and shortening fraction are two critical measures for assessing contractility Association (2023).

### 2.3.2 Heart rate

The second indicator is heart rate, with Table 2.3 showcasing the normal heart rate (fréquence cardiaque) ranges according to the patient's age. The values were extracted from Pulse (2019). Using these ranges, we can determine whether to classify a patient as having a normal heart rate.

Tableau 2.3 Heart rate Table

| **Age (years)** | **Heart rate (bpm)** |
|---|---|
| < **1 month** | 70 - 190 |
| **1-11 months** | 80 - 160 |
| **1-2 years** | 80 - 130 |
| **3-4 years** | 80 - 120 |
| **5-6 years** | 75 - 115 |
| **7-9 years** | 70 - 110 |
| > **10 years** | 60 - 100 |

### 2.3.3 Pulmonary artery diameter

The third indicator involves measuring the diameter of the pulmonary artery (diamètre pulmonaire). Table 2.4 displays the normal pulmonary artery diameter ranges based on the patient's weight. The values were extracted from Acar (2008)

Tableau 2.4 Pulmonary artery diameter

| Weight (kg) | Diameter (mm) | Weight (kg) | Diameter (mm) |
|---|---|---|---|
| **3** | 4.2 | **12** | 9.2 |
| **4** | 5.3 | **14** | 9.5 |
| **5** | 6 | **16** | 10.2 |
| **6** | 6.7 | **18** | 10.6 |
| **7** | 7 | **20** | 11 |
| **8** | 7.8 | **25** | 11.7 |
| **9** | 8.2 | **30** | 12.4 |
| **10** | 8.5 | **35** | 12.8 |

### 2.3.4 Oxygen saturation

Oxygen saturation also serves as an essential indicator. According to Aubertin *et al.* (2013), the recommended level of oxygen saturation for children should exceed 96% for children.

Although the APGAR score, ventricular gradient (**G**), and the size of the atrial-ventricular septal (**CIA-CIV**) defect often provide indications of potential cardiac failure or vital distress, our healthcare collaborators advise against assigning standard values to these parameters. This caution is due to the complexity of factors involved, emphasizing that these parameters should be evaluated by experts on an individual basis.

In summary, the task comprises two steps : the initial step involves identifying the physiological parameter to which a specific numerical value pertains, followed by the second step, which involves evaluating the criticality of these numerical values. These combined steps enable clinicians to promptly identify key indicators suggesting a cardiac failure.

## 2.4 Methodology

The methodology adopted in this study is illustrated in Figure 2.1. It consists of two key phases : (i) Phase 1 involves detecting numerical values and classifying them based on their attributes

using eight predefined classes, and (ii) Phase 2 focuses on comparing each detected class against a corresponding range to determine if it falls within the normal range or outside of it. Technically, to achieve the first phase, we initially replace every numerical value within the text with a specific placeholder word, prompting the model to derive as much information as possible from the text's contextual clues. After this substitution, the modified text and certain class-related keywords are embedded and introduced into the Label Embedding for Self-Attention (LESA) layer. The aim here is to produce word representations that are more distinct and informative. These refined representations are then processed by a classifier to determine the category of each word, marking the completion of phase 1. Finally, during the second phase, each classified numerical value is evaluated as critical or not, according to the benchmark outlined in section 2.3. By successfully completing these two phases, we can implement an end-to-end automatic algorithm that detects and classifies the attributes of numerical values and verifies whether these values fall inside or outside their corresponding ranges. This capability significantly enhances the efficiency of healthcare professionals by boosting their review of key indicators, thereby supporting their clinical decision-making process.

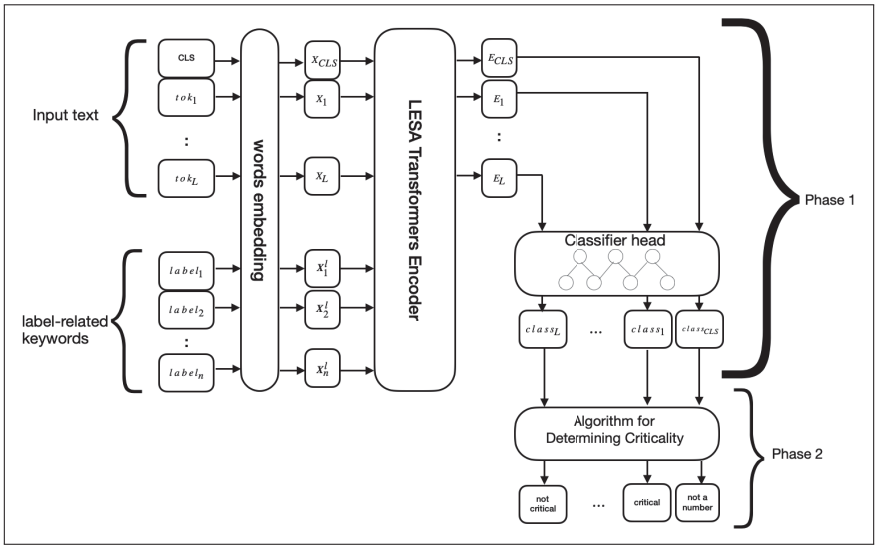


Figure 2.1 An overview of the proposed methodology to classify numerical values from medical notes

### 2.4.1 LESA-BERT for token classification

The main innovation of BERT is the integration of transformer encoder layers. Each of the 12 transformer layers contains a multi-head self-attention sub-layer and a feed-forward sub-layer. In plain terms, let's consider a text $\boldsymbol{t} = [token_1, token_2, \cdots, token_L]$ of $L$ tokens, with initial embeddings $[\boldsymbol{x}_{token_1}, \cdots, \boldsymbol{x}_{token_L}]$. Those initial embeddings prepended with the special token $[CLS]$ are inputted to the transformers layers in this form :

$$\boldsymbol{X} = [\boldsymbol{x}_{CLS}, \boldsymbol{x}_{token_1}, \cdots, \boldsymbol{x}_{token_L}] \in \mathbb{R}^{(L+1)\times D} \tag{2.1}$$

where $D$ is the dimension of the embedding space. The $[CLS]$ embedding plays the role of a sentence aggregator which will not be useful in our work given that we are focusing on the tokens themselves (and not on the global sentence). Drawing inspiration from the methodology outlined in LESA-BERT (Si *et al.*, 2020), we leverage the expertise of healthcare professionals to gather a collection of class-related keywords chosen to be representative of the various classes under consideration. To this end, we have compiled a list of the final 8 classes, presented in Table 2.5 for reference.

Tableau 2.5 Representative keywords per labels

| **Label** | **Key Words** |
|---|---|
| Out of class | mot, patient, historique |
| Contractilité | fraction, ejection, raccourcissement |
| Fréquence cardiaque | cardiaque, coeur, frequence |
| Diamètre pulmonaire | diamètre, pulmonaire, artère |
| Saturation en oxygène | oxygène, O2, sat |
| APGAR | apgar, minute, nombre |
| Gradients | gradient, pulmonaire, ventricule |
| CIA-CIV | cia, civ, inter |

For every individual class, we calculate the mean value of the keyword initial embeddings. This computation yields an embedding matrix denoted as $\boldsymbol{X}^l \in \mathbb{R}^{n \times D}$ ($n$ is the number of classes), encompassing distinct label embeddings for each class.

Subsequently, the input sequence and the label embeddings are then mapped to the key, query, and value triplets, denoted as matrices $(\boldsymbol{K}, \boldsymbol{Q}, \boldsymbol{V})$ and $(\boldsymbol{K}^l, \boldsymbol{Q}^l, \boldsymbol{V}^l)$ respectively :

$$\boldsymbol{K} = \boldsymbol{X}W_K, \boldsymbol{Q} = \boldsymbol{X}W_Q, \boldsymbol{V} = \boldsymbol{X}W_V \tag{2.2}$$

$$\boldsymbol{K}^l = \boldsymbol{X}^l W_K, \boldsymbol{Q}^l = \boldsymbol{X}^l W_Q, \boldsymbol{V}^l = \boldsymbol{X}^l W_V \tag{2.3}$$

where $\{W_K, W_Q, W_V\} \in \mathbb{R}^{D \times D}$ are learnable parameters. From this point on, these matrices are split into multiple heads

$$\boldsymbol{K}_h^l = \boldsymbol{K}^l[:, hd - d : hd], \boldsymbol{Q}_h^l = \boldsymbol{Q}^l[:, hd - d : hd], \boldsymbol{V}_h^l = \boldsymbol{V}^l[:, hd - d : hd] \tag{2.4}$$

$$\boldsymbol{K}_h = \boldsymbol{K}[:, hd - d : hd], \boldsymbol{Q}_h = \boldsymbol{Q}[:, hd - d : hd], \boldsymbol{V}_h = \boldsymbol{V}[:, hd - d : hd] \tag{2.5}$$

for $h = 1, \cdots, 12$. where $d$ is the size of each head. The operation $[:, a : b]$ consists in extracting the columns from index $a$ to index $b - 1$. Therefore $\{\boldsymbol{K}_h, \boldsymbol{Q}_h, \boldsymbol{V}_h\} \in \mathbb{R}^{(L+1) \times d}$.

The point of subdividing the data into multiple heads is to prevent bias propagation by having multiple independent text representations. Now we can define the self-attention according to each head as :

$$\boldsymbol{A}_h = \frac{\boldsymbol{Q}_h \boldsymbol{K}_h^T}{\sqrt{d}} \in \mathbb{R}^{(L+1) \times (L+1)} \tag{2.6}$$

$$\textbf{Self-attention}_h = \text{Softmax}(A_h) \in \mathbb{R}^{(L+1) \times (L+1)} \tag{2.7}$$

for $h = 1, \cdots, 12$,

and the cross-attention between the label embeddings and the text tokens :

$$\boldsymbol{A}_h^l = \frac{\boldsymbol{Q}_h^l \boldsymbol{K}_h^T}{\sqrt{d}} \in \mathbb{R}^{n \times (L+1)} \tag{2.8}$$

for $h = 1, \cdots, 12$,

where Softmax$(\cdot)$ is softmax function applied row-wise. Now, unlike Si *et al.* (2020), we want to employ the cross-attention to improve the whole self-attention matrix, which was designed only to update the $[CLS]$ embedding from Si *et al.* (2020). Therefore, we introduce an intermediate matrix

$$\textbf{CoSim}_h = \text{norm}(\boldsymbol{A}_h^l)^T \text{norm}(\boldsymbol{A}_h^l) \in \mathbb{R}^{(L+1)\times(L+1)} \tag{2.9}$$

for $h = 1, \cdots, 12$,

where norm$(\cdot)$ is the row-wise normalizing function. $\boldsymbol{A}_h^l$ calculates an affinity score for each pair of (token, label). Then, those affinity scores are used to compute a cosine similarity-like matrix denoted by **CoSim**. **CoSim** measures the similarity of each pair of tokens based on their affinities with the labels. Finally, we can compute the new self-attention matrix :

$$\textbf{New-Self-attention}_h = \textbf{Self-attention}_h + \textbf{CoSim}_h \tag{2.10}$$

for $h = 1, \cdots, 12$,

and then the output features as the weighted average :

$$\boldsymbol{O}_h = \textbf{New-Self-attention}_h V_h \in \mathbb{R}^{(L+1)\times d} \tag{2.11}$$

$$O = [O_1, O_2, \cdots, O_{12}] \in \mathbb{R}^{(L+1)\times D} \tag{2.12}$$

for $h = 1, \cdots, 12$.

The whole process is summarized in Figure 2.2

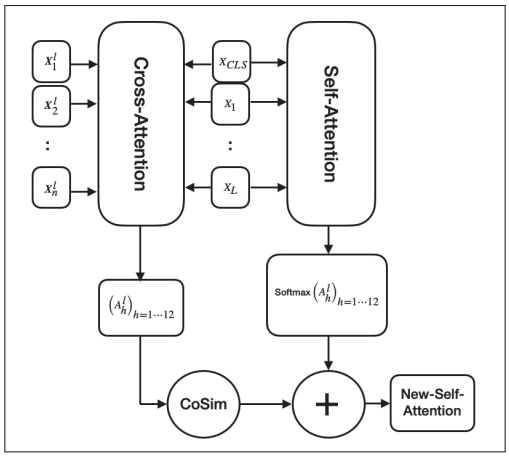


Figure 2.2 An illustration of the Label Embedding for Self-Attention (LESA). The input of this layer are the tokens emebeddings $[X_{CLS}, X_1, \cdots X_L]$ and the keywords embeddings $[X_1^l, X_2^l, \cdots, X_n^l]$. This layer outputs the enhanced self-attention

### 2.4.2 Blind dataset

Handling numerical values presents an additional challenge, as the numbers themselves offer no direct indication for their categorization, unlike regular words. For example, in the sentences "*The heart rate was 180*" and "*The white cells count was 180*", the token *180* represents different entities, heart rate, and white blood cell count. This reliance on context necessitates approaches that are agnostic to the numerical value. Consequently, we implemented a technique where all numerical values were uniformly replaced with the keyword **"nombre"**. This is based on the expectation that the model would develop its embeddings primarily from the surrounding context rather than depending heavily on its pre-existing vocabulary. This method aligns well with the concept of label embeddings, as it encourages the model to identify relationships between the context of the substituted term and the keywords related to the labels fed into the model. Through this method, we aspire to augment the model's aptitude for contextual comprehension. This will result in a refined capacity to extract meaningful insights from the given data. A similar idea has been employed in prior research. For example, in Griffith & Kalita (2019), the authors substituted all numerical values within a mathematical problem with symbols before proceeding to solve it.

In order to implement this method, we distinguish two types of numerical values :

- Quantitative numbers, denoting measurements, time, dates, etc., which are the primary focus of our classification task and undergo the blinding process.
- Code numbers, representing specific medical terms (e.g., "B1B2", "G1P3", "22q11", etc.) and units (e.g., "mm2", "cm3", etc.), which remain unaltered during the blinding process.

This approach improved performance regarding both accurate responses and the comprehensibility of the answer.

## 2.5 Experiments

This section presents our models, baselines, experimental setup, and the results we obtain.

### 2.5.1 Our proposed models

We propose 2 models incorporating the different solutions introduced in this work : label embeddings and blind datasets. A visual description is presented in Fig. 2.3.

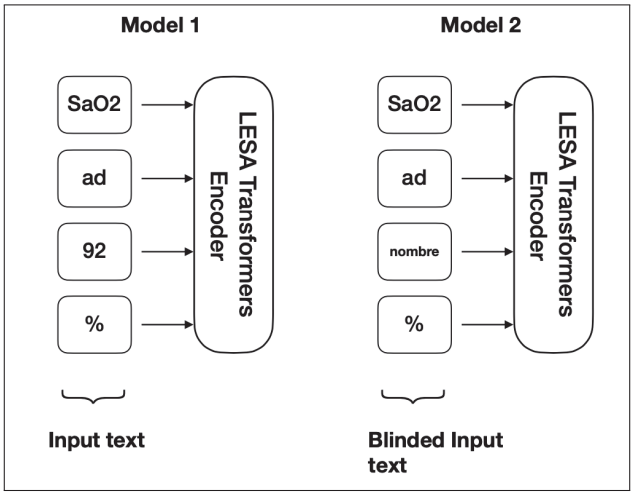


Figure 2.3 An overview of the two proposed models

**Model 1** : We augmented **CamemBERT-bio** with LESA, initializing it with pre-trained parameters from the original **CamemBERT-bio** model. Subsequently, we fine-tuned the model

on the token classification task using our small dataset. Additionally, we incorporated LayerNorm in the token classification head of **CamemBERT-bio** to improve its generalization capabilities across diverse inputs, stabilizing the distributions of hidden layers (Xu, Sun, Zhang, Zhao & Lin, 2019).

**Model 2** : Similar to **Model 1**, this model utilizes the Blind Dataset during the fine-tuning phase. The Blind Dataset comprises our small dataset with numerical values replaced by the keyword "nombre".

Tableau 2.6 A short recap of all the models we proposed, along with the baseline models. This includes the Tokenizer used for preprocessing, the Embedding layer, the Datasets used for training, and number of parameters

| Model | Tokenizer | Embedding | Training Dataset | Model parameters |
|---|---|---|---|---|
| GPT-4 | Tiktoken | Transformer | Unknown | 1.8 trillions |
| bi-LSTM (v1) | WordPiece | Linear | Standard Dataset | 230k |
| bi-LSTM (v2) | RegEx | GloVe | Standard Dataset | 230k |
| DistilCamemBERT | WordPiece | RoBERTa | Standard Dataset | 55 M |
| Camembert-bio | WordPiece | RoBERTa | Standard Dataset | 110 M |
| Camembert-bio (ComNum) | WordPiece | RoBERTa | ComNum Dataset for prefinetuning + Standard Dataset for finetuning | 110 M |
| Model 1 | WordPiece | RoBERTa + LESA ( 2.4.1) | Standard Dataset | 110 M |
| GloVe + SVM | RegEx | GloVe | Standard Dataset | 1 M |
| bi-LSTM (v1)(b) | WordPiece | Linear | Blind Dataset | 230k |
| bi-LSTM (v2)(b) | RegEx | GloVe | Blind Dataset | 230k |
| Camembert-bio(b) | WordPiece | RoBERTa | Blind Dataset | 110 M |
| Model 2 | WordPiece | RoBERTa + LESA ( 2.4.1) | Blind Dataset | 110 M |

### 2.5.2 Baselines

In this study, we aim to benchmark our models against a variety of baseline models sourced from previously published works. These baselines encompass a spectrum from conventional NLP models to cutting-edge deep learning-based models.

GPT-4 : For these experiments, we selected a subset of 200 entries from our dataset and manually input them into ChatGPT using a few-shot learning approach. The model's responses were generated as a sequence of numerical values, each associated with a corresponding category. The aggregated outputs were then utilized to compute the evaluation metrics presented below.

**GloVe + SVM :** Global Vector (GloVe)(Pennington *et al.*, 2014) proposes a way of extracting features from the global words co-occurrence counts instead of local context windows. Their idea is to cast the word-embedding problem as a hand-crafted weighted least squares regression. On top of this embedding, a Support Vector Machine (SVM) is applied for the classification task. In our implementation, we used a linear kernel and the Hinge loss formula adapted to multiclass classification as introduced in Crammer & Singer (2002). This combination of GloVe for feature extraction and SVM for classification presents a compelling baseline that underscores the viability of statistical approaches in comparison to transformer-based models.

**DistillCamemBERT :** DistilCamemBERT (Delestre & Amar, 2022), uses Knowledge Distillation (Bucila *et al.*, 2006), (Hinton *et al.*, 2015) in order to reduce the parameters of CamemBERT. Basically, Knowledge Distillation is a training technique involving an expert model (teacher) and a smaller model (student). The goal is to make the student reproduce the behavior of the teacher. DistillCamemBERT reached very similar performances to CamemBERT on many NLP tasks. This baseline shows how our solution compares to the reduction of parameters. For this model, the public version available on Huggingface was used to conduct the experiment (Wolf *et al.*, 2019).

**LSTM :** Long Short-Term Memory (LSTM) is a particular type of recurrent neural network (RNN) that enables the model to invoke information from nearby or distant elements within a given input sequence. We have replicated the most effective architecture as in Ezen-Can (2020) :

- Bi-LSTM (v1) : Linear embedding + 1 bidirectional + 1 unidirectional (50 neurons per layer)
- Bi-LSTM (v2) : GloVe embedding + 1 bidirectional + 1 unidirectional (50 neurons per layer)

The Linear embedding is the trainable class nn.Embedding from PyTorch. We also added a LayerNorm, since it enhances generalization (Xu *et al.*, 2019). The GloVe model for the French language is imported from Kocaman & Talby (2021). We conducted empirical experiments to determine that increasing the model's size did not yield performance improvements. Consequently, these LSTM-based architectures give the optimal outcomes.

We trained both versions of Bi-LSTM, denoted as Bi-LSTM (v1) and Bi-LSTM (v2), using both our dataset and the Blinded dataset. The results obtained after training with the Blinded dataset are designated as Bi-LSTM (v1)(b) and Bi-LSTM (v2)(b).

**CamemBERT-bio :** This is a french version of BioBERT (Lee *et al.*, 2019) introduced by Touchent *et al.* (2023). We trained Camembert-bio using both our dataset and the Blinded dataset. The results obtained after training with the Blinded dataset are labeled as CamemBERT-bio(b).

**CamemBERT-bio + ComNum :** This model adheres to the training methodology outlined by Chen *et al.* (2023), which involves representing numbers in scientific notation and pre-finetuning the model on the Comparing Number Dataset. The Comparing Number Dataset facilitates a binary classification task based on assertions that compare two numbers. We adjusted the value range to suit our specific needs.

All the models we proposed, along with the baseline models, are summarized in Table 2.6.

### 2.5.3 Setup and Evaluation

For each of the models, we ran the training 10 times with different random seeds. To prevent overfitting, we used early-stopping with 4 epochs of patience for the transformers-based models and 10 epochs for the other models with a maximum of 100 epochs. The early stopping is monitored by $F_1$ score, a class-wise metric well-suited for imbalanced datasets. The $F_1$ implementation is imported from *Scikit learn* (version 1.2.2).

We used a learning rate of $3 \times 10^{-5}$ with the AdamW optimizer Loshchilov & Hutter (2017), following the setup in Benjamin Muller,Nathan Godey (2022). All hyperparameters—including early stopping criteria, learning rate, weight decay, and learning rate scheduler parameters—were selected based on extensive experimentation. For each hyperparameter, we sampled values within standard ranges and evaluated several combinations to identify the most effective configuration. The final settings are available in the project repository.

The dataset was split into training, validation, and testing sets with a target distribution of 70%, 15%, and 15% across all eight classes. To preserve class balance, samples were sequentially allocated to each subset to ensure that the class distribution remained consistent across all three sets. While we considered implementing $k$-fold cross-validation, maintaining this balanced distribution across all folds proved challenging ; as a result, we opted not to apply cross-validation in this study. The training on our two proposed models lasted approximately 106 hours on an NVIDIA Tesla V100 GPU (32GB). In comparison, training all the baseline models took roughly 50 hours. The complete codebase, excluding the dataset and the weights of the trained models, is accessible on our GitHub repository.

## 2.6 Results and Discussions

As shown in Table 2.8, all the models achieve nearly perfect $F_1$ scores for the Out-of-class category, which is expected since it comprises a large number of examples. However, the GloVe + SVM model consistently performs the poorest across almost all classes. As pointed out by Thawani *et al.* (2021), this can be attributed to its embedding layer (GloVe) which consistently assigns a zero embedding to words not present in its training data. These words include many numerical values and medical terms. In contrast, RNN-based (respectively transformer-based) models inherently possess the capability to capture context through their recurrent architecture (respectively attention mechanisms). Despite the efficiency of the SVM layer, this limitation of the Glove embedding remains insurmountable.

Since all the models exhibit identical performance for the Out-of-class category, and considering that the remaining classes are approximately equally important in terms of size, we calculate the overall $F_1$ score by averaging the $F_1$ scores of each category including Out-of-class, without applying any weighting. This approach is adopted because the introduction of weighting would not yield a more nuanced evaluation. Table 2.7 displays the overall $F_1$ scores of each method considered in this study.

Tableau 2.7 Overall $F_1$ score per model

| **Model** | **Overall $F_1$ score** |
|---|---|
| GPT-4 | 0.95 |
| bi-LSTM (v1) | 0.54 |
| bi-LSTM (v2) | 0.42 |
| DistilCamemBERT | 0.68 |
| Camembert-bio | 0.66 |
| Camembert-bio (ComNum) | 0.82 |
| Model 1 | 0.78 |
| GloVe + SVM | 0.26 |
| bi-LSTM (v1)(b) | 0.77 |
| bi-LSTM (v2)(b) | 0.71 |
| Camembert-bio(b) | 0.73 |
| Model 2 | **0.89** |

Based on this ranking, it is clear that our two approaches (**Model 1** and **Model 2**) consistently surpass all other baseline methods except for CamemBERT-bio + ComNum and GPT-4.

### 2.6.1 Results OF LESA

One of the important features for interpretability is the attention matrix Wiegreffe & Pinter (2019). In simple terms, each row of this matrix represents a scaled, positive quantification of the relationship between the token corresponding to that row and all other tokens. In the camemBERT-bio's architecture multiple attention heads are generated in order to have diverse representations of the tokens. In Clark, Khandelwal, Levy & Manning (2019), the authors show that certain of these attention heads correspond to some linguistic notions of syntax. For example, certain attention heads attend to the direct objects of verbs, determiners of nouns, objects of prepositions, etc. with remarkably high accuracy. In our work, we analyzed how the attention heads from CamemBERT-bio and Model 1 detect and associate numerical values with other tokens in the input text. To achieve this, we analyze the attention patterns in the following input text :

> *"Née à terme, Grossess et accouchement sans complication PN 3.23 Kg, APGAR 8-9-9."*

To ensure clarity, we will dissect the sentence into its initial and final segments.

For the first segment, "Née à terme, Grossess et accouchement sans complication," we visually represented the attention patterns of two attention heads from CamemBERT-bio and Model 1 in Figure 2.4. Each cell in these attention heads denotes the relationship between the tokens indexed by the row and column. A darker cell indicates a weaker relationship, while a lighter cell indicates a stronger relationship. Upon examining Figure 2.4(a), we observe an uniformly low attention in CamemBERT-bio hidden states, suggesting a lack of meaningful word relationships learned from our dataset. This trend persists across other attention heads, implying overfitting in CamemBERT-bio. Conversely, Figure 2.4(b) reveals some activated cells linking specific tokens, such as "accouchement" and "grossesse," here preceded by a space, as typical in WordPiece tokenization. This pattern is consistent across most attention heads in Model 1, indicating that our LESA technique facilitated the establishment of more token relationships.

For the final segment of the sentence, "PN 3.23 Kg, APGAR 8-9-9.", we conducted a similar analysis focusing specifically on the numerical values. In Figure 2.4, we presented the most activated attention heads from both CamemBERT-bio and Model 1. Figures 2.4(c) and (d) illustrate that both models effectively identified the numbers, yet Model 1 demonstrated greater precision. Notably, Figure 2.4(c) displays a broader area (upper left quadrant) denoting the dependencies of the number "3.23 Kg," whereas Figure 2.4(d) concentrates this information into a more compact area, excluding extraneous dependencies. Additionally, in the lower portion of Figure 2.4(c), CamemBERT-bio exhibits a dependency on "-" within "8-9-9" and the rest of the text. This imprecision is rectified in Figure 2.4(d) by Model 1.

Our hypothesis to explain these findings is that the dataset's small size, in comparison to the CamemBERT-bio complexity, leads the model to resort to shortcuts rather than genuinely focus on meaningful features, resulting in non-optimal generalization as Table 2.7 shows. The dataset's scarcity exacerbates this poor generalization. The LESA technique aims to equip CamemBERT-bio with more language cues to reach a semantical word embedding more easily.

Moreover, in Figure 2.5, we depicted the $F_1$ scores along with their standard deviations for all models trained without the Blind Dataset except GPT-4, enabling us to assess its impact independently of the Blind Dataset. The figure illustrates that Model 1 achieved the second-highest performance among the models that did not utilize the Blind Dataset, trailing only behind CamemBERT-bio + ComNum. It is important to note that CamemBERT-bio + ComNum received a pre-finetuning on a specific dataset before being fine-tuned on our target dataset. Despite this discrepancy, the LESA technique narrowed the performance gap between CamemBERT-bio and CamemBERT-bio + ComNum by enhancing CamemBERT-bio's overall $F_1$ scores by 0.12.

### 2.6.2 Results of BLIND DATASET

The concept of the blind dataset, aims at encouraging the model to prioritize context over individual tokens for classification purposes. This approach was implemented on CamemBERT-bio, Model 1, bi-LSTM (v1), and bi-LSTM (v2). These models were chosen for various reasons. Firstly, they produce contextual word embeddings, unlike GloVe, whose embeddings are context-independent. This capability of contextual embeddings is crucial for training with the Blind Dataset, as it ensures that occurrences of "nombre" are classified based on their context. Additionally, the four selected models encompass different types and complexities of models. Therefore, our analysis will assess the impact as the model complexity increases.

Figure 2.6 illustrates the $F_1$ scores for each model trained with both Normal and Blind Datasets. The figure demonstrates a significant improvement in both bi-LSTM (v1) and bi-LSTM (v2) across all classes, characterized by higher $F_1$ scores and reduced standard deviations. This improvement indicates that the benefits of the Blind Dataset technique are consistent and applicable across various training scenarios, not reliant on specific parameter optimization paths during training. A similar positive trend is observed for **Model 2** compared to **Model 1**. However, while **CamemBERT-bio (b)** shows improved $F_1$ scores compared to Camembert-bio, there is also an increase in standard deviations. Nevertheless, in all cases, this strategy results in better overall $F_1$ scores, as shown in Figure 2.6. Based on these observations, we can confidently assert that this approach has been successful.

It is noteworthy to mention that, with the exception of GPT-4, the data in Table 2.7 **Model 2** significantly outperformed all the other approaches with a $F_1$ score margin of 0.07 over the second-best,**CamemBERT-bio + ComNum**. This consistent superiority is observed across nearly all categories, underscoring the effectiveness of our method. Furthermore, despite having several orders of magnitude more parameters, GPT-4 and **Model 2** exhibit comparable performance overall.

### 2.6.3 GPT-4 VS MODEL 2

The category-wise and overall F1 score comparison demonstrates a clear advantage for GPT-4, with an overall F1 score that is 0.06 units higher. However, this performance gain appears marginal when considering the substantial difference in model size—**Model 2** with 110M parameters versus the several trillion parameters of GPT-4.

To further analyze the performances of GPT-4, we conducted a failure case study to assess the recurrence of errors. Acoording to Table 2.8, the model consistently underperformed in the Contractibility (Cp) and Gradient (G) categories. In the first category, errors predominantly occurred when ambiguous abbreviations were encountered. For example, "FR" was misinterpreted as Fréquence respiratoire instead of Fraction de raccourcissement, despite the fact that these two measurements operate on different numerical scales. This suggests potential weaknesses in GPT-4's numerical reasoning capabilities. In the second category, the model frequently misclassified different types of gradients (e.g., aortic, atrial/ventricular) as being related to the interventricular gradient, indicating limitations in its contextual understanding of medical terminology. Table 2.8 indicates that Model 2 exhibits a similar limitation in the Cp category. However, it significantly outperforms GPT-4 in the G category, demonstrating superior contextual understanding of medical notes.

From a practical standpoint, our approach offers significant advantages in terms of system integration and computational efficiency. Unlike large-scale LLMs such as GPT-4 and LLaMA, our model can be seamlessly incorporated into medical decision-making systems while requiring

minimal computational resources and giving comparable results. This not only enhances accessibility but also improves data privacy, as all computations can be performed locally without reliance on external cloud-based models.

Tableau 2.8 $F_1$ score results of the first 4 classes. The mention "(b)" means that the model was trained on blinded datasets

| | **O** | **Cp** | **FC** | **D** |
|---|---|---|---|---|
| GPT-4 | 0.99 | 0.93 | 0.99 | 0.97 |
| bi-LSTM (v1) | 0.99 ± 0.00 | 0.18 ± 0.25 | 0.31 ± 0.05 | 0.44 ± 0.13 |
| bi-LSTM (v2) | 0.97 ± 0.00 | 0.10 ± 0.12 | 0.44 ± 0.12 | 0.25 ± 0.12 |
| DistilCamemBERT | 0.99 ± 0.00 | 0.67 ± 0.10 | 0.56 ± 0.08 | 0.50 ± 0.04 |
| Camembert-bio | 0.99 ± 0.00 | 0.58 ± 0.10 | 0.51 ± 0.07 | 0.47 ± 0.08 |
| **Camembert-bio (ComNum)** | 0.99 ± 0.00 | 0.65 ± 0.11 | 0.81 ± 0.09 | 0.65 ± 0.17 |
| **Model 1** | 0.99 ± 0.00 | **0.67 ± 0.08** | 0.79 ± 0.10 | 0.64 ± 0.08 |
| GloVe + SVM | 0.98 ± 0.00 | 0.00 ± 0.00 | 0.43 ± 0.04 | 0.15 ± 0.15 |
| bi-LSTM (v1)(b) | 0.99 ± 0.00 | 0.61 ± 0.23 | 0.75 ± 0.08 | 0.74 ± 0.08 |
| bi-LSTM (v2)(b) | 0.99 ± 0.00 | 0.38 ± 0.20 | 0.53 ± 0.12 | 0.68 ± 0.14 |
| Camembert-bio(b) | 0.99 ± 0.00 | 0.45 ± 0.25 | 0.65 ± 0.09 | 0.63 ± 0.16 |
| **Model 2** | 0.99 ± 0.00 | 0.56 ± 0.12 | **0.84 ± 0.05** | **0.92 ± 0.04** |

Tableau 2.9 $F_1$ score results for the last classes. The mention "(b)" means that the model was trained on blinded datasets

| | **SO2** | **AGPR** | **G** | **CIA/CIV** |
|---|---|---|---|---|
| GPT-4 | 0.99 | 0.99 | 0.80 | 0.99 |
| bi-LSTM (v1) | 0.52 ± 0.06 | 0.79 ± 0.03 | 0.59 ± 0.10 | 0.51 ± 0.10 |
| bi-LSTM (v2) | 0.34 ± 0.15 | 0.50 ± 0.23 | 0.26 ± 0.11 | 0.48 ± 0.22 |
| DistilCamemBERT | 0.64 ± 0.03 | 0.90 ± 0.07 | 0.50 ± 0.07 | 0.67 ± 0.10 |
| Camembert-bio | 0.69 ± 0.02 | 0.91 ± 0.09 | 0.50 ± 0.06 | 0.65 ± 0.07 |
| Camembert-bio (ComNum) | 0.72 ± 0.05 | 0.98 ± 0.11 | 0.85 ± 0.02 | 0.96 ± 0.11 |
| **Model 1** | 0.72 ± 0.04 | **0.98 ± 0.04** | 0.50 ± 0.03 | 0.95 ± 0.08 |
| GloVe + SVM | 0.03 ± 0.04 | 0.23 ± 0.02 | 0.00 ± 0.00 | 0.22 ± 0.04 |
| bi-LSTM (v1)(b) | 0.82 ± 0.04 | 0.92 ± 0.02 | 0.76 ± 0.07 | 0.67 ± 0.08 |
| bi-LSTM (v2)(b) | 0.71 ± 0.07 | 0.78 ± 0.13 | 0.80 ± 0.10 | 0.82 ± 0.11 |
| Camembert-bio(b) | 0.73 ± 0.06 | 0.89 ± 0.13 | 0.84 ± 0.11 | 0.69 ± 0.15 |
| **Model 2** | **0.85 ± 0.03** | 0.97 ± 0.05 | **0.98 ± 0.03** | **0.99 ± 0.03** |

## 2.7 Limitations

It should be noted that the blind dataset method's heavy dependence on contextual data can lead to less than optimal results when applied to numerical data in medical notes, especially when these notes contain scarce or unclear contextual indicators. To illustrate, let's consider the following excerpt from a note :

*"[...] FR 21 FC 100-110 FR 50 [...]."*

and it's blinded version :

*"[...] FR* ***nombre*** *FC* ***nombre*** *FR* ***nombre*** *[...]."*

This note contains a list of physiological measurements with abbreviations. The initial mention of *FR* likely denotes the breathing frequency (fréquence respiratoire) given that 21 is usually too low for a shortening fraction. The second occurrence of *FR* is expressed as a percentage, therefore it likely refers to the shortening fraction (fraction de raccourcissement). Here, it wasn't the context but the values themselves that offered the necessary distinctions. Without these specific indicators, a text with masked numerical values would struggle to achieve precise classification in such instances. This situation is common in the medical field, characterized by the use of terminology that often involves implied or subtly expressed references. It is noteworthy to mention that even GPT-4 failed to perform the right classification here. In fact it is one of the failure case scenario discussed in 2.6.3.

Moreover, this approach cannot be utilized for a question-answering task that involves numbers, as the model would never be exposed to numerical values.

The primary limitation of this project is the decision to use Camembert-bio instead of large language models (LLMs). This choice was made as a trade-off between performance and computational resource constraints. A direct comparison with GPT-4 confirms its overall superior performance ; however, its computational requirements are several orders of magnitude higher.

We acknowledge that the experiments conducted in this study focus on a specific task and are based on a single dataset. However, we believe that the methodologies proposed in this work are generalizable and can be effectively applied to similar scenarios involving limited datasets. Furthermore, the second phase of our approach relies on predefined thresholds specifically

designed for our medical task—detecting cardiac failure. This inherently limits its generalizability compared to advanced LLMs, which are built to autonomously interpret and adapt to diverse contexts without requiring manually set thresholds, thereby improving both scalability and accuracy. However, we acknowledge this limitation, as maintaining control over these parameters is essential due to the rigor and sensitivity required in this medical application.

## 2.8 Application Perspective

The Clinical Decision Support System (CDSS) at CHUSJ remains under development. By integrating this NLP algorithm for clinical text—with previously developed algorithms for detecting the absence of heart failure Le, Noumeir, Rambaud, Sans & Jouvet (2022b, 2023b); Le, Jouvet & Noumeir (2023a), hypoxemia detection Sauthier, Tuli, Jouvet, Brownstein & Randolph (2021) and chest X-ray analysis Zaglam, Jouvet, Flechelles, Emeriaud & Cheriet (2014); Yahyatabar, Jouvet & Cheriet (2020), our next step is to deploy the combined CDSS (merging all algorithms) into the cyberinfrastructure of the pediatric intensive care unit (PICU) at CHUSJ for the early diagnosis of acute respiratory distress syndrome (ARDS). Following the completion of integration with the PICU's e-Medical infrastructure, we will prospectively assess the system's ability to detect ARDS accurately. This evaluation will guide further system refinements and improvements

## 2.9 Conclusion

This study aimed to optimize training methodologies for Camembert-bio, specifically focusing on classifying numerical values within small medical datasets. Despite the promise of transformer-based models, our investigation revealed significant challenges when applied to smaller datasets, including difficulty in learning meaningful word relationships and handling numbers effectively without specific prefinetuning. To address these issues, we introduced two straightforward yet impactful strategies : an adapted version of the Label Embedding for Self-Attention (LESA) technique and a number-blinding method called Blind Dataset specifically designed to handle numbers. Both methods are informed by insights from healthcare professionals to ensure

their applicability in medical contexts. Evaluation of these strategies demonstrated notable enhancements in Camembert-bio's performance. LESA narrowed the performance gap between Camembert-bio and CamemBERT-bio + ComNum by improving overall $F_1$ scores by 0.12, while the Blind Dataset resulted in significant improvements across all models tested, regardless of complexity and architecture. Notably, the combination of Camembert-bio, LESA, and Blind Dataset (referred to as Model 2) consistently outperformed other approaches, achieving a $F_1$ score margin of 0.07 over the second-best model, CamemBERT-bio + ComNum, across various categories. However, while our results remain lower than those of GPT-4, they are still comparable despite the substantial difference in model size. This highlights the effectiveness of our approach in balancing performance and computational efficiency. We believe this work has the potential to serve as a viable alternative to large-scale LLMs for medical note processing in terms of computational resources and limited data availability constraint in hospital environment.

Additionally, we developed an algorithm capable of assessing the criticality of each numerical value identified by our models using established medical benchmarks. This advancement greatly assists healthcare practitioners by simplifying the evaluation of essential health indicators, thereby enabling more informed clinical decisions.

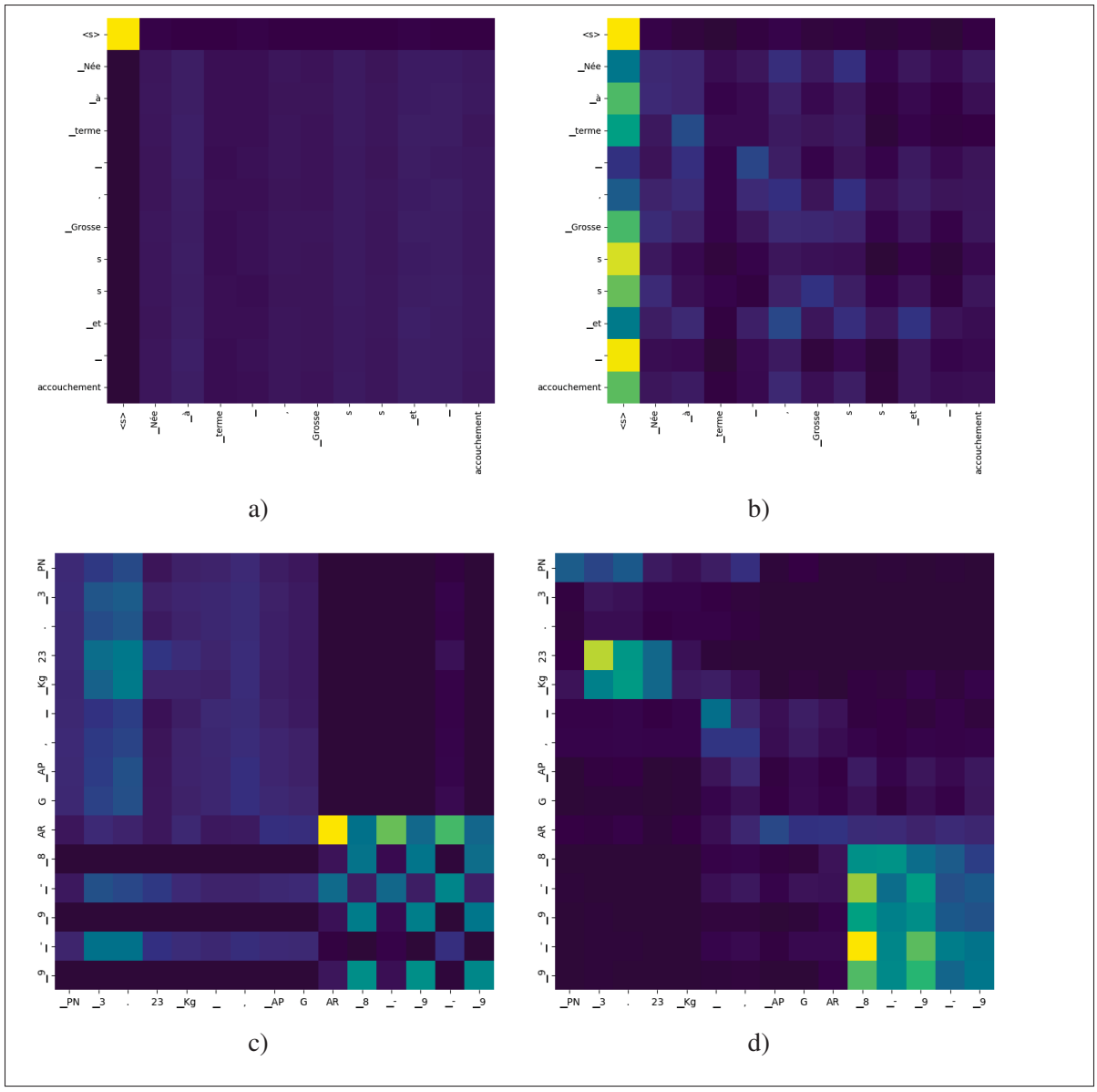


Figure 2.4 Representative Attention Head from **CamemBERT-bio** and **Model 1** last Attention Layer. It was derived from the sentence : *"Née à terme, Grossess et accouchement sans complication PN 3.23 Kg, APGAR 8-9-9."* (a) Within this plot, we have the Attention generated by **CamemBERT-bio** on the initial segment of the sentence *"Née à terme, Grossess et accouchement sans complication*. (b) Within this plot, we have the Attention generated by **Model 1** on the same sentence segment. **Model 1** is **CamemBERT-bio** augmented with the implementation of LESA. (c) Within this plot, we have the Attention generated by **CamemBERT-bio** on the last segment of the sentence *"PN 3.23 Kg, APGAR 8-9-9."*. (d) Within this plot, we have the Attention generated by **Model 1** on the same sentence segment. We used **Seaborn** color map *viridis*

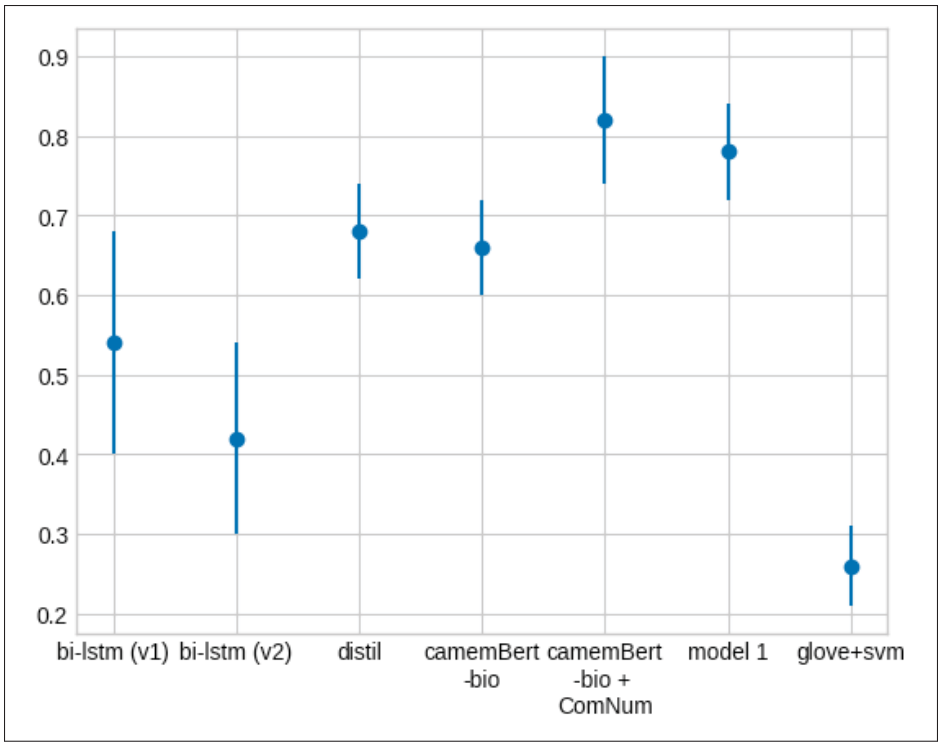


Figure 2.5 Comparison of $F_1$ scores with their standard deviations for all models trained without the Blind Dataset except GPT-4. The plot reveals that Model 1 outperforms all other baselines except for CamemBERT-bio + ComNum

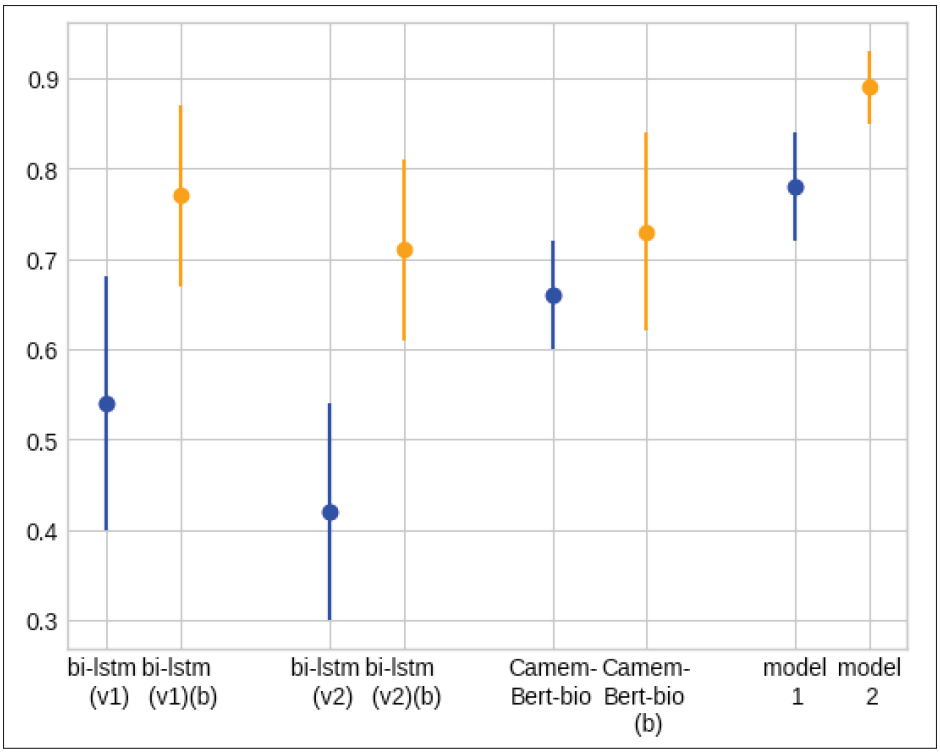


Figure 2.6 Comparison of $F_1$ scores for each model when training with Normal (color blue) and Blind Dataset (color orange). The figure highlights the improvement in terms of $F_1$ scores and standard deviations achieved by the Blind Dataset technique. Model 2 reaches the highest performances and best standard deviations

# CHAPITRE 3

# MULTI-OBJECTIVE REPRESENTATION FOR NUMBERS IN CLINICAL NARRATIVES : A CAMEMBERT-BIO-BASED ALTERNATIVE TO LARGE-SCALE LLMS

Boammani Aser Lompo[1] , Thanh-Dung Le[1] , Philippe Jouvet[2] , Rita Noumeir[1]

[1] Electrical Engineering Department, École de Technologie Supérieure,
1100 Notre-Dame Street West, Montreal, Quebec, Canada H3C 1K3
[2] CHU Sainte-Justine Research Center, CHU Sainte-Justine,
3175 Chemin de la Côte-Sainte-Catherine, Montreal, Quebec, Canada H3T 1C5

Article Submitted to the «IEEE Transactions on Artificial Intelligence» Journal
on March 7th 2025.

## 3.1 Abstract

The processing of numerical values is a rapidly developing area in the field of Language Models (LLMs). Despite numerous advancements achieved by previous research, significant challenges persist, particularly within the healthcare domain. This paper investigates the limitations of Transformers Models in understanding numerical values. *Objective :* The aim of this research is to categorize numerical values extracted from medical documents into eight specific physiological categories using CamemBERT-bio. *Methods :* In a context where scalable methods and Large Language Models (LLMs) are emphasized, we explore the option of lifting the limitations of transformer-based models. We examine two strategies : fine-tuning CamemBERT-bio on a small medical dataset with the integration of Label Embedding for Self-Attention (LESA), and combining LESA with additional enhancement techniques such as Xval. Given that CamemBERT-bio is already pre-trained on a large medical dataset, the first approach aims to update its encoder with newly added label embeddings technique, while the second approach seeks to develop multiple representations of numbers (contextual and magnitude-based) to achieve more robust number embeddings. *Results :* As anticipated, fine-tuning the standard CamemBERT-bio on our small medical dataset did not improve F1 scores. However, significant improvements were observed with CamemBERT-bio + LESA, resulting in an over 13% increase. Similar enhancements were noted when combining LESA with Xval, outperforming conventional

methods and giving comparable results to those of GPT-4 *Conclusions and Novelty :* This study introduces two innovative techniques for handling numerical data, which are also applicable to other modalities. We illustrate how these techniques can improve the performance of Transformer-based models, achieving more reliable classification results even with small datasets.

***Clinical and Translational Impact Statement*— The use of CamemBERT-bio to numerical value classification holds significant promise, particularly given the limited availability of clinical data. By developing multiple representation-based embeddings for numbers, we enhance our models robustness and generalizability. This approach enables the creation of diagnostic algorithms from unstructured medical notes, as our model provides the necessary structure. The methodology can be extended to various other NLP challenges and modalities, leveraging the core idea of constructing embeddings based on multiple data representations. Additionally, this work addresses the common healthcare issue of handling small and imbalanced datasets.**

## 3.2 Introduction

A major challenge in processing medical texts is understanding numerical data. Numbers are ubiquitous in medical documents, appearing in various forms such as physiological measurements, DNA coding sequences, event occurrences (e.g., number of births, cardiac failures), durations (e.g., pregnancy length, medication duration), and medical protocol codes (e.g., G1P2 standing for Gravida 1 Para 2). In the context of Clinical Decision Support Systems (CDSS), it is crucial to recognize and interpret the meaning of these numbers within their specific context to determine if they indicate a potential problem. For example, in the text "Patient agé de 12 ans, rythme cardiaque de 130 bpm," it is essential to identify "12" as the age of the patient and "130" as the heart rate, and then understand that a heart rate of "130" is problematic for a 12-year-old patient.

Most machine learning efforts in the medical domain must contend with small and imbalanced datasets due to the high cost of data annotation in healthcare. Additionally, medical notes are often unstructured, filled with medical jargon, and sometimes contain incomplete sentences.

This necessitates handling shorthand notations, typographical errors, and various forms of textual noise commonly found in medical documentation, making systematic approaches like case-by-case algorithms ineffective. In addition, numerical values in medical datasets typically span a wide range. For example, the dataset provided by Centre Hospitalier Universitaire Sainte-Justine (CHUSJ) for our study includes numerical values ranging from $10^{-4}$ to $10^{5}$. Therefore, careful gradient management is essential, as a single large outlier can negatively impact the overall performance of the model.

This paper contributes to a broader project aimed at leverage Language Models for diagnosing cardiac failure. According to the National Heart, Lung, and Blood Institute (2023), cardiac failure occurs when the heart is unable to pump sufficient blood to meet the body's needs, either due to inadequate filling or diminished pumping capacity. Cardiac failure presents through a range of physiological indicators, including blood pressure, ventricular gradients, and ejection fractions. Our research focuses on classifying numerical values extracted from medical notes into one of eight specific physiological categories using CamemBERT-bio Touchent *et al.* (2023).

### 3.2.1 Goal statement

Our objective is to train CamemBERT-bio on a limited and uneven dataset with the aim of accomplishing the following outcome :

**Input medical note** :

> *"Norah, BB né à 37+6 par AVS, APGAR 8-8-8. Dx Anténatale : d-TGV avec septum intact, sat 50-65% à la naissance."*

**Expected output** :

| Value | Attributes | Unit |
|---|---|---|
| 37 + 6 | Divers (âge du patient) | semaines |
| 8 – 8 – 8 | score APGAR | |
| 50 – 65 | saturation en oxygène | % |

### 3.2.2 Contribution

In this work, we outline the subsequent contributions :

- We demonstrate that CamemBERT-bio + LESA as introduced in Lompo *et al.* (2025) can be further enhanced through language modeling prefinetuning. This prefinetuning does not improve CamemBERT-bio alone, indicating that integrating the Label Embedding for Self-Attention (LESA) technique necessitates pretraining to adjust the model's parameters. Prefinetuning on a small medical dataset significantly improves the F1 score for CamemBERT-bio + LESA.
- We develop a new model by extending finetuned CamemBERT-bio + LESA with the Xval architecture Golkar *et al.* (2023). The main idea of Xval is to incorporate the magnitude of numbers into their embeddings before the attention layers. This new model provides a multi-objective representation of numbers, enhancing generalizability and robustness.
- We assess and juxtapose our tailored model against conventional counterparts including LLMs such as GPT-4. Findings demonstrate that our model surpasses standard approaches and adeptly manages datasets characterized by limited size and imbalance. Additionally, our model delivers performance on par with GPT-4 despite the significant disparity in parameter size.

## 3.3 The task and the dataset

We used the exact dataset as in the previous chapter. The number distribution is displayed in Figure 3.1. All the numbers beyond 500 are regrouped because they are scattered from 500 to 35000.

## 3.4 A few limitations of end-to-end learning for LLM in medical datasets

Most pretrained Transformers Foundational Models such as GPT-4, LLAMA and BERT are pretrained in a end-to-end paradigm. End-to-end learning involves training a model to perform a task from raw input to final output, without intermediate steps or feature engineering. While this is

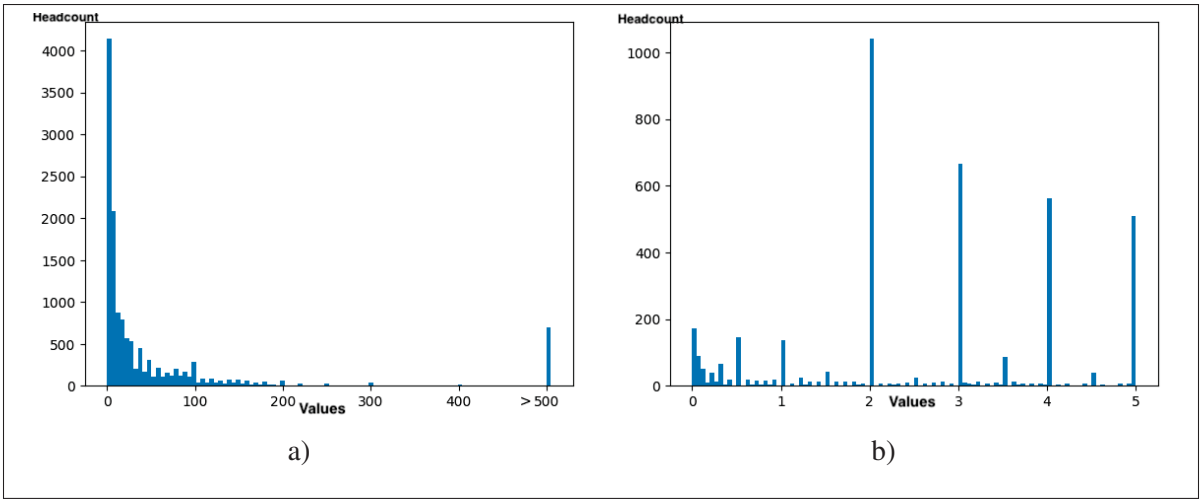


Figure 3.1 (a) Overlook to the whole numbers distribution (b) A focus on the small numbers distribution. Non integers numbers are effectively represented in the dataset

the standard approach for training most Transformers, our work faces the significant limitation of working with small and imbalanced datasets, unlike many state-of-the-art approaches mentioned earlier. This section explores other major limitations of end-to-end training for Transformers Models in the medical domain.

### 3.4.1 The limitations of Tokenization

A fundamental aspect of recent transformer-based models revolves around the introduction of highly efficient tokenization techniques, notably exemplified by WordPiece Wu *et al.* (2016) and SentencePiece Kudo & Richardson (2018). These tokenizer variants share a common trait : the capability to deconstruct intricate or infrequent words into subsequences comprising their constituent characters, an important ability in managing Out-of-vocabulary (OOV) words. Let's consider the word **"working"**, which undergoes tokenization into **"work" + "ing"**. This resultant sequence forms a good starting point to model the word, as **"work"** and **"ing"** independently hold semantic significance and individual meaning. However, as mentionned by Loukas *et al.* (2022), we claim that this approach exhibits limitations when dealing with numerical values. For instance, consider the instance of **"123.7g/L"**, which would be disassembled into **"12" + "3"** + . + **"7" + "g/L"**. This segmentation raises several issues. Predominantly, in many classification

tasks, only the first token of a segmented word is subject to classification. Only **“12"** would be classified in our example. However, **“12"** is markedly ambiguous, vastly deviating from the initial term’s inherent meaning. Furthermore, this method of breaking down numerical tokens might fail to capture their magnitudes. The difference in magnitude between **“12"** and **“123.7"** is substancial. While the self-attention mechanism and index position embedding hold the potential to discern this magnitude discrepancy, we claim that these mechanisms alone are insufficient.

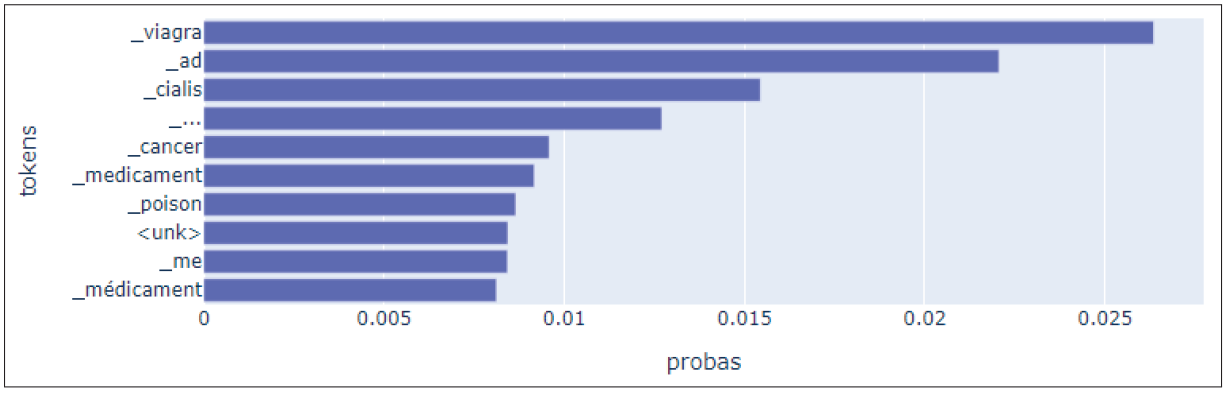


Figure 3.2 Output of a completion task by **CamemBERT-bio**. The input sentence is *“Patient en détresse respiratoire, gradient VG-VD ad ⟨ mask ⟩ mmgh."*. The figure contains the possible words to fill the mask with their corresponding probabilities

To illustrate this point, let’s examine CamemBERT-bio’s performance in a task focused on completing sentences that contain numerical values (Figure 3.2). A word completion task involves taking a sentence and replacing one or more words with the token ⟨mask⟩. The model is then asked to predict the missing words. The input sentence, *“Patient en détresse respiratoire, gradient VG-VD ad ⟨ mask ⟩ mmgh."*, was crafted to align with the language style of our target dataset. It becomes evident from this illustration that the model struggles to provide meaningful numerical predictions as the predicted tokens are not even numbers.

### 3.4.2 The structure issue

Some earlier research focused on studying transformer-based models to understand how they function and evaluate their ability to recognize different syntactic relations. Clark *et al.* (2019) revealed that, surprisingly, without explicit supervision on syntax and co-reference, certain heads in the BERT model specialize in grammatical tasks such as finding direct objects of verbs, determiners of nouns, objects of prepositions, and objects of possessive pronouns with more than 75% accuracy. This reveals that the structure of sentences matters for these models. However, our medical notes landscape stands apart in this regard. A significant proportion of sentences within these notes lack well-defined structures, with instances wherein clear subjects and verbs are even absent, as aptly exemplified by the following illustrations :

- "*CARDIOMYOPATHIE 1) S'est présenté le 28/04 à l'hôpital - Admis ad 07/05 aux soins [...] Depuis, céphalée on/off, fatigue, essoufflement.[...] Vomissement le soirx 5 jours. Pincement thoracique à 2-3 reprises depuis 1mois. l'a réveillé 1 fois.*"
- "*ATCDp : #PNA droite à 1 mois de vie et en 2004 Atrophie/asymétrie rénale 2aire Vu en néphro : Écho (11/2014) : Rein Droit : 8.8 cm, contour supérieur légèrement irrégulier + vague zone kystique, de 5 mm, [...].*"
- *"[...] FR 21 FC 100-110 FR 50 [...]."*

As noted in Lompo *et al.* (2025), the last excerpt illustrate the lack of contextual information in sentence structures. This excerpt reads a sequence of physiological indicators. The initial reference to *FR* most likely indicates the respiratory rate (fréquence respiratoire) since a value of 21 is usually too low for the shortening fraction (fraction de raccourcissement). Conversely, the second occurrence of *FR* refers to a value above 50, suggesting it pertains to the shortening fraction. In this case, it was the values themselves, rather than the context, that provided the necessary distinctions, highlighting a flaw in the sentence structure. This issue echoes the challenges of tokenization, emphasizing the need for a thorough understanding of the numerical values. However, it is worth mentioning that the impact of this issue is very limited on advanced LLMs such as LLAMA, GPT-4, Claude that are engineered to be able to work with data

without requiring any preprocessing. Nevertheless, their ability in handling the lack of structure significantly on their prompts Hu *et al.* (2024).

## 3.5 Methodology

The methodology adopted in this study involves developing a multi-objective representation of numbers to achieve a more generalizable and robust model Collobert & Weston (2008). Our representations are obtained through the Masked Language Modelling (MLM) Loss and the Number Prediction Loss. The application of these two representations result in the two following models :

- CamemBERT-bio + LESA prefinetuned via MLM on our small medical dataset
- CamemBERT-bio + LESA + Xval trained with both MLM loss and Number Prediction Loss.

We anticipate that this multi-objective training approach will produce number representations that incorporate both contextual and magnitude information.

### 3.5.1 Masked Language Modelling (MLM) objective with LESA

We summarize the model description from Lompo *et al.* (2025), the previous chapter. Let's consider a text $\boldsymbol{t} = [token_1, token_2, \cdots, token_L]$ of $L$ tokens, with initial embeddings $[\boldsymbol{x}_{token_1}, \cdots, \boldsymbol{x}_{token_L}]$. These initial embeddings, prefixed with the special token $[CLS]$, are fed into the transformer layers in the following manner :

$$\boldsymbol{X} = [\boldsymbol{x}_{CLS}, \boldsymbol{x}_{token_1}, \cdots, \boldsymbol{x}_{token_L}] \in \mathbb{R}^{(L+1)\times D} \tag{3.1}$$

where $D$ is the dimension of the embedding space.

For every individual class, some descriptive keywords are selected, and their initial embeddings are computed and then averaged. This computation produces an embedding matrix, represented

as $\boldsymbol{X}^l \in \mathbb{R}^{n \times D}$, where $n$ corresponds to the number of classes. This matrix includes unique label embeddings for each class.

Subsequently, the authors compute the cross-attention between $\boldsymbol{X}^l$ and $\boldsymbol{X}$, $\boldsymbol{A}^l \in \mathbb{R}^{n \times (L+1)}$, and the self-attention of $\boldsymbol{X}$ **Self-attention** $\in \mathbb{R}^{(L+1) \times (L+1)}$.

Then, after computing the cosine similarity matrix **CoSim** = $\text{norm}(\boldsymbol{A}^l)^T \text{norm}(\boldsymbol{A}^l)$, we therefore get the enhanced self-attention :

$$\textbf{New-Self-attention} = \textbf{Self-attention} + \textbf{CoSim} \tag{3.2}$$

In this work, we use this approach to address the limited size of our dataset as it reduces the risks of overfitting while making the model more robust. We then train the model via the Masked Language Modelling (MLM) task as a prefinetuning step. The MLM task involves randomly masking tokens and asking the model to predict the masked tokens, thereby training the model to build contextual representations of tokens.

### 3.5.2 Number Prediction Objective or Xval

This method consists of four steps, as described in Golkar *et al.* (2023). Let $t$ be an input text containing both numbers and words.

1. All numeric values in $t$ are replaced with a placeholder keyword *NUM*, producing a masked text $t_{\text{masked}}$. The original numbers are saved for later use.
2. The masked text $t_{\text{masked}}$ is tokenized and embedded, resulting in a matrix of contextual embeddings $h_{\text{masked}}$.
3. The embedding vectors in $h_{\text{masked}}$ corresponding to the masked numeric positions are then multiplied by their original numerical values, producing a new matrix $h$.
4. This modified embedding matrix $h$ is fed into the encoder of the model.

The full pipeline is illustrated in Figure 3.3.

By masking all numbers with a common placeholder, we solve the tokenization limitation mentioned earlier. The objective is to train the model to encode not only the context but also the **magnitude** of each number — a crucial feature, as numerical scale often serves as a strong signal for classification. For instance, heart rate values are typically higher than those in the APGAR score class.

### 3.5.3 Multi Objective Loss

Working with multiple objectives losses require to merge them into one single loss

$$L = \sum w_i L_i \tag{3.3}$$

where $w_i$ represents the weight of the loss $L_i$. In our case we have two losses $L_1$ which corresponds to the MLM objective and $L_2$ which corresponds to the Number Prediction Objective. Our training dataset $\mathcal{D}$ is made of couples $(x, y_1, y_2)$ where $x$ is a sequence of tokens with one token masked using the keyword *Mask*, $y_1$ is the index of the masked token in the model vocabulary, and $y_2$ is the value of the masked token if it was a number. Otherwise $y_2$ is set to 1. Figure 3.3 illustrates the inference process of the model. In this illustration, the first masked token is decoded as the standard word "Saturation," and the default number 1 is assigned to it. The second masked token is decoded as the special token "NUM," and the number 65 is predicted by the number prediction head.

In Kendall *et al.* (2018), the authors propose an automatic way to compute each loss weight $w_i$. For the MLM objective, the predicted token distribution is usually given as a Softmax :

$$p(z|f_1(x,\Theta),\sigma_2) = \text{Softmax}\left(\frac{1}{\sigma_1^2} f_1(x,\Theta)\right)_z \tag{3.4}$$

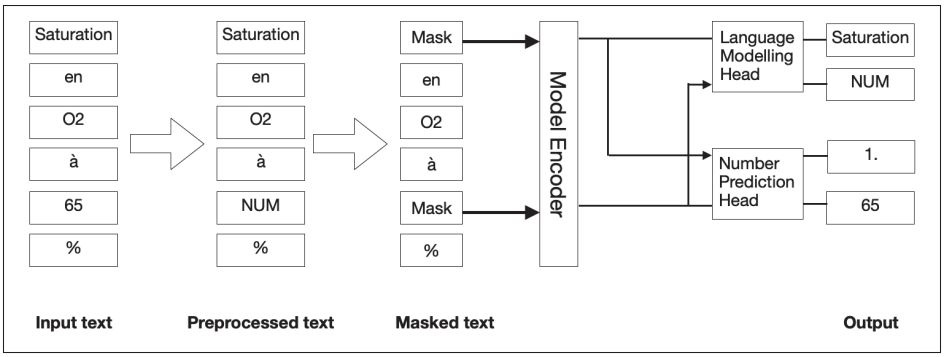


Figure 3.3 Brief illustration of the inference process using multiple representations

where $x$ is the input sequence, $f_1(x, \Theta)$ is the vector of logits outputted by the Language Head, $\Theta$ is the model parameters, and $\sigma_1$ is artificially added as a *temperature* parameter. Then, after a few approximations, they get to write the log likelihood as

$$\log p(z|f_1(x,\Theta)) \propto \frac{1}{\sigma_1^2} \log \text{Softmax}\left(f_1(x,\Theta)\right)_z - \log \sigma_1 \tag{3.5}$$

Here, one may recognize the cross entropy loss with its weight $\frac{1}{\sigma_1^2}$ and a regularization term $\log \sigma_1$. Therefore we take $L_1$ as the cross entropy loss function and we define $w_1 = \frac{1}{\sigma_1^2}$.

For the number prediction task, the authors predict the masked number using a Gaussian distribution :

$$p\left(s|f_2(x,\Theta)\right) = \mathcal{N}\left(f_2(x,\Theta)|\sigma_2^2\right) \tag{3.6}$$

where $f_2(x, \Theta)$ is Number Prediction Head's output, and $\mathcal{N}$ refers to the Gaussian probability distribution function and $\sigma_2$ its standard deviation. This leads to the log likelihood

$$\log p(s|f_1(x,\Theta)) \propto -\frac{1}{2\sigma_2^2}\left(s - f_2(x,\Theta)\right)^2 - \log \sigma_2 \tag{3.7}$$

Then one can recognize the square error $(s - f_2(x, \Theta))^2$ with its weight $\frac{1}{2\sigma_2^2}$ and a regularization term $\log \sigma_2$. Therefore, we take $L_2$ as the square error loss and we define $w_2 = \frac{1}{2\sigma_2^2}$.

Finally, when we sum up everything, we get the following loss function :

$$\begin{aligned} L(x, y_1, y_2) &= -\frac{1}{\sigma_1^2} \log \text{Softmax}\left(f_1(x, \Theta)\right)_{y_1} + \frac{1}{2\sigma_2^2} \left(y_2 - f_2(x, \Theta)\right)^2 \\ &\quad + \log \sigma_1 + \log \sigma_2 \\ &= \frac{1}{\sigma_1^2} L_1(x, y_1) + \frac{1}{2\sigma_2^2} L_2(x, y_2) + \log \sigma_1 + \log \sigma_2 \end{aligned} \tag{3.8}$$

More precisely the global objective function is

$$\mathcal{L}(\Theta) = \mathbb{E}_{\mathcal{D}} \left[L(x, y_1, y_2)\right]$$

with respect to $\Theta, \sigma_1, \sigma_2$ Kendall *et al.* (2018). The first-order conditions with respect to $\sigma_1$ and $\sigma_2$, give

$$\sigma_1^2 = 2\mathbb{E}_{\mathcal{D}} \left[L_1(x, y_1)\right] \tag{3.9}$$

$$\sigma_2^2 = \mathbb{E}_{\mathcal{D}} \left[L_2(x, y_2)\right] \tag{3.10}$$

Since the exact numbers within a medical note are often not predictable from the context, and given the wide range of values present in the dataset (Figure 3.1), it is natural to assume that $\mathbb{E}_{\mathcal{D}} \left[L_2(x, y_2)\right]$ is significantly larger than $\mathbb{E}_{\mathcal{D}} \left[L_1(x, y_1)\right]$. Therefore by Equations 3.9 and 3.10, $w_2$ is significantly smaller than $w_1$. Our experiments confirms this assumption, placing greater emphasis on contextual representation over magnitude-based representation. As a result, we observe poor performance.

$\mathbb{E}_{\mathcal{D}}\left[L_2(x, y_2)\right]$ depends on the numbers distribution in $\mathcal{D}$ and more precisely on the range of the distribution. In order to lower the impact of this parameter we substituted $L_2$ with its log scaled version $\log(L_2)$. Additionnally, we removed the regularization terms $\log \sigma_1$ and $\log \sigma_2$ and we adopted $w_1 = w_2 = \frac{1}{2}$

$$\tilde{L}_2(x, y_2) = \left(\log(y_2 + 1) - \log\left(f_2(x, \Theta) + 1\right)\right)^2 \tag{3.11}$$

We therefore get the new multi-objective loss

$$\tilde{L}(x, y_1, y_2) = \frac{1}{2} L_1(x, y_1) + \frac{1}{2} \tilde{L}_2(x, y_2) \tag{3.12}$$

instead of 3.8.

This substitution is motivated by the following considerations. Let $\theta \in \Theta$ represents a parameter that requires updating. For simplicity, we'll denote the partial derivative with respect to $\theta$ as $\partial_\theta$. We derive the two following gradients of $L_2$ and $\tilde{L}_2$ with respect to $\theta$ :

$$\begin{aligned} \partial_\theta L_2(x, y_2) &= \partial_\theta \left\{(y_2 - f_2(x, \theta))^2\right\} \\ &= 2\partial_\theta f_2(x, \theta)\left(f_2(x, \theta) - y_2\right) \end{aligned} \tag{3.13}$$

$$\begin{aligned} \partial_\theta \tilde{L}_2(x, y_2) &= \partial_\theta \left\{\left(\log(y_2 + 1) - \log\left(f_2(x, \theta) + 1\right)\right)^2\right\} \\ &= 2\frac{\partial_\theta f_2(x, \theta)}{f_2(x, \theta) + 1} \log\left(\frac{f_2(x, \theta) + 1}{y_2 + 1}\right) \end{aligned} \tag{3.14}$$

Now, the two gradients can be compared by examining their ratio.

$$\left|\frac{\partial_\theta \tilde{L}_2}{\partial_\theta L_2}\right| = \left|\frac{1}{(f_2(x,\theta)+1)\,(f_2(x,\theta)-y_2)} \log\left(\frac{f_2(x,\theta)+1}{y_2+1}\right)\right| \tag{3.15}$$

Here, the term $\frac{1}{f_2(x,\theta)-y_2} \log\left(\frac{f_2(x,\theta)+1}{y_2+1}\right)$ is contingent on the relative comparison between the prediction $f_2(x,\theta)$ and $y_2$. However, the term $\frac{1}{f_2(x,\theta)+1}$ exhibits a dependency on the magnitude of the prediction. In simpler terms, when high numbers (outliers) are predicted, $L_2$ will significantly update the weights compared to $\tilde{L}_2$, potentially undermining the model's performance on the real data distribution. This characteristic underscores why $\tilde{L}_2$ is more effective in handling a wide range of numbers.

## 3.6 Experiments

This section introduces the baseline models, outlines our experimental configuration, and presents the results we achieved.

### 3.6.1 Our models

We present two models that incorporate the diverse solutions proposed in this research.

Model 1 : The architecture is exactly the one presented in Lompo *et al.* (2025). The training processus comprises the 2 following steps :

1. We pre-finetune the pre-trained CamemBERT-bio augmented with the Label Embedding for Self-Attention (LESA) technique on the unannotated medical notes with the Masked Language Modeling (MLM) task.
2. We then fine-tune the pre-finetuned model on the token classification task with the annotated notes.

We kept the same architecture as in Lompo *et al.* (2025)

Model 2 : We incorporated Xval architecture to Model 1, in order to have both contextual and magnitude representations of numbers.

### 3.6.2 Baselines

GPT-4 : In these experiments, we selected a subset of 200 entries from our dataset and manually provided them to ChatGPT using a few-shot learning prompting approach. The model generated responses as sequences of numerical values, each mapped to its respective category. These outputs were then aggregated and used to compute the evaluation metrics detailed below.

DistillCamemBERT : DistilCamemBERT, as described in Delestre & Amar (2022), employs Knowledge Distillation Bucila *et al.* (2006), Hinton *et al.* (2015) in order to reduce the parameters of CamemBERT. This technique is a training method where a larger, expert model (teacher) guides a smaller model (student) to mimic its behavior. DistillCamemBERT has achieved performance levels comparable to CamemBERT on various NLP tasks, providing a benchmark for evaluating our approach against parameter reduction techniques. For this experiment, we utilized the version available on Huggingface Wolf *et al.* (2019).

Camembert-bio : This transformer based model is a french version of BioBERT Lee *et al.* (2019) introduced by Touchent *et al.* (2023) and trained on a large medical corpus.

CamemBERT-bio + LESA : This model follows the training approach detailed in Lompo *et al.* (2025), which consists in incorporating LESA to CamemBERT-bio. The authors derived two models from this approach : one model trained on the raw dataset and another where numbers are replaced with a keyword placeholder before training. We will only keep the second one as it got the best performances. We will refer to this baseline as CamemBERT-bio + LESA

CamemBERT-bio + Xval : This model follows the training approach detailed in the work by Golkar *et al.* (2023), which consists in incorporating some magnitude information into the numbers embeddings. The architecture was imported from their GitHub repository.

NumBERT : As intruduced in Zhang *et al.* (2020), it is a version of the BERT model trained by substituting each numerical instance in the training dataset with its scientific notation representation, comprising both an exponent and mantissa. We used the implementation provided in their GitHub repository.

ELMO : As introduced in Sarzynska-Wawer *et al.* (2021), Embeddings from Languages Models (ELMO) is a bidirectional LSTM that is trained with a coupled language model (LM) objective on a large text corpus. ELMO is very effective in practice in medical related classification tasks. We used the implementation from their GitHub repository.

### 3.6.3 Setup and Evaluation

The performance evaluation metric utilized is the $F_1$ score, a class-wise metric particularly suitable for imbalanced datasets. For the pre-finetuning step, we trained all our models for 100 epochs as a warm up with AdamW optimizer Loshchilov & Hutter (2017) optimizer and a learning rate of $3e-5$. Then we used cosine learning scheduler until early-stopping with 4 epochs of patience. For the classification downstream task, we performed 10 sets of runs with different random seeds until early stopping. We compute the mean F1 score over the 10 runs along with their standard deviations, with all results summarized in Table 3.3. For all the methods considered in this study, we used cross-entropy loss with appropriate weights for each classes in order to compensate the distribution imbalance. However, such weights were not necessary for our proposed Model 2 which performed better without them. The training, validation, and testing datasets were structured to maintain a distribution of 70%, 15%, and 15% respectively across all eight classes. Our models' training process took 20 hours on a Tesla V100 GPU with 32 GB of memory. In contrast, training the baseline models required approximately 10 hours.

## 3.7 Results and Discussions

Table 3.3 shows that all models exhibit nearly perfect $F_1$ scores for the Out-of-class class. This was expected given its extensive representation. To evaluate overall performance, we compute the

average $F_1$ score across all classes without applying weighting, considering that the remaining classes hold equal importance in terms of size. This choice is made to ensure a straightforward evaluation without introducing unnecessary complexity. The overall $F_1$ scores of all models are presented in Table 3.1.

Tableau 3.1 Overall $F_1$ score per model

| **Model** | **Overall $F_1$ score** |
|---|---|
| GPT-4 | 0.95 |
| **ELMO** | 0.72 |
| **DistilCamemBERT** | 0.74 |
| **Camembert-bio** | 0.69 |
| **Camembert-bio + LESA** | 0.83 |
| **Camembert-bio + Xval** | 0.80 |
| **NumBERT** | 0.77 |
| **Model 1** | 0.83 |
| **Model 2** | **0.90** |

Using this ranking as our basis, it becomes evident that our two models (models 1 and 2) consistently surpass all the other baseline methods except for GPT-4

#### 3.7.0.1 Impact of PRE-FINETUNING via MLM

Table 3.2 presents a performance comparison of CamemBERT-bio and CamemBERT-bio + LESA before and after prefinetuning. The results show that Prefinetuning has no significant impact on CamemBERT-bio model, which is expected since CamemBERT-bio was already trained on diverse medical datasets similar to our current dataset, which means that most of the medical knowledge was already encoded in its parameters. However, the results show a significant performance boost for CamemBERT-bio + LESA after prefinetuning. This improvement is attributed to the LESA technique and it indicates that incorporating Label Embeddings in the

model's architecture requires fine-tuning in order to adjust the parameters of CamemBERT-bio's embedding layers. Essentially, the latent space derived from CamemBERT-bio + LESA evolves to better discriminate the eight categories, thereby enhancing the $F_1$ score.

Tableau 3.2 Impact of prefinetuning on $F_1$ score

| **Model** | **Without** | **With** |
|---|---|---|
| **Camembert-bio** | 0.69 | 0.70 |
| **Camembert-bio + LESA** | 0.79 | 0.83 |

#### 3.7.0.2 Results of MULTIPLE OBJECTIVE LOSS

:

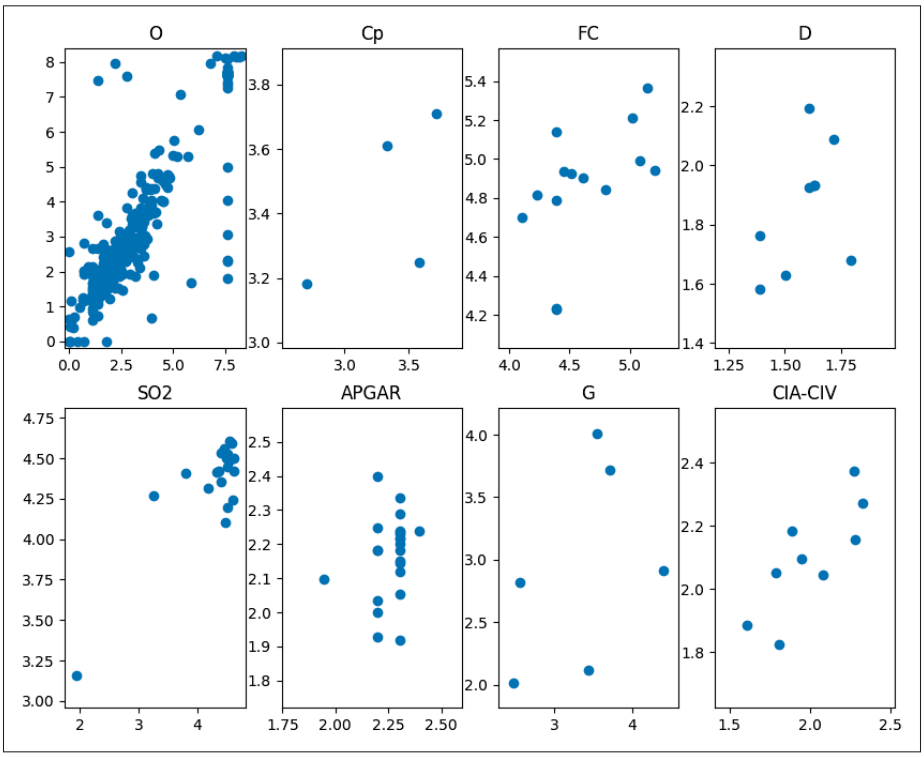


Figure 3.4 The numbers predicted by Model 2. The $x$ axis represents the ground truth values, and the $y$ axis represents the predicted values. Both axes are displayed on a logarithmic scale

The primary advantage of the number prediction loss function is that it introduces intrinsic magnitude information into the model's embeddings, an aspect significantly enhanced by incorporating Xval. Through this loss function, we aim for the model to reconstruct numbers with similar magnitudes to the original values. While predicting the exact numbers would lead to memorization, we expect the predictions to be close but not identical. As shown in Figure 3.4, Model 2 the model predictions on the $y$ axis mostly aligns with the ground truth numbers on the $x$ axis. This shows that the goal was successfully achieved. We did not include the mean square error, as it is not a key metric in our evaluation. Since our goal is not to have predictions overly close to the ground truth, emphasizing this metric could encourage the model to memorize rather than generalize.

Table 3.3 presents the F1 score performance of all studied models, broken down by category. For each category, the highest F1 score achieved by a model, excluding GPT-4, is highlighted in bold. A detailed examination of the results in Table 3.3 reveals that Model 2 outperforms all other models except GPT-4 in nearly every class. Notably, all models, including GPT-4 struggle with the *ventricular Gradient* class. This difficulty arises because this class is ambiguous and can be confused with other similar classes, such as *aortic gradient, aortic ring, pulmonary gradient, and gradient between the heart earcups*. This highlights a limitation of context-based token representation and explains the poor performance of the other models. Indeed, in certain failure cases, the surrounding context provides insufficient cues to accurately determine the gradient type. Instead, the key distinguishing factor lies in the magnitude of the gradient values. For example, gradients between heart earcups are notably lower than ventricular gradients. This magnitude-based distinction gives Model 2 a significant advantage over the other models, resulting in a higher F1 score for the gradient category.

When we subject Model 2 to the same sentence completion test as in section 3.4.1, the results are shown in Figure 3.5.

The side-by-side comparison of Figure 3.2 and 3.5 is highly illustrative. Unlike CamemBERT-bio, Model 2 clearly recognizes that a numerical value is expected in this context. Moreover, when we

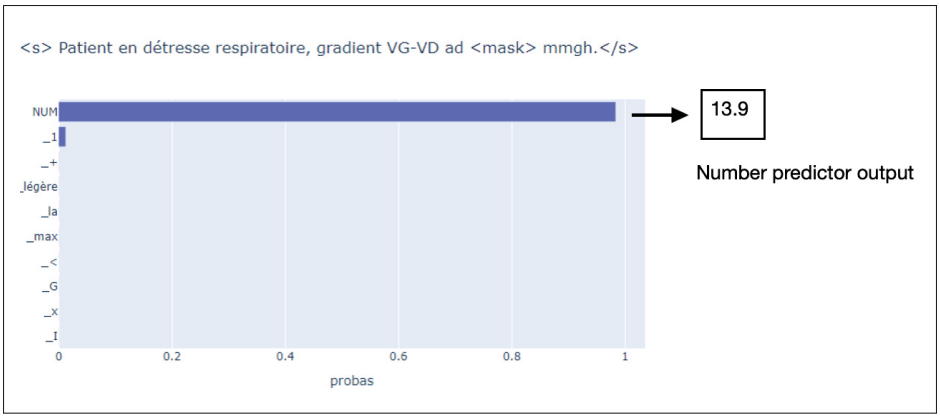


Figure 3.5 Output of a completion task by Model 2. The input sentence is *"Patient en détresse respiratoire, gradient VG-VD ad ⟨ mask ⟩ mmgh."*. The figure contains the possible words to fill the mask with their corresponding probabilities. Here the model successfully predict a number which is given by the number prediction head

Tableau 3.3 $F_1$ score results for the first 4 classes

| | **O** | **Cp** | **FC** | **D** |
|---|---|---|---|---|
| GPT-4 | 0.99 | 0.93 | 0.99 | 0.97 |
| **ELMO** | 0.99 ± 0.00 | 0.76 ± 0.25 | 0.62 ± 0.18 | 0.68 ± 0.13 |
| **DistilCamemBERT** | 0.99 ± 0.00 | 0.80 ± 0.16 | 0.62 ± 0.09 | 0.69 ± 0.14 |
| **Camembert-bio** | 0.99 ± 0.00 | 0.78 ± 0.27 | 0.68 ± 0.07 | 0.52 ± 0.11 |
| **CamemBERT-bio + LESA** | 0.99 ± 0.00 | 0.85 ± 0.07 | 0.87 ± 0.04 | 0.84 ± 0.13 |
| **CamemBERT-bio + Xval** | 0.99 ± 0.00 | 0.85 ± 0.14 | 0.64 ± 0.07 | **0.92 ± 0.06** |
| **NumBERT** | 0.99 ± 0.00 | 0.78 ± 0.10 | 0.64 ± 0.04 | 0.67 ± 0.15 |
| **Model 1** | 0.99 ± 0.00 | 0.85 ± 0.07 | 0.87 ± 0.04 | 0.84 ± 0.13 |
| **Model 2** | 0.99 ± 0.00 | **0.86 ± 0.07** | **0.89 ± 0.11** | 0.90 ± 0.08 |

take the prediction from the Number Head layer, we obtain 13.9, which falls within the correct range for this type of gradient. We attribute this improvement to the implementation of multiple objective number representations. By replacing numbers with placeholders, the unstructured and noisy text becomes clearer, making it easier for the model to learn where numbers are expected. Furthermore, incorporating LESA and Xval enhances the model's ability to more accurately predict the value range of the masked numbers. We aim to illustrate our claim by measuring the similarity of the embeddings vectors produced by various models derived from

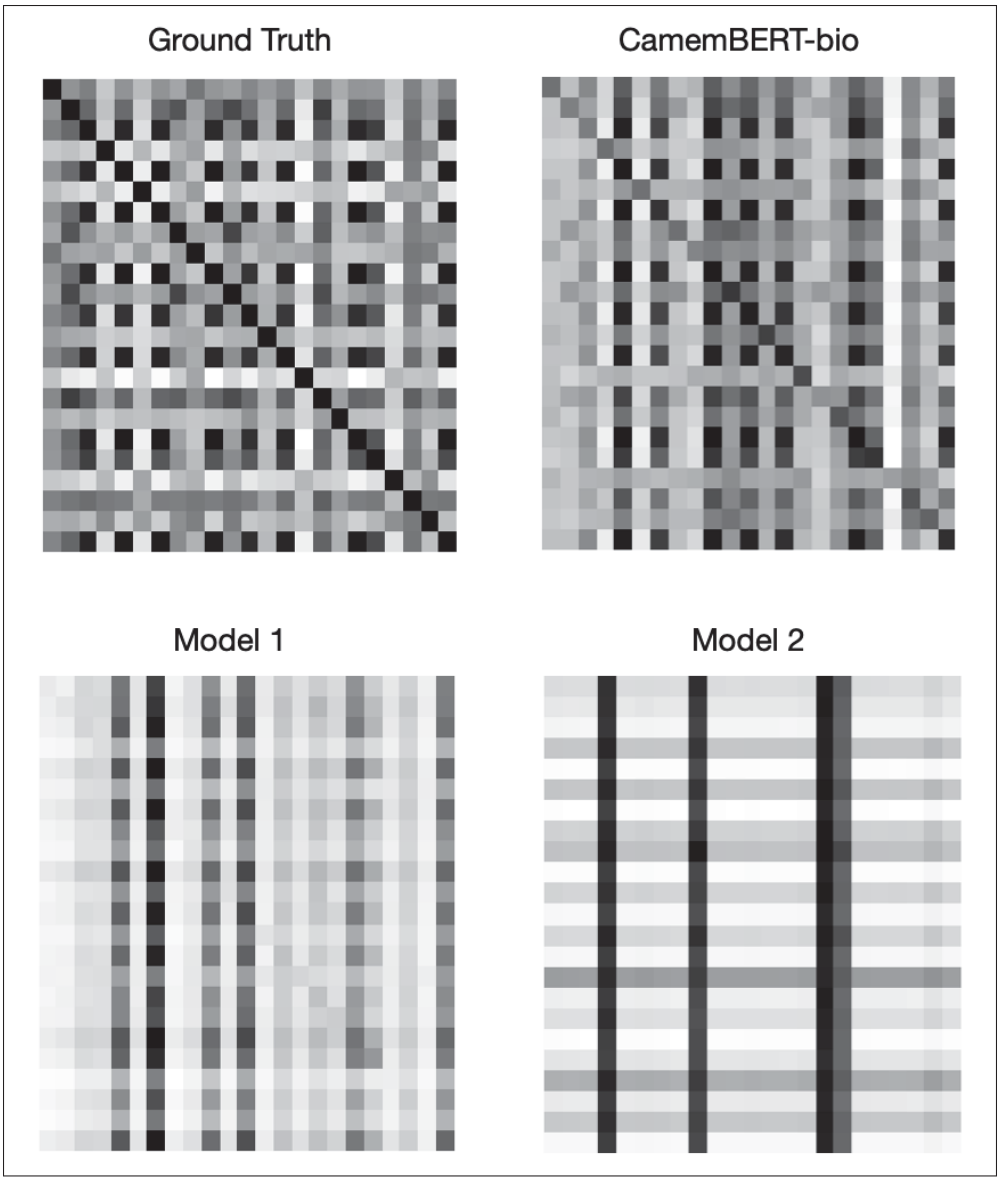


Figure 3.6 Evaluation of the impact of Multiple Objective Representation on the words embeddings. From left to right we computed the cosine similarity of multiple medical terms embeddings generated by CamemBERT-bio and itself, finetuned CamemBERT-bio, Model 1 and Model 2 respectively. The darker a cell is, the higher its value

CamemBERT-bio and those generated by CamemBERT-bio itself (our reference). Specifically, we evaluate three derived models : CamemBERT-bio fine-tuned on unannotated medical notes using masked language modeling (MLM), Model 1, and Model 2. This experiment is designed to assess how the different methodologies developed in this study influence the embedding space

Tableau 3.4 $F_1$ score results for the last 4 classes

| | **SO2** | **AGPR** | **G** | **CIA/CIV** |
|---|---|---|---|---|
| GPT-4 | 0.99 | 0.99 | 0.78 | 0.99 |
| **ELMO** | 0.83 ± 0.15 | 0.87 ± 0.12 | 0.24 ± 0.14 | 0.81 ± 0.16 |
| **DistilCamemBERT** | 0.84 ± 0.05 | 0.95 ± 0.02 | 0.23 ± 0.02 | 0.81 ± 0.15 |
| **Camembert-bio** | 0.85 ± 0.05 | 0.93 ± 0.02 | 0.22 ± 0.02 | 0.80 ± 0.08 |
| **CamemBERT-bio + LESA** | 0.91 ± 0.02 | 0.99 ± 0.01 | 0.27 ± 0.01 | 0.91 ± 0.05 |
| **CamemBERT-bio + Xval** | 0.90 ± 0.03 | 0.95 ± 0.01 | 0.23 ± 0.15 | 0.91 ± 0.08 |
| **NumBERT** | 0.83 ± 0.07 | 0.90 ± 0.13 | 0.24 ± 0.10 | 0.82 ± 0.11 |
| **Model 1** | 0.91 ± 0.02 | 0.99 ± 0.01 | 0.27 ± 0.01 | 0.91 ± 0.05 |
| **Model 2** | **0.93 ± 0.02** | **0.98 ± 0.03** | **0.61 ± 0.12** | **0.94 ± 0.06** |

of CamemBERT-bio. To conduct this analysis, we selected twenty frequently occurring medical terms from our dataset. As a similarity metric, we employed cosine similarity, following the insights from Steck, Ekanadham & Kallus (2024), which highlights that cosine similarity is a meaningful measure primarily for models trained with respect to this metric—a condition that can be facilitated by layer normalization. Since our models satisfy this prerequisite, cosine similarity was chosen as the most appropriate metric for our evaluation.

The cosine similarity graphics are displayed in Figure 3.6. In this figure, each plot represents a cosine similarity matrix $M$ in grayscale. Each row $i$ and column $j$ is indexed by the medical terms $\text{tok}_i$ and $\text{tok}_j$, respectively. The corresponding cell $M_{i,j}$ contains the cosine similarity score of the embeddings of $\text{tok}_i$ obtained by the reference model and the embedding of $\text{tok}_j$ produced by the model we're comparing. The darker a cell is, the higher the cosine similarity score.

From left to right, the matrices progressively lighten, indicating increasing divergence of the embedding vectors from the reference model. Specifically, the second matrix demonstrates that CamemBERT-bio embeddings remain largely unchanged after fine-tuning, corroborating the results in Table 3.2, which suggest that fine-tuning alone does not introduce additional knowledge into the model. In contrast, the third matrix reveals a notable difference in embeddings between Model 1 and CamemBERT-bio. This observation aligns with the analysis in Table 3.2, which suggests that fine-tuning Model 1 via MLM modifies the embedding space due to the integration

of label information into CamemBERT-bio through the LESA technique. Finally, the last matrix in Figure 3.6 exhibits the most substantial deviation from the reference, highlighting the impact of Multiple Objective Representations. This result supports the hypothesis that incorporating multiple losses leads to a more significant transformation of the embedding space compared to using LESA alone.

### 3.7.1 GPT-4 vs MODEL 2

The category-wise and overall F1 score analysis highlights GPT-4's superior performance, with an overall F1 score surpassing Model 2 by 0.05 units. However, this improvement is relatively minor given the vast disparity in model size—Model 2, with 110 million parameters, compared to GPT-4's multi-trillion parameter scale.

A failure case analysis conducted by Lompo *et al.* (2025) identified recurring misclassifications by GPT-4, particularly in the Contractibility (Cp) and Gradient (G) categories. Misclassifications in the Cp category often stemmed from ambiguous abbreviations, whereas errors in the G category frequently involved confusion between different gradient types, such as aortic and atrial/ventricular gradients, which were mistakenly linked to the interventricular gradient. These systematic errors suggest limitations in GPT-4's contextual comprehension of medical terminology. Similarly, Table 3.3 shows that Model 2 exhibits difficulties in the Cp and G categories. However, it significantly reduced the performance gap compared to the difference between CamemBERT-bio and GPT-4, narrowing it from 0.52 to 0.17 F1 score units in the G category.

From a practical perspective, our approach offers notable advantages in terms of computational efficiency and system integration. Unlike large-scale LLMs such as GPT-4 and LLaMA, Model 2 can be seamlessly deployed in medical decision-support systems while maintaining competitive performance. Its reduced computational demands enable local execution, eliminating reliance on external cloud-based models and thereby enhancing data privacy and accessibility.

## 3.8 Limitations and Application Perspectives

The main limitation of this study is the choice of CamemBERT-bio over large language models (LLMs). In a context where scalable models are emphasized, this decision was driven by the need to balance performance with computational efficiency. While a direct comparison with GPT-4 confirms its superior overall performance, its computational demands are significantly higher, making it less practical for integration into resource-constrained environments.

Additionally, we recognize that our experiments focus on a specific task and rely on a single dataset. However, we believe that the methodologies introduced in this work are broadly applicable and can be adapted to similar scenarios involving limited data availability.

The Clinical Decision Support System (CDSS) at CHUSJ is still under development. Our next step is to integrate this NLP model for clinical text analysis with previously developed algorithms for heart failure detection Le *et al.* (2022b, 2023b,a), hypoxemia detection Sauthier *et al.* (2021), and chest X-ray analysis Zaglam *et al.* (2014); Yahyatabar *et al.* (2020). Ultimately, we aim to deploy a comprehensive CDSS within the pediatric intensive care unit (PICU) at CHUSJ to facilitate early diagnosis of acute respiratory distress syndrome (ARDS). Once integrated into the PICU's e-Medical infrastructure, we will conduct a prospective evaluation of the system's ARDS detection capabilities, using the results to guide further refinements and enhancements.

## 3.9 Conclusion

This research focused on optimizing numerical understanding by CamemBERT-bio, particularly for classifying numerical values in small medical datasets. We addressed the challenges posed by tokenization and the lack of text structure, which required special handling of numbers. To tackle these issues, we implemented multiple approaches. First, we demonstrated that CamemBERT-bio + LESA could be further improved through prefinetuning with the MLM task, an improvement not seen with CamemBERT-bio alone. Additionally, we developed a multiple-objective numbers representation by combining LESA Lompo *et al.* (2025) and Xval Golkar *et al.* (2023), resulting in embeddings that incorporate both contextual and magnitude information. Rigorous evaluations

of these strategies showed significant enhancements in the performance of CamemBERT-bio. While our results do not surpass those of GPT-4, they remain comparable despite the significant difference in model size. This demonstrates the effectiveness of our approach in achieving a balance between performance and computational efficiency. We believe that this work presents a promising alternative to large-scale LLMs for medical note processing, particularly in hospital environments where computational resources are limited and data availability constraints must be considered.

# CONCLUSION ET RECOMMANDATIONS

Notre travail aborde le problème de la classification des valeurs numériques en utilisant des modèles basés sur les Transformers sur un jeu de données limité. Ce projet est crucial car il aide à structurer les vastes quantités de manuscrits générés quotidiennement par les professionnels de la santé, qui autrement nécessitent beaucoup de temps et d'efforts pour être convertis dans un format adapté aux algorithmes d'apprentissage automatique. L'automatisation de ce processus atténuera les contraintes liées à la disponibilité des données annotées, améliorant ainsi la performance des Systèmes de Soutien à la Décision Clinique (CDSS). De plus, ce travail contribue au diagnostic automatique des risques de crises cardiaques chez les patients de moins de 18 ans, soulignant ainsi l'importance et la pertinence de cette recherche.

La principale difficulté était d'éviter le surapprentissage de notre modèle de réference **CamemBERT-bio**, étant donné qu'il possède 100 000 fois plus de paramètres que de points de données. Dans le deuxième chapitre, nous avons introduit une approche centrée sur le contexte appelée Etiquettage de Label pour l'Auto-Attention (LESA), combinée avec une technique d'anonymisation des nombres. Les deux méthodes visaient à orienter l'attention du modèle vers le contexte entourant les nombres, favorisant ainsi une représentation plus robuste et généralisable. Bien que ces méthodes aient considérablement amélioré les performances du modèle, elles présentaient encore une limitation cruciale : l'absence d'information sur la magnitude des nombres encodés. Comme discuté dans le troisième chapitre, le processus de tokenisation dépouille les nombres de leurs informations de magnitude. Pour remédier à cela, nous avons proposé d'incorporer des informations de magnitude dans les représentations des nombres en utilisant la méthode Xval décrite dans Golkar *et al.* (2023), ce qui a conduit à une représentation des nombres plus robuste et généralisable.

Pour les recherches futures, nous prévoyons d'explorer le raisonnement numérique avec des modèles Transformers de taille moyenne.

Ces dernières années, les grands modèles de langage (LLMs) se sont imposés comme des outils puissants dans le domaine du raisonnement médical, offrant des applications prometteuses pour améliorer les processus de diagnostic et de prise de décision. L'un de leurs principaux atouts réside dans leur capacité à raisonner sur des connaissances implicites Conneau *et al.* (2018); Talmor, Tafjord, Clark, Goldberg & Berant (2020). Cette capacité permet aux LLMs d'exploiter l'immense quantité d'informations encodées dans leurs paramètres afin de générer des analyses, même en l'absence de données directes, imitant ainsi les processus de raisonnement humain dans des scénarios médicaux complexes.

Le raisonnement médical avec les LLMs peut être abordé à travers différents cadres de raisonnement, tels que le raisonnement abductif, inductif et révisable. Le raisonnement abductif Young, Bao, Bensemann & Witbrock (2022), souvent utilisé dans un contexte diagnostique, permet aux LLMs de formuler des hypothèses à partir de données incomplètes ou ambiguës. Le raisonnement inductif permet au modèle d'identifier des tendances à partir des observations de plusieurs patients, ce qui peut être précieux en recherche médicale. Lorsqu'il est combiné avec le raisonnement abductif, le modèle peut proposer des explications plausibles pour des observations inattendues. Le raisonnement inductif Yang *et al.* (2024) permet également aux modèles de généraliser à partir de cas spécifiques, soutenant ainsi les recommandations en matière de pronostic et de traitement. En revanche, le raisonnement révisable consiste à réévaluer la certitude d'une hypothèse à la lumière de nouvelles informations, ce qui le rend particulièrement utile pour ajuster les diagnostics à mesure que de nouveaux résultats médicaux ou événements apparaissent.

Une application du raisonnement inductif dans Lompo *et al.* (2025) pourrait être de remplacer la tâche de classification de tokens par une tâche de classification de phrases en langage naturel. Par exemple, dans la phrase « La FR est de 45% », les auteurs cherchaient à classifier « 45% » pour en comprendre la pertinence clinique. Cependant, en suivant l'approche de Talmor *et al.*

(2020), une autre méthode consisterait à formuler une question comme « 45% est-il un rythme respiratoire ? » et à laisser le modèle répondre « oui » ou « non ». Cette approche, qui repose sur une entrée en langage naturel, pourrait permettre au modèle de mieux exploiter ses connaissances internes.

Malgré ce potentiel, il n'existe actuellement aucun jeu de données spécifiquement conçu pour entraîner les modèles à ces types de raisonnement en santé. Bien que des études antérieures, telles que Talmor *et al.* (2020); Conneau *et al.* (2018); Young *et al.* (2022); Rudinger *et al.* (2020); Yang *et al.* (2024), aient proposé des méthodes de création de jeux de données pour ces approches de raisonnement, leur développement dans un contexte médical nécessiterait une expertise médicale approfondie, rendant le processus à la fois long et coûteux en ressources.

Pour améliorer leurs performances, les LLMs s'appuient également sur des techniques d'auto-amélioration. Les méthodes de Chain of Thought (CoT) Wei *et al.* (2022) guident les LLMs à structurer systématiquement leurs étapes de raisonnement, améliorant ainsi la transparence et l'interprétabilité. Les approches de Chain of Hindsight (CoH) affinent encore davantage le raisonnement en permettant aux modèles de revisiter et d'affiner leurs étapes précédentes, simulant ainsi un processus réflexif Liu, Sferrazza & Abbeel (2023). La Self-Consistency Wang *et al.* (2022) renforce la robustesse des prédictions des modèles en agrégeant plusieurs chemins de raisonnement, ce qui conduit à des résultats plus précis et fiables.

Les techniques de Few-shot learning Kojima, Gu, Reid, Matsuo & Iwasawa (2022) sont particulièrement précieuses dans le domaine médical, car elles permettent aux modèles d'apprendre de nouvelles tâches ou de s'adapter à des situations inédites avec un minimum de données annotées, souvent rares dans les domaines spécialisés. Ces techniques ne sont pas seulement complémentaires, elles peuvent aussi générer leurs propres données d'entraînement de haute qualité, créant ainsi un cycle d'amélioration auto-entretenu. De plus, elles sont généralement plus robustes et fiables que les approches supervisées du raisonnement et offrent une meilleure

interprétabilité, car elles abordent les problèmes en les décomposant et en raisonnant étape par étape.

Dans ce sens, Singhal *et al.* (2023) a montré que ces techniques sont particulièrement efficaces pour encoder les connaissances médicales, rapprochant ainsi les performances des modèles de celles des cliniciens. Cependant, comme l'a démontré Zhang, Zeng, Wang & Lu (2024), l'écart de performance entre les grands LLMs et les petits modèles de langage (SLMs) reste significatif ; même optimisés, les SLMs sont encore loin d'atteindre les capacités des LLMs. Ainsi, les grands LLMs restent essentiels pour les tâches de raisonnement de haut niveau, en particulier dans des domaines complexes comme la santé.

Par conséquent, entraîner des Transformers de taille moyenne pour atteindre des performances similaires pourrait ouvrir la porte à de nombreuses applications médicales.

## LISTE DE RÉFÉRENCES


Acar, P. (2008). *Échocardiographie pédiatrique et foetale* [Version]. doi : 10.1016/b978-2-294-70348-5.x5000-2.

Agarwal, P., Rahman, A. A., St-Charles, P.-L., Prince, S. J. & Kahou, S. E. (2023). Transformers in reinforcement learning : a survey. *arXiv preprint arXiv :2307.05979*.

Association, A. H. (2023). diagnosing-heart-failure [Format]. Repéré à https://www.heart.org/en/health-topics/heart-failure/diagnosing-heart-failure/ejection-fraction-heart-failure-measurement.

Aubertin, G., Marguet, C., Delacourt, C., Houdouin, V., Leclainche, L., Lubrano, M., Marteletti, O., Pin, I., Pouessel, G., Rittié, J.-L., Saulnier, J.-P., Schweitzer, C., Stremler, N., Thumerelle, C., Toutain-Rigolet, A. & Beydon, N. (2013). Recommandations pour l'oxygénothérapie chez l'enfant en situations aiguës et chroniques : évaluation du besoin, critères de mise en route, modalités de prescriptions et de surveillance. *Revue des Maladies Respiratoires*, 30(10), 903–911. doi : 10.1016/j.rmr.2013.03.002.

Benjamin Muller,Nathan Godey, R. C. (2022). Hands on CamemBERT [Notes de cours]. Repéré à https://camembert-model.fr/posts/tutorial_part2/.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877–1901.

Bucila, C., Caruana, R. & Niculescu-Mizil, A. (2006). Model compression. *Knowledge Discovery and Data Mining*.

Charton, F. (2021). Linear algebra with transformers. *arXiv preprint arXiv :2112.01898*.

Chen, C.-C., Huang, H.-H. & Chen, H.-H. (2021). NQuAD : 70,000+ Questions for Machine Comprehension of the Numerals in Text. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 2925–2929.

Chen, C.-C., Takamura, H., Kobayashi, I. & Miyao, Y. (2023). Improving Numeracy by Input Reframing and Quantitative Pre-Finetuning Task. *Findings of the Association for Computational Linguistics : EACL 2023*, pp. 69–77.

Chen, R., Stewart, W. F., Sun, J., Ng, K. & Yan, X. (2019). Recurrent Neural Networks for Early Detection of Heart Failure From Longitudinal Electronic Health Record Data : Implications for Temporal Modeling With Respect to Time Before Diagnosis, Data Density, Data Quantity, and Data Type. *Circulation : Cardiovascular Quality and Outcomes*, 12(10). doi : 10.1161/circoutcomes.118.005114.

Cheng, N., Yan, Z., Wang, Z., Li, Z., Yu, J., Zheng, Z., Tu, K., Xu, J. & Han, W. (2024). Potential and Limitations of LLMs in Capturing Structured Semantics : A Case Study on SRL. *International Conference on Intelligent Computing*, pp. 50–61.

Clark, K., Khandelwal, U., Levy, O. & Manning, C. D. (2019). What does bert look at ? an analysis of bert's attention. *arXiv preprint arXiv :1906.04341*.

Collobert, R. & Weston, J. (2008). A unified architecture for natural language processing : Deep neural networks with multitask learning. *Proceedings of the 25th international conference on Machine learning*, pp. 160–167.

Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S. R., Schwenk, H. & Stoyanov, V. (2018). XNLI : Evaluating cross-lingual sentence representations. *arXiv preprint arXiv :1809.05053*.

Crammer, K. & Singer, Y. (2002). On the Algorithmic Implementation of Multiclass Kernel-Based Vector Machines. *J. Mach. Learn. Res.*, 2, 265–292.

Cui, M., Bai, R., Lu, Z., Li, X., Aickelin, U. & Ge, P. (2019). Regular expression based medical text classification using constructive heuristic approach. *IEEE Access*, 7, 147892–147904.

Delestre, C. & Amar, A. (2022). DistilCamemBERT : a distillation of the French model CamemBERT. *arXiv preprint arXiv :2205.11111*.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

Ezen-Can, A. (2020). A Comparison of LSTM and BERT for Small Corpus. *arXiv preprint arXiv :2009.05451*.

Golkar, S., Pettee, M., Eickenberg, M., Bietti, A., Cranmer, M., Krawezik, G., Lanusse, F., McCabe, M., Ohana, R., Parker, L. et al. (2023). xval : A continuous number encoding for large language models. *arXiv preprint arXiv :2310.02989*.

Griffith, K. & Kalita, J. (2019). Solving arithmetic word problems automatically using transformer and unambiguous representations. *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 526–532.

Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., Mirjalili, S. et al. (2023). A survey on large language models : Applications, challenges, limitations, and practical usage. *Authorea Preprints*, 3.

Hinton, G., Vinyals, O. & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv :1503.02531*.

Hu, Y., Chen, Q., Du, J., Peng, X., Keloth, V. K., Zuo, X., Zhou, Y., Li, Z., Jiang, X., Lu, Z. et al. (2024). Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31(9), 1812–1820.

Kendall, A., Gal, Y. & Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7482–7491.

Kocaman, V. & Talby, D. (2021). Spark NLP : Natural Language Understanding at Scale. *Software Impacts*, 8, 100058. doi : https://doi.org/10.1016/j.simpa.2021.100058.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35, 22199–22213.

Kudo, T. & Richardson, J. (2018). SentencePiece : A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, pp. 66–71. doi : 10.18653/v1/D18-2012.

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. *bioRxiv*, 2023–04.

Le, T.-D., Noumeir, R., Rambaud, J., Sans, G. & Jouvet, P. (2022a). Detecting of a Patient's Condition From Clinical Narratives Using Natural Language Representation. *IEEE Open Journal of Engineering in Medicine and Biology*, 3, 142–149. doi : 10.1109/ojemb.2022.3209900.

Le, T.-D., Noumeir, R., Rambaud, J., Sans, G. & Jouvet, P. (2022b). Detecting of a patient's condition from clinical narratives using natural language representation. *IEEE open journal of engineering in medicine and biology*, 3, 142–149.

Le, T.-D., Jouvet, P. & Noumeir, R. (2023a). Improving transformer performance for french clinical notes classification using mixture of experts on a limited dataset. *arXiv preprint arXiv :2303.12892*.

Le, T.-D., Noumeir, R., Rambaud, J., Sans, G. & Jouvet, P. (2023b). Adaptation of autoencoder for sparsity reduction from clinical notes representation learning. *IEEE Journal of Translational Engineering in Health and Medicine*, 11, 469–478.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H. & Kang, J. (2019). BioBERT : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240. doi : 10.1093/bioinformatics/btz682.

Li, Y., Yao, L., Mao, C., Srivastava, A., Jiang, X. & Luo, Y. (2018). Early Prediction of Acute Kidney Injury in Critical Care Setting Using Clinical Notes. *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. doi : 10.1109/bibm.2018.8621574.

Liu, H., Sferrazza, C. & Abbeel, P. (2023). Chain of hindsight aligns language models with feedback. *arXiv preprint arXiv :2302.02676*.

Lompo, B. A., Le, T.-D., Jouvet, P. & Noumeir, R. (2025). Are Medium-Sized Transformers Models still Relevant for Medical Records Processing ? Repéré à https://arxiv.org/abs/2404.10171.

Loshchilov, I. & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv :1711.05101*.

Loukas, L., Fergadiotis, M., Chalkidis, I., Spyropoulou, E., Malakasiotis, P., Androutsopoulos, I. & Paliouras, G. (2022). FiNER : Financial Numeric Entity Recognition for XBRL Tagging. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. doi : 10.18653/v1/2022.acl-long.303.

Martin, L., Muller, B., Suá rez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D. & Sagot, B. (2020). CamemBERT : a Tasty French Language Model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. doi : 10.18653/v1/2020.acl-main.645.

Mascio, A., Kraljevic, Z., Bean, D., Dobson, R., Stewart, R., Bendayan, R. & Roberts, A. (2020). Comparative analysis of text classification approaches in electronic health records. *arXiv preprint arXiv :2005.06624*.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*.

Nguyen, P., Nguyen, K., Ichise, R. & Takeda, H. (2018). EmbNum : Semantic labeling for numerical values with deep metric learning. *ArXiv*, abs/1807.01367. Repéré à https://api.semanticscholar.org/CorpusID:49567815.

Pennington, J., Socher, R. & Manning, C. D. (2014). Glove : Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1532–1543. doi : 10.3115/v1/d14-1162.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L. (2018). Deep contextualized word representations. NAACL-HLT. *arXiv*.

Pulse. (2019). diagnosing-heart-failure [Format]. Repéré à https://www.ucsfbenioffchildrens.org/medical-tests/pulse#:~:text=Normal%20Results&text=Infants%201%20to%2011%20months,to%20115%20beats%20per%20minute.

Rudinger, R., Shwartz, V., Hwang, J. D., Bhagavatula, C., Forbes, M., Le Bras, R., Smith, N. A. & Choi, Y. (2020). Thinking Like a Skeptic : Defeasible Inference in Natural Language. *Findings of the Association for Computational Linguistics : EMNLP 2020*, pp. 4661–4675. doi : 10.18653/v1/2020.findings-emnlp.418.

Sanh, V., Debut, L., Chaumond, J. & Wolf, T. (2019). DistilBERT, a distilled version of BERT : smaller, faster, cheaper and lighter. arXiv. Repéré à https://arxiv.org/abs/1910.01108.

Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M. & Okruszek, L. (2021). Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304, 114135.

Sauthier, M., Tuli, G., Jouvet, P. A., Brownstein, J. S. & Randolph, A. G. (2021). Estimated Pao2 : A continuous and noninvasive method to estimate Pao2 and oxygenation index. *Critical care explorations*, 3(10), e0546.

Shah, S. V. (2024). Accuracy, consistency, and hallucination of large language models when analyzing unstructured clinical notes in electronic medical records. *JAMA Network Open*, 7(8), e2425953–e2425953.

Si, S., Wang, R., Wosik, J., Zhang, H., Dov, D., Wang, G. & Carin, L. (2020). Students need more attention : Bert-based attention model for small data with application to automatic patient message triage. *Machine Learning for Healthcare Conference*, pp. 436–456.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S. et al. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972), 172–180.

Steck, H., Ekanadham, C. & Kallus, N. (2024). Is cosine-similarity of embeddings really about similarity ? *Companion Proceedings of the ACM on Web Conference 2024*, pp. 887–890.

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N. & Kroeker, K. I. (2020). An overview of clinical decision support systems : benefits, risks, and strategies for success. *npj Digital Medicine*, 3(1). doi : 10.1038/s41746-020-0221-y.

Talmor, A., Tafjord, O., Clark, P., Goldberg, Y. & Berant, J. (2020). Leap-of-thought : Teaching pre-trained models to systematically reason over implicit knowledge. *Advances in Neural Information Processing Systems*, 33, 20227–20237.

Thawani, A., Pujara, J., Ilievski, F. & Szekely, P. (2021). Representing Numbers in NLP : a Survey and a Vision. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pp. 644–656.

Touchent, R., Romary, L. & de La Clergerie, E. (2023). CamemBERT-bio : a Tasty French Language Model Better for your Health. *arXiv preprint arXiv :2306.15550*.

Ullah, E., Parwani, A., Baig, M. M. & Singh, R. (2024). Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology–a recent scoping review. *Diagnostic pathology*, 19(1), 43.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wallace, E., Wang, Y., Li, S., Singh, S. & Gardner, M. (2019). Do NLP models know numbers ? probing numeracy in embeddings. *arXiv preprint arXiv :1909.07940*.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A. & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv :2203.11171*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D. et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35, 24824–24837.

Wiegreffe, S. & Pinter, Y. (2019). Attention is not not explanation. *arXiv preprint arXiv :1908.04626*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M. et al. (2019). Huggingface's transformers : State-of-the-art natural language processing. *arXiv preprint arXiv :1910.03771*.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K. et al. (2016). Google's neural machine translation system : Bridging the gap between human and machine translation. *arXiv preprint arXiv :1609.08144*.

Xu, J., Sun, X., Zhang, Z., Zhao, G. & Lin, J. (2019). Understanding and improving layer normalization. *Advances in Neural Information Processing Systems*, 32.

Yahyatabar, M., Jouvet, P. & Cheriet, F. (2020). Dense-Unet : a light model for lung fields segmentation in Chest X-Ray images. *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 1242–1245.

Yang, Z., Dong, L., Du, X., Cheng, H., Cambria, E., Liu, X., Gao, J. & Wei, F. (2024). Language Models as Inductive Reasoners. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pp. 209–225. Repéré à https://aclanthology.org/2024.eacl-long.13.

Young, N., Bao, Q., Bensemann, J. & Witbrock, M. (2022). AbductionRules : Training Transformers to Explain Unexpected Inputs. *Findings of the Association for Computational Linguistics : ACL 2022*, pp. 218–227. doi : 10.18653/v1/2022.findings-acl.19.

Zaglam, N., Jouvet, P., Flechelles, O., Emeriaud, G. & Cheriet, F. (2014). Computer-aided diagnosis system for the acute respiratory distress syndrome from chest radiographs. *Computers in biology and medicine*, 52, 41–48.

Zhang, P., Zeng, G., Wang, T. & Lu, W. (2024). Tinyllama : An open-source small language model. *arXiv preprint arXiv :2401.02385*.

Zhang, X., Ramachandran, D., Tenney, I., Elazar, Y. & Roth, D. (2020). Do language embeddings capture scales ? *arXiv preprint arXiv :2010.05345*.

Zhou, M., Duan, N., Liu, S. & Shum, H.-Y. (2020). Progress in Neural NLP : Modeling, Learning, and Reasoning. *Engineering*, 6(3), 275-290. doi : https://doi.org/10.1016/j.eng.2019.12.014.