Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Comment évaluer un grand modèle de langue dans le domaine médical en français ?

Christophe Servan Cyril Grouin Aurélie Névéol Pierre Zweigenbaum
Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 91400, Orsay
prenom.nom@lisn.fr

RÉSUMÉ

Les récentes avancées en Traitement Automatique des Langues liées aux grands modèles de langue (LLM) auto-régressifs investissent également les domaines spécialisés dont celui de la santé. Cette étude examine les questions qui se posent dans l'évaluation de LLM appliqués au domaine de la santé en se focalisant sur le français. Après un bref tour d'horizon des tâches et des données d'évaluation disponibles pour ce domaine de spécialité, l'article examine le mode d'évaluation des LLM dans des tâches de nature discriminante (détection d'entités nommées, classification de textes) et génératives (résumé de comptes rendus, génération de cas cliniques). L'article n'a pas vocation à rapporter une évaluation concrète, mais à discuter et préparer la méthodologie pour le faire.

ABSTRACT

How to evaluate a large language model in healthcare for the French language ?

Recent advances in Natural Language Processing related to autoregressive large language models (LLMs) are also spreading to specialized fields, including healthcare. This study examines the issues in evaluating LLMs applied to healthcare, focusing on French. After a brief overview of the tasks and evaluation data available for this specialty field, the article examines how LLMs are evaluated in discriminative tasks (named entity detection, text classification) and generative tasks (summarizing reports, generating clinical cases). This paper is not intended to report a concrete evaluation, but to discuss and prepare methods for such an evaluation.

MOTS-CLÉS : Grands Modèles de Langue, Santé, Évaluation, État de l'art.

KEYWORDS: Large Language Models, Health, Evaluation, State-of-the-Art.

1 Introduction

Les grands modèles de langue (LLM) ont montré des capacités qui dépassent celles attendues dans le traitement automatique des langues en offrant des performances notables pour la génération de texte, la traduction, les tâches de question-réponse, etc. (OpenAI *et al.*, 2023; Grattafiori *et al.*, 2024; DeepSeek-AI *et al.*, 2024, 2025; Ben Allal *et al.*, 2025). Dans le domaine médical, et notamment en français, diverses initiatives ont vu le jour pour créer des modèles de langue : des modèles à masquage (Labrak *et al.*, 2023b; Touchent & de la Clergerie, 2024) puis des modèles auto-régressifs (Labrak *et al.*, 2024). D'autres initiatives en cours visent la création d'autres modèles francophones auto-régressifs pour ce domaine, comme dans le cadre du projet PARTAGES.

Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM), pages 51–67.

L'une des questions centrales de ces initiatives concerne la question de l'évaluation de ces modèles et plus particulièrement : comment évaluer un LLM dans le cadre du domaine médical en français ?

Lors de l'évaluation dans le cadre d'un domaine spécialisé comme le domaine médical, on peut bien sûr se demander s'il y a une spécificité à évaluer un LLM francophone adapté au domaine par rapport à un LLM non-spécialisé. Dans notre étude, nous nous focalisons sur le cas spécifique de l'évaluation d'un LLM francophone spécialisé dans le domaine médical. Cependant, certains points abordés sont de fait généraux, et une grande partie de ces considérations s'appliqueront à d'autres situations.

Considérant ce cadre particulier, cela impacte notre étude selon deux axes principaux. D'une part, l'usage du français et d'un domaine spécialisé restreignent de fait les corpus disponibles, ce qui est source de difficultés qu'il faudra résoudre. D'autre part, cibler des tâches particulières, choisies selon les besoins des utilisateurs de ce domaine, amène à se placer dans un contexte concret qui peut se départir des tâches « académiques »[1] et qui peut amener des points particuliers dans les méthodes d'évaluation.

Dans cet article, nous étudions les problématiques d'évaluation rencontrées dans ce contexte (section 2), dont le besoin de jeux de données (section 3). Nous rappelons les méthodes classiques pour l'évaluation de tâches discriminantes comme la détection d'entités et la classification de textes (section 4) et discutons celles qui sont utiles pour l'évaluation de tâches génératives comme la génération de textes ou le résumé (section 5). Nous terminons par l'examen de plusieurs campagnes d'évaluation récentes portant sur des textes médicaux (section 6), puis concluons sur des recommandations générales (section 7).

2 Problématiques

Ces problématiques d'évaluation d'un LLM concernent les méthodes d'évaluation et les données sur lesquelles elles s'appliquent.

2.1 Qu'évalue-t-on ?

L'évaluation peut être « intrinsèque » à la constitution du modèle. Cela concerne ici typiquement la capacité du modèle à prédire le mot suivant (et par extension la séquence de mot suivante), à travers de mesures de perplexité ou entropie sur des textes du domaine visé (Bahl *et al.*, 1983).

L'évaluation est plus souvent « extrinsèque » et examine alors la capacité du modèle à contribuer à la réalisation d'une tâche utile dans le domaine visé. Dans le domaine du Traitement Automatique des Langues, nous pouvons distinguer deux grandes familles de tâches.

D'une part, des tâches de nature « discriminante », qui produisent des données prises dans des catégories prédéfinies. Ainsi, nous nous intéresserons à la détection d'entités, à des fins de désidentification (aussi appelée pseudonymisation) ou de détection de biomarqueurs dans des comptes rendus médicaux ; et à la classification de textes, pour déterminer par exemple la réponse à un traitement, d'où vient l'infection d'un patient, ou pour produire des codes diagnostiques à partir d'un compte rendu médical.

1. En entendant par là les tâches qui reviennent le plus souvent dans les publications.

D'autre part, des tâches « génératives » qui produisent un nouveau texte. Nous nous pencherons sur la génération de textes possédant des caractéristiques données : la création de textes du même genre que des comptes rendus médicaux existants, utile pour étendre la taille du corpus initial dans une démarche d'augmentation de données (Vakili *et al.*, 2025), ou encore la production de descriptions de cas cliniques utiles à l'enseignement. Nous examinerons aussi la génération du paragraphe final d'un compte rendu médical, qui joue le rôle particulier de conclusion ou résumé du compte rendu (Afzal *et al.*, 2020; Singh *et al.*, 2021; Tsai *et al.*, 2022; Chuang *et al.*, 2024).

Les critères d'évaluation extrinsèques portent en général essentiellement sur la performance du système testé : est-ce que les sorties produites sont adéquates pour la tâche visée et avec quelle performance ? Cette performance se mesure automatiquement à l'aide de scores par comparaison à une vérité terrain (les sorties attendues idéalement), ou éventuellement par examen humain des sorties produites. Notons que l'évaluation humaine dans le domaine médical demande souvent une expertise du domaine, coûteuse et difficile à mobiliser. De plus, comme dans toutes les tâches, il est crucial de définir des consignes d'évaluation précises dans un guide d'évaluation. Il est important également de publier ce guide pour faciliter l'interprétation des données et la réplication des études. L'intérêt d'un système peut également s'examiner sous d'autres angles. Par exemple, en génération de textes, la diversité des textes produits peut être un aspect pertinent (Rodriguez-Almeida *et al.*, 2022; Nicholas *et al.*, 2023).

En conditions réelles, la qualité des données initiales peut ne pas être parfaite. Par exemple, dans beaucoup d'hôpitaux, une partie non négligeable des documents textuels est disponible sous forme de documents PDF. Leur conversion en un format textuel analysable est une source potentielle de bruit. La réduction de ce bruit demande de mettre en place des techniques d'analyse précises comme EDS-PDF mentionné par Gérardin *et al.* (2023), qui sépare le corps d'un compte rendu médical de tout ce qui concerne l'en-tête, le pied de page et plus généralement l'entour d'une page, non pertinents d'un point de vue médical.

Au-delà de ces évaluations intrinsèques et extrinsèque, les systèmes entraînés sur des données s'appuient généralement sur les distributions observées dans ces données. Dans les corpus de pré-entraînement des modèles de langue, ces distributions peuvent correspondre à divers biais présents dans nos sociétés. On observe que les grands modèles de langue ont tendance à amplifier ces biais, y compris concernant des catégories médicales (Ducel *et al.*, 2025). Il est donc important de mesurer les biais qu'un modèle de langue est susceptible d'apporter lorsqu'on l'utilise dans une tâche donnée.

Enfin, le gigantisme des LLM est cause d'un coût computationnel important, lui-même à l'origine d'un impact écologique nettement plus grand que les techniques antérieures. Il est donc particulièrement important d'évaluer également cet impact écologique (Mytton, 2021; Morand *et al.*, 2024). En parallèle avec ce coût écologique, le coût financier d'équipement et de fonctionnement est aussi à prendre en compte pour l'organisation qui veut s'équiper d'un service sous-tendu par un tel système. Ce coût est également source d'un manque d'équité dans les possibilités d'usage des grands modèles de langue (Sarker *et al.*, 2024). Dans le contexte qui nous intéresse ici, la question de l'équipement est une nécessité réglementaire pour assurer la protection des données qui ne peuvent être traitées que chez des hébergeurs de données de santé agrées. En pratique, actuellement, seuls les plus grands établissements de santé possèdent un centre de calcul équipé de GPU.

2.2 Comment évaluer ?

Un des écueils provient de la masse de données textuelles sur lesquelles les LLM sont pré-entraînés. De ce fait, les grands modèles de langue peuvent avoir vu le corpus de test de certains jeux de données au cours de leur pré-entraînement. Cela fausse l'évaluation de ces mêmes modèles sur ces jeux de données. Le manque de documentation sur les corpus employés dans ce pré-entraînement rend difficile la détection de la *contamination* de ces modèles par ces jeux de test (Balloccu *et al.*, 2024; Deng *et al.*, 2024; Fu *et al.*, 2024; Palavalli *et al.*, 2024; Xu *et al.*, 2024; Ravaut *et al.*, 2025).

De plus, l'évaluation doit essentiellement porter sur des jeux de données pertinents pour la tâche visée, donc pour la langue et le domaine visé[2]. La versatilité des LLM fait qu'il est difficile d'anticiper l'ensemble des tâches sur lesquelles il pourraient être utilisés, et de créer un jeu d'évaluation pour chaque cas d'usage (Raji *et al.*, 2021).

Comme indiqué plus haut, travailler sur une langue autre que l'anglais et en domaine spécialisé réduit la quantité et la variété des données disponibles (Névéol *et al.*, 2018). Nous faisons le point sur le français médical en section 3.

3 Jeux de données existants

Plus d'une centaine de jeux de données ont été créés pour évaluer des tâches de traitement automatique des langues en anglais dans le domaine biomédical (He *et al.*, 2025) : recherche d'information, reconnaissance d'entités nommées, classification de textes, traduction, etc. Cependant, en français, peu de données sont disponibles. Cette rareté est fortement liée aux contraintes d'utilisation des données médicales en Europe : ces données relèvent de la vie privée et sont donc protégées (RGPD, 2016). On peut citer les corpus QuaeroFrenchMed, E3C, CAS, et FrenchMedMCQA (tableau 1). Ces jeux de données ont été principalement utilisés dans des campagnes d'évaluation comme CLEF (Neveol *et al.*, 2015, 2016) et DEFT (Grabar *et al.*, 2019; Cardon *et al.*, 2020; Grouin *et al.*, 2021; Labrak *et al.*, 2023a; Bazoge *et al.*, 2024a). À noter, certaines campagnes d'évaluation ont pu mettre à disposition d'autres jeux de données temporairement, sous conditions, comme dans CLEF eHealth 2017 (Névéol *et al.*, 2017). Ils ne peuvent de ce fait pas être considérés comme « disponibles ».

QuaeroFrenchMed (Névéol *et al.*, 2014) est composé de deux jeux de données : EMEA et MEDLINE. EMEA est une collection de notices patient concernant des médicaments commercialisés en Europe, tandis que MEDLINE est une collection de titres d'articles scientifiques indexés dans la base de données bibliographique MEDLINE[3]. Ces deux parties sont annotées en dix types d'entités nommées (Anatomy, Chemical and Drugs, Devices, Disorders, Geographic Areas, Living Beings, Objects, Phenomena, Physiology, Procedures).

E3C (Magnini *et al.*, 2023) est un corpus multilingue européen de descriptions de cas de patients publiés dans des revues (*case report*) collectés à partir de multiples sources scientifiques telles que PubMed et SciELO. Ce corpus est annoté en entités de type événement, expression temporelle, acteur, partie du corps, mesure et résultat d'examen, et en relations temporelles et aspectuelles.

2. En complément, si le modèle est multilingue, il pourrait être utile de vérifier sa performance dans d'autres langues sur le même domaine ou dans la même langue en « domaine général », mais la priorité de cette évaluation est nettement plus basse.

3. La base MEDLINE (Medical Literature Analysis and Retrieval System Online), gérée par la Bibliothèque nationale de médecine des États-Unis (United States National Library of Medicine), couvre les sciences de la vie avec un accent sur la biomédecine.

jeu de données	type	tâches
QuaeroFrenchMed	notices, titres d'articles	REN
E3C	cas cliniques	REN, RA, GCC
CAS	cas cliniques	REN, LR, RA, GCC
DIAMED	cas cliniques	classification
FRASIMED	cas cliniques	REN, LR
FrenchMedMCQA	question & réponses	RQ

TABLE 1 – Description des corpus disponibles en français pour la reconnaissance d'entité nommées (REN), le liage référentiel (LR), le résumé automatique (RA), la réponse à des questions (RQ) et la génération de cas cliniques (GCC).

CAS (Grabar *et al.*, 2018; Grouin *et al.*, 2019) est un corpus français de cas cliniques collectés à partir d'articles scientifiques et de documents de formation. Les annotations de ce corpus comprennent des identifiants de concepts de l'UMLS [4] (liage référentiel), des marqueurs de négation et d'hypothèse, et la portée de ces marqueurs.

DIAMED (Bazoge *et al.*, 2024b) est un corpus de cas cliniques collecté après les deux précédents. 69 % de son contenu est dans E3C (Hiebel *et al.*, 2023b). Il fournit des annotations selon les chapitres de la Classification internationale des maladies (CIM-10) (ICD-10, 2011).

FRASIMED (Zaghir *et al.*, 2024a) est un corpus obtenu par traduction automatique des deux corpus espagnols de cas cliniques CANTEMIST (Miranda-Escalada *et al.*, 2020) et DISTEMIST (Miranda-Escalada *et al.*, 2022). Les deux corpus sont annotés par des entités nommées de type maladie, liées plus précisément à des concepts normalisés (liage référentiel). Dans CANTEMIST, les concepts sont des descriptions de morphologie tumorale provenant de la classification internationale des maladies pour l'oncologie (CIM-O-3) dans sa version espagnole (eCIE-O-3.1). Dans DISTEMIST il s'agit de concepts de l'ontologie SNOMED CT.

Enfin, FrenchMedMCQA (Labrak *et al.*, 2022) répertorie 3105 questions à choix multiple issues d'archives d'examens de pharmacie. Chaque question est associée à cinq réponses possibles parmi lesquelles entre une et cinq sont correctes.

Ces jeux de données peuvent être employés dans différentes tâches, par exemple, le corpus CAS est employé dans une tâche de reconnaissance d'entités nommées et dans une tâche de résumé automatique.

4 Tâches discriminantes

Les tâches de nature « discriminante » produisent des données prises dans des catégories prédéfinies, comme la détection d'entités ou la classification de textes.

La question d'évaluation principale soulevée par l'usage d'un modèle génératif dans une tâche discriminante est le mode d'évaluation du texte qu'il génère. En effet, un LLM peut ne pas respecter le format de sortie qui lui est demandé. Cela rend difficile l'exploitation de ses résultats (Beurer-Kellner *et al.*, 2024; Tam *et al.*, 2024). Il se peut alors que le score automatiquement assigné à ce résultat lors

4. *Unified Medical Language System* (Lindberg *et al.*, 1993).

d'une évaluation automatique soit faible, voire nul, alors qu'une interprétation humaine de la sortie du système aurait pu y déceler une réponse correcte. Certains auteurs concluent qu'il faut évaluer les LLM différemment.

Dans notre contexte, l'usage visé est la production de données structurées, pas une lecture humaine. De ce fait, si la forme produite n'est pas celle attendue et que cela empêche d'interpréter automatiquement les résultats fournis, il est pertinent que l'évaluation automatique produise un score bas.

Donner des consignes pour que le LLM produise une sortie structurée (JSON, etc.) est généralement possible, mais cela ne garantit pas que ces consignes soient respectées. La génération contrainte est une autre piste (Beurer-Kellner *et al.*, 2024; Tam *et al.*, 2024). En tout état de cause, ces méthodes relèvent de la conception d'un système et pas de son évaluation. Ce que l'évaluation doit prendre en compte est la sortie du système global formé du modèle, des consignes qui lui sont données, et du post-traitement du texte généré. Les méthodes employées pour réaliser ce système global sont du ressort des personnes qui l'ont conçu, pas de celles qui l'évaluent.

Les métriques automatiques d'évaluation restent celles utilisées dans les autres domaines, à savoir, la précision, le rappel, la F-mesure et l'exactitude.

4.1 Reconnaissance d'entités nommées

La reconnaissance d'entités nommées est employée dans les comptes rendus médicaux pour y détecter des informations médicales. Elle sert aussi à y repérer des informations personnelles à des fins de pseudonymisation.

La tâche de pseudonymisation a pour objectif de rendre les informations personnelles contenues dans un texte non-identifiables. Celles-ci peuvent alors le cas échéant être remplacées par de nouvelles informations personnelles générées artificiellement, appelées substituts plausibles. Ces informations incluent par exemple les nom et prénom du patient, son âge, éventuellement son genre lorsque cela n'est pas nécessaire pour le diagnostic. La motivation de la pseudonymisation est le respect du Règlement Général de Protection des Données (RGPD, 2016) qui demande de protéger les données à caractère privé, en empêchant l'identification des personnes par un traitement et recoupement d'informations.

Peu de travaux ont fait état de performances notables des LLM dans le cadre de la reconnaissance d'entités. Zaghir *et al.* (2024b) ont mené une étude comparative des performances des modèles Transformers appliqués au domaine bio-médical et en français. Ces derniers montrent très clairement que les capacités des LLM génératifs sont largement en deçà des modèles à masquage affinés sur les données. Cependant il est à noter que les conditions d'évaluation ne sont pas comparables. En effet, dans cette étude, les LLM utilisent quelques exemples dans leurs amorces alors que les modèles de langue à masquage (MLM) utilisent l'ensemble des données d'entraînement disponibles.

Naguib *et al.* (2024) prennent en compte cette dimension en comparant ces deux familles de modèles dans le cadre d'un apprentissage à peu de coups (*few-shot learning*) avec le même nombre de coups. Même dans ce cadre, les LMM génératifs sont significativement moins performants que les MLM. Les auteurs recommandent l'utilisation des LLM dans le cadre d'une approche à zéro coup (*zero shot*) ou pour pré-annoter les données dans l'objectif d'utiliser des MLM (*bootstrapping*).

Ces travaux font ressortir un aspect souvent mis en avant dans les LLM : leur capacité à fonctionner avec pas ou peu d'exemples d'entraînement. Ils rappellent qu'en général, un minimum d'exemples

sont de toute façon disponibles et peuvent servir à entraîner un modèle (masqué ou autorégressif) ; et que la mise au point d'une amorce pour un LLM est elle-même consommatrice d'exemples d'entraînement, qu'il faut eux aussi comptabiliser dans le nombre d'exemples annotés nécessaires à son usage.

4.2 Classification de textes cliniques

La classification de textes cliniques est utile notamment pour déterminer les diagnostics principaux concernant un patient lors d'un séjour hospitalier et les exprimer sous forme de classes de la Classification internationale des maladies (CIM-10) (ICD-10, 2011). Cette tâche a fait l'objet de très nombreux travaux, par exemple Song *et al.* (2020). Un autre exemple concerne la détermination de la présence d'obésité (Uzuner, 2009) ou d'autres déterminants sociaux de la santé (Patra *et al.*, 2021).

Au contraire de la reconnaissance d'entités, les LLM offrent des performances comparables voire meilleures que les approches plus classiques fondées sur les modèles masqués (Guevara *et al.*, 2024). Les auteurs se posent cependant la question de la contamination des modèles par les données, déjà évoquée plus haut. Une autre considération importante est le risque de biais qui peut être plus important sur ce type de tâches dans les modèles neuronaux que dans d'autres modèles statistiques (Lwowski & Rios, 2021).

5 Tâches génératives

L'évaluation de tâches génératives s'appuie souvent sur un texte de référence qui permet de calculer des métriques automatiques telles que le score ROUGE (Lin, 2004), ou d'entraîner des modèles évaluant la corrélation entre un texte généré et un texte de référence (Ke *et al.*, 2022). Cependant, dans de nombreux contextes, de telles références ne sont pas disponibles, ce qui complique le travail d'évaluation (Deutsch *et al.*, 2022; Ito *et al.*, 2025).

5.1 Génération de textes cliniques

Le terme *texte clinique* fait généralement référence à des textes décrivant des informations médicales sur des patients, comme le compte rendu d'hospitalisation dans un dossier hospitalier. Il inclut aussi par extension les descriptions de cas de patients publiés dans des revues (*case report*). La génération de textes cliniques est notamment employée pour répondre à une problématique centrale dans le domaine : la difficulté d'accès en recherche aux documents médicaux concernant des patients, qui motive la création de textes fictifs générés automatiquement (Ive *et al.*, 2020; Hiebel *et al.*, 2023a; Boulanger *et al.*, 2024; Hackl *et al.*, 2025; Hahn, 2025; Vakili *et al.*, 2025; Meoni *et al.*, 2025). Ces textes fictifs sont ensuite utilisés à la place de textes réels pour constituer des corpus annotés et entraîner des systèmes. Des travaux portent aussi sur la génération de comptes rendus d'entretiens entre patient et praticien (Eremeev *et al.*, 2023; Ben Abacha *et al.*, 2023; Asada & Miwa, 2023). Parmi les travaux cités ci-dessus, qui portent majoritairement sur l'anglais, Hiebel *et al.* (2023a) et Boulanger *et al.* (2024) génèrent des cas cliniques en français.

Hiebel *et al.* (2023a) évaluent l'utilisabilité des textes générés dans le contexte de la reconnaissance d'entités cliniques. Dans une préoccupation de protection de la confidentialité des données d'entraîne-

ment, ils recensent la présence de n-grammes longs en commun entre le corpus d'entraînement et les textes générés (n-gram overlap). Boulanger *et al.* (2024) réalisent une tâche de génération contrôlée qu'ils évaluent avec la fluidité (adéquation linguistique) du texte généré (perplexité), la diversité des textes générés en utilisant *self-BLEU* (Zhu *et al.*, 2018), l'adéquation aux contraintes imposées, la proximité avec des textes de référence comme avec *corpus-BLEU* (Yu *et al.*, 2017).

D'autres aspects sont également évalués, indépendamment des corpus d'évaluation. Boulanger *et al.* (2024) montrent par exemple que les modèles encodeurs seuls utilisent des ressources de calcul moins importantes que des modèles encodeur-décodeur, ce qui se traduit par un avantage matériel (compatibilité avec des cartes graphiques plus petites) et un impact environnemental moindre. Ducel *et al.* (2025) proposent des métriques pour évaluer la prévalence du genre des personnes décrites dans les cas cliniques générés, ce qui permet une comparaison par rapport à des données de santé publique comme les corpus de cas cliniques réels ou le rapport des sexes documenté pour les pathologies. Enfin, Hiebel *et al.* (2024) proposent un *jeu ayant un but* permettant de collecter des jugements humains sur des textes générés afin d'en évaluer la plausibilité.

5.2 Résumé de textes cliniques

Un compte rendu hospitalier résume les éléments clés du séjour d'un patient dans l'hôpital : motif d'hospitalisation, antécédents, etc., jusqu'au traitement prévu. Sa conclusion est une synthèse de tous ces éléments, qui ont été accumulés au cours du séjour. Générer cette conclusion à partir du reste du compte rendu ou directement à partir de notes cliniques prises au cours du séjour est donc une tâche particulièrement utile. Cette tâche correspond à la génération d'un résumé de notes cliniques (Singh *et al.*, 2021; Chuang *et al.*, 2024). Certains travaux portent également sur le résumé des informations fournies par un patient lors d'une consultation (Tsai *et al.*, 2022). Jain *et al.* (2022) recensent plus largement des situations de résumé de textes cliniques ou d'articles scientifiques dans le domaine médical. La plupart de ces travaux portent sur l'anglais, et nous n'en connaissons pas sur le français.

Les évaluations classiques du résumé automatique comparent le résumé produit à un ou plusieurs résumés de référence. Comme dans la génération de texte, cette comparaison se fait en examinant les mots ou n-grammes communs comme ROUGE (Lin, 2004), BLEU (Papineni *et al.*, 2002) ou à travers une similarité sémantique comme le BERTScore (Zhang *et al.*, 2020). Au-delà de cette similarité à un résumé préexistant considéré comme une vérité terrain, il est important de vérifier que l'information contenue dans le résumé est pertinente. D'une part, est-ce que tous les éléments clés qui devraient y figurer sont présents (par exemple, les diagnostics, traitements, etc.). Une façon d'y parvenir pourrait être de vérifier que certaines entités reconnues dans le texte sont aussi dans le résumé. D'autre part, est-ce que les éléments présents dans le résumé sont bien mentionnés dans le texte à résumer (absence d'« hallucination »). La même méthode de détection d'entités, appliquée dans la direction inverse, constitue une piste de mode de vérification.

6 Travaux apparentés : campagnes d'évaluation récentes

De nombreuses campagnes d'évaluation internationales ont porté sur des tâches de traitement automatique des langues dans le domaine clinique dans d'autres langues que le français, les plus anciennes étant i2b2 (Uzuner *et al.*, 2007) et le CMC Medical NLP Challenge (Pestian *et al.*, 2007). La montée en puissance des grands modèles autorégressifs a-t-elle influencé ces campagnes d'évaluation ? Les

campagnes d'évaluation récentes portant sur des textes cliniques aident à examiner ce point et à compléter les méthodes que nous avons mentionnées plus haut.

CLEF eRisk 2024 a proposé trois tâches concernant la dépression dans des médias sociaux : une tâche de recherche d'information concernant des symptômes de dépression dans des phrases, une tâche de détection de signes d'anorexie et une tâche de détermination de la sévérité de symptômes associés à des troubles du comportement alimentaire (Parapar *et al.*, 2024). Seuls deux des systèmes participants ont utilisé des LLM (GPT-3 et ChatGPT-4).

Dans eRisk 2025, une tâche pilote concerne la détection de signes de dépression d'une personne sur la base d'un dialogue à réaliser avec un agent conversationnel qui joue le rôle de cette personne. L'agent conversationnel est implémenté sous forme d'un LLM auquel a été donnée la consigne appropriée. On peut s'attendre également à ce que les systèmes des participants emploient des LLM. Un système peut renvoyer son verdict de détection au bout d'un nombre de tours de parole aussi grand qu'il le souhaite. L'évaluation portera sur la correction de la détection, mais *en pénalisant les décisions tardives*.

Cette campagne d'évaluation montre un exemple d'évaluation d'un système, qui pourrait être fondé sur un LLM, *à travers un dialogue interactif*, un type de tâche que nous n'avons pas envisagé. L'article portant sur cette campagne d'évaluation n'était pas encore publié au moment de l'écriture de cet article.

ClinIQLink 2025 – LLM Lie Detector Test est l'une des quatre tâches proposées par BioNLP-ST 2025 et concerne spécifiquement les LLM. Les participants doivent soumettre un LLM, qui est évalué par sa capacité à répondre à des questions sur des concepts médicaux fondamentaux correspondant au niveau d'un médecin généraliste. Pour les questions à réponse fermée, l'évaluation est classique. Pour les questions à réponse ouverte, dont la réponse est un texte court, l'évaluation consiste à comparer la réponse fournie à une réponse de référence, par appariement exact s'il réussi, ou sinon *par similarité* (BLEU, ROUGE, METEOR). Dans le cas où les résultats des mesures de similarité seraient incertains ou incohérents, des jugements humains et une similarité fondée sur des représentations vectorielles (de mots, phrases et paragraphes) seront employés. Ces représentations vectorielles sont similaires à celles obtenues par *BERTScore* (Zhang *et al.*, 2020) et *SemScore* (Aynetdinov & Akbik, 2024), ce qui rejoint les directions que nous avons indiquées pour l'évaluation du résumé de textes cliniques en ajoutant une mesure de similarité supplémentaire.

DEFT 2024 portait sur la détection automatique des réponses correctes à des questions à choix multiple provenant d'archives d'examens de pharmacie (Bazoge *et al.*, 2024a) utilisant un nouveau sous-ensemble du corpus FrenchMedMCQA (Labrak *et al.*, 2022). Pour permettre de tester des méthodes de génération assistée par la recherche d'information (RAG), deux collections de textes ont été mises à disposition : Wikipédia et NACHOS. Plusieurs des participants ont utilisé ce type de méthode et ont obtenu les meilleurs résultats. Pour examiner l'influence de la taille des grands modèles de langue sur les résultats, deux évaluations ont été organisées : l'une réservée aux systèmes de moins de trois milliards de paramètres, l'autre non limitée, mais aucun participant n'a employé de système de la seconde catégorie.

Cette campagne d'évaluation montre que pour évaluer des systèmes utilisant la recherche d'information (RAG), il faut *fournir les collections de textes dans lesquels doit se faire cette recherche d'information*, ce que nous n'avons pas discuté plus haut.

EvalLLM 2024 et 2025 sont des campagnes d'évaluation des LLM. La seconde inclut une tâche d'extraction d'information à des fins de veille sanitaire dans des documents journalistiques. Ces campagnes ont pris pour principe de *donner peu d'exemples annotés*. Elles ont notamment demandé aux participants d'*évaluer l'empreinte carbone des LLM*, ce qui rejoint nos recommandations, et de décrire les amorces (*prompts*) utilisées. L'édition 2025 inclut également une tâche d'affinage de LLM pour un domaine spécialisé dont l'évaluation inclut *un test de non-régression sur le domaine général* que nous évoquions en note 2.

7 Conclusion

Nous avons discuté des modes d'évaluation des grands modèles de langue (LLM) pour le domaine de la santé en français pour plusieurs tâches pertinentes dans ce domaine.

Nous avons indiqué en préalable que comme pour l'évaluation d'autres méthodes, le manque de données de référence en français et les restrictions d'accès aux données médicales rendent cette évaluation d'autant plus difficile.

Concernant l'évaluation de tâches discriminantes, nous considérons que l'évaluation de systèmes à base de LLM ne doit pas être différente de celle d'autres méthodes : c'est le système global que l'on évalue, indépendamment de son mode de fonctionnement interne. L'utilisation de méthodes pour assurer que le système génère une sortie respectant le format attendu, comme la génération contrainte ou la production de JSON ou XML, est du ressort des auteurs du système.

Dans le cadre de l'évaluation de tâches génératives, la notion de diversité dans la génération de cas cliniques n'est que peu abordée. Dans la génération de résumé, outre les mesure de similarité de surface et celles fondées sur des représentations vectorielles, la complétude et la véracité du contenu du résumé sont particulièrement importants dans un domaine sensible comme la santé.

Les systèmes à apprentissage supervisé sont sujets aux biais des corpus sur lesquels ils sont entraînés. L'exacerbation des biais dans les grands modèles de langue amène à porter une attention particulière à l'évaluation de ces biais, dans les tâches discriminantes comme génératives.

Enfin, du fait du coût computationnel nettement plus élevé de l'usage de très grands modèles de langue, il faut ajouter aux évaluations de qualité l'estimation du coût computationnel du système et de son impact écologique. Cela devrait aider les utilisateurs de systèmes à choisir le meilleur compromis entre la qualité des sorties produites et les coûts induits.

Remerciements

Ce travail a bénéficié d'un soutien dans le cadre de l'appel BPI France FRANCE 2030 « Communs numériques pour l'intelligence artificielle générative » (projet PARTAGES, contrat d'aide DOS0245197).

Nous remercions les relecteurs pour leurs remarques constructives que nous nous sommes efforcés de prendre toutes en compte.

Références

AFZAL M., ALAM F., MALIK K. M. & MALIK G. M. (2020). Clinical context–aware biomedical text summarization using deep neural network : model development and validation. *Journal of medical Internet research*, **22**(10), e19810.

ASADA M. & MIWA M. (2023). BioNART : A biomedical non-autoregressive transformer for natural language generation. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, p. 369–376.

AYNETDINOV A. & AKBIK A. (2024). SemScore : Automated evaluation of instruction-tuned LLMs based on semantic textual similarity.

BAHL L. R., JELINEK F. & MERCER R. L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, **PAMI-5**(2), 179–190. DOI : 10.1109/TPAMI.1983.4767370.

BALLOCCU S., SCHMIDTOVÁ P., LANGO M. & DUSEK O. (2024). Leak, cheat, repeat : Data contamination and evaluation malpractices in closed-source LLMs. In Y. GRAHAM & M. PURVER, Éds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 67–93, St. Julian's, Malta : Association for Computational Linguistics.

BAZOGE A., LABRAK Y., DUFOUR R., FAVRE B. & ROUVIER M. (2024a). Tâches et systèmes de sélection automatique de réponses à des QCM dans le domaine médical : Présentation de la campagne DEFT 2024. In R. DUFOUR, B. FAVRE, M. ROUVIER, A. BAZOGE & Y. LABRAK, Éds., *Actes du Défi Fouille de Textes@TALN 2024*, p. 1–10, Toulouse, France : ATALA and AFPC.

BAZOGE A., MORIN E., DAILLE B. & GOURRAUD P. (2024b). Adaptation of biomedical and clinical pretrained models to French long documents : A comparative study. *CoRR*, **abs/2402.16689**. DOI : 10.48550/ARXIV.2402.16689.

BEN ABACHA A., YIM W.-W., FAN Y. & LIN T. (2023). An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, p. 2291–2302.

BEN ALLAL L., LOZHKOV A., BAKOUCH E., MARTÍN BLÁZQUEZ G., PENEDO G., TUNSTALL L., MARAFIOTI A., KYDLÍCEK H., PIQUERES LAJARÍN A., SRIVASTAV V., LOCHNER J., FAHLGREN C., NGUYEN X., FOURRIER C., BURTENSHAW B., LARCHER H., ZHAO H., ZAKKA C., MORLON M., RAFFEL C., VON WERRA L. & WOLF T. (2025). SmolLM2 : When Smol goes big – data-centric training of a small language model. DOI : 10.48550/ARXIV.2502.02737.

BEURER-KELLNER L., FISCHER M. & VECHEV M. (2024). Guiding LLMs the right way : fast, non-invasive constrained generation. In *Proceedings of the 41st International Conference on Machine Learning*, p. 3658–3673.

BOULANGER H., HIEBEL N., FERRET O., FORT K. & NÉVÉOL A. (2024). Using structured health information for controlled generation of clinical cases in French. In T. NAUMANN, A. BEN ABACHA, S. BETHARD, K. ROBERTS & D. BITTERMAN, Éds., *Proceedings of the 6th Clinical Natural Language Processing Workshop*, p. 172–184, Mexico City, Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2024.clinicalnlp-1.14.

CARDON R., GRABAR N., GROUIN C. & HAMON T. (2020). Présentation de la campagne d'évaluation DEFT 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques. In *Actes de l'atelier Défi Fouille de Textes@JEP-TALN 2020 similarité sémantique et extraction d'information fine. Atelier DÉfi Fouille de Textes*, p. 1–13, Nancy, France : Association pour le Traitement Automatique des Langues.

CHUANG Y.-N., TANG R., JIANG X. & HU X. (2024). SPeC : A soft prompt-based calibration on performance variability of large language model in clinical notes summarization. *Journal of Biomedical Informatics*, **151**, 104606. DOI : https ://doi.org/10.1016/j.jbi.2024.104606.

DEEPSEEK-AI *et al.* (2024). DeepSeek-V3 technical report.

DEEPSEEK-AI *et al.* (2025). DeepSeek-R1 : Incentivizing reasoning capability in LLMs via reinforcement learning.

DENG C., ZHAO Y., HENG Y., LI Y., CAO J., TANG X. & COHAN A. (2024). Unveiling the spectrum of data contamination in language model : A survey from detection to remediation. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Findings of the Association for Computational Linguistics : ACL 2024*, p. 16078–16092, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.findings-acl.951.

DEUTSCH D., DROR R. & ROTH D. (2022). On the limitations of reference-free evaluations of generated text. In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 10960–10977, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.753.

DUCEL F., HIEBEL N., FERRET O., FORT K. & NÉVÉOL A. (2025). "Women do not have heart attacks !" gender biases in automatically generated clinical cases in French. In L. CHIRUZZO, A. RITTER & L. WANG, Éds., *Findings of the Association for Computational Linguistics : NAACL 2025*, p. 7145–7159, Albuquerque, New Mexico : Association for Computational Linguistics.

EREMEEV M., VALMIANSKI I., AMATRIAIN X. & KANNAN A. (2023). Injecting knowledge into language generation : a case study in auto-charting after-visit care instructions from medical dialogue. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2373–2390, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.133.

FU Y., UZUNER Ö., YETISGEN M. & XIA F. (2024). Does data contamination detection work (well) for LLMs ? A survey and evaluation on detection assumptions. DOI : 10.48550/ARXIV.2410.18966.

GÉRARDIN C., WAJSBÜRT P., DURA B., CALLIGER A., MOUCHER A., TANNIER X. & BEY R. (2023). Detecting automatically the layout of clinical documents to enhance the performances of downstream natural language processing. *arXiv preprint arXiv :2305.13817*.

GRABAR N., CLAVEAU V. & DALLOUX C. (2018). CAS : French corpus with clinical cases. In A. LAVELLI, A.-L. MINARD & F. RINALDI, Éds., *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, p. 122–128, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-5614.

GRABAR N., GROUIN C., HAMON T. & CLAVEAU V. (2019). Recherche et extraction d'information dans des cas cliniques. présentation de la campagne d'évaluation DEFT 2019. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Défi Fouille de Textes (atelier TALN-RECITAL)*, p. 7–16, Toulouse, France : Association pour le Traitement Automatique des Langues. Information Retrieval and Information Extraction from Clinical Cases.

GRATTAFIORI A. *et al.* (2024). The Llama 3 Herd of Models. DOI : 10.48550/arXiv.2407.21783.

GROUIN C., GRABAR N., CLAVEAU V. & HAMON T. (2019). Clinical case reports for NLP. In D. DEMNER-FUSHMAN, K. B. COHEN, S. ANANIADOU & J. TSUJII, Éds., *Proceedings of the 18th BioNLP Workshop and Shared Task*, p. 273–282, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/W19-5029.

GROUIN C., GRABAR N. & ILLOUZ G. (2021). Classification de cas cliniques et évaluation automatique de réponses d'étudiants : présentation de la campagne DEFT 2021 (clinical cases clas-

sification and automatic evaluation of student answers : Presentation of the DEFT 2021 challenge). In C. GROUIN, N. GRABAR & G. ILLOUZ, Éds., *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier DÉfi Fouille de Textes (DEFT)*, p. 1–13, Lille, France : ATALA.

GUEVARA M., CHEN S., THOMAS S., CHAUNZWA T. L., FRANCO I., KANN B. H., MONINGI S., QIAN J. M., GOLDSTEIN M., HARPER S. *et al.* (2024). Large language models to identify social determinants of health in electronic health records. *NPJ digital medicine*, **7**(1), 6.

HACKL W. O., NEURURER S. B., RICHTER S., TAHA H., MUEHLBOECK H., HICKMANN C., GSCHEIDLINGER P., DANLER M., SCHWEITZER M., UEBEREGGER M. & PFEIFER B. (2025). Development of a synthetic oncology pathology dataset for large language model evaluation in medical text classification. *Stud Health Technol Inform*, **324**, 215–220.

HAHN U. (2025). Clinical document corpora-real ones, translated and synthetic substitutes, and assorted domain proxies : a survey of diversity in corpus design, with focus on german text data. *JAMIA Open*, **8**(3), ooaf024. DOI : 10.1093/jamiaopen/ooaf024.

HE Y., HUANG F., JIANG X., NIE Y., WANG M., WANG J. & CHEN H. (2025). Foundation model for advancing healthcare : Challenges, opportunities and future directions. *IEEE Reviews in Biomedical Engineering*, **18**, 172–191. DOI : 10.1109/RBME.2024.3496744.

HIEBEL N., FERRET O., FORT K. & NÉVÉOL A. (2023a). Can synthetic text help clinical named entity recognition ? a study of electronic health records in French. In A. VLACHOS & I. AUGENSTEIN, Éds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, p. 2320–2338, Dubrovnik, Croatia : Association for Computational Linguistics. DOI : 10.18653/v1/2023.eacl-main.170.

HIEBEL N., FERRET O., FORT K. & NÉVÉOL A. (2023b). Similarité surfacique et similarité sémantique dans des cas cliniques générés. In *Journée d'étude sur la Similarité entre Patients, ATALA, SimPa 2023*.

HIEBEL N., REMY B., GUILLAUME B., FERRET O., NÉVÉOL A. & FORT K. (2024). Hostomytho : A GWAP for synthetic clinical texts evaluation and annotation. In C. MADGE, J. CHAMBERLAIN, K. FORT, U. KRUSCHWITZ & S. LUKIN, Éds., *Proceedings of the 10th Workshop on Games and Natural Language Processing @ LREC-COLING 2024*, p. 14–20, Torino, Italia : ELRA and ICCL.

ICD-10 (2011). *ICD-10. International Statistical Classification of Diseases and Related Health Problems. 10th Revision. Volume 2. Instruction manual*. World Health Organization.

ITO T., VAN DEEMTER K. & SUZUKI J. (2025). Reference-free evaluation metrics for text generation : A survey. DOI : 10.48550/ARXIV.2501.12011.

IVE J., VIANI N., KAM J., YIN L., VERMA S., PUNTIS S., CARDINAL R. N., ROBERTS A., STEWART R. & VELUPILLAI S. (2020). Generation and evaluation of artificial mental health records for natural language processing. *NPJ digital medicine*, **3**(1), 69.

JAIN R., JANGRA A., SAHA S. & JATOWT A. (2022). A survey on medical document summarization. *CoRR*, **abs/2212.01669**. DOI : 10.48550/ARXIV.2212.01669.

KE P., ZHOU H., LIN Y., LI P., ZHOU J., ZHU X. & HUANG M. (2022). CTRLEval : An unsupervised reference-free metric for evaluating controlled text generation. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2306–2319, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.164.

LABRAK Y., BAZOGE A., DAILLE B., DUFOUR R., MORIN E. & ROUVIER M. (2023a). Tâches et systèmes de détection automatique des réponses correctes dans des qcms liés au domaine médical : Présentation de la campagne DEFT 2023. In *Actes de CORIA-TALN 2023. Actes du Défi Fouille*

de Textes@TALN2023, p. 57–67, Paris, France : Association pour le Traitement Automatique des Langues.

LABRAK Y., BAZOGE A., DUFOUR R., DAILLE B., GOURRAUD P.-A., MORIN E. & ROUVIER M. (2022). FrenchMedMCQA : A French multiple-choice question answering dataset for medical domain. In A. LAVELLI, E. HOLDERNESS, A. JIMENO YEPES, A.-L. MINARD, J. PUSTEJOVSKY & F. RINALDI, Éds., *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, p. 41–46, Abu Dhabi, United Arab Emirates (Hybrid) : Association for Computational Linguistics. DOI : 10.18653/v1/2022.louhi-1.5.

LABRAK Y., BAZOGE A., DUFOUR R., ROUVIER M., MORIN E., DAILLE B. & GOURRAUD P.-A. (2023b). DrBERT : A robust pre-trained model in French for biomedical and clinical domains. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 16207–16221, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.896.

LABRAK Y., BAZOGE A., MORIN E., GOURRAUD P.-A., ROUVIER M. & DUFOUR R. (2024). BioMistral : A collection of open-source pretrained large language models for medical domains. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Findings of the Association for Computational Linguistics : ACL 2024*, p. 5848–5864, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.findings-acl.348.

LIN C.-Y. (2004). ROUGE : A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, p. 74–81, Barcelona, Spain : Association for Computational Linguistics.

LINDBERG D. A., HUMPHREYS B. L. & MCCRAY A. T. (1993). The unified medical language system. *Yearbook of medical informatics*, **2**(01), 41–51.

LWOWSKI B. & RIOS A. (2021). The risk of racial bias while tracking influenza-related content on social media using machine learning. *Journal of the American Medical Informatics Association*, **28**(4), 839–849.

MAGNINI B., ALTUNA B., LAVELLI A., SPERANZA M. & ZANOLI R. (2023). European clinical case corpus. In *European Language Grid*, p. 143–148, Switzerland : Springer, Cham.

MEONI S., DE LA CLERGERIE É. & RYFFEL T. (2025). Synthetic documents for medical tasks : Bridging privacy with knowledge injection and reward mechanism. In S. ANANIADOU, D. DEMNER-FUSHMAN, D. GUPTA & P. THOMPSON, Éds., *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, p. 12–25, Albuquerque, New Mexico : Association for Computational Linguistics.

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

MIRANDA-ESCALADA A., GASCÓ L., LIMA-LÓPEZ S., FARRÉ-MADUELL E., ESTRADA D., NENTIDIS A., KRITHARA A., KATSIMPRAS G., PALIOURAS G. & KRALLINGER M. (2022). Overview of DisTEMIST at BioASQ : Automatic detection and normalization of diseases from clinical texts : results, methods, evaluation and multilingual resources. In G. FAGGIOLI, N. FERRO, A. HANBURY & M. POTTHAST, Éds., *Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (CLEF)*, volume 3180 de CEUR Workshop Proceedings, p. 179–203, Aachen.

MORAND C., LIGOZAT A.-L. & NÉVÉOL A. (2024). How green can AI be ? a study of trends in machine learning environmental impacts. Working paper or preprint, HAL : hal-04839926.

MYTTON D. (2021). Data centre water consumption. *npj Clean Water*, **4**(1), 11.

NAGUIB M., TANNIER X. & NÉVÉOL A. (2024). Few-shot clinical entity recognition in English, French and Spanish : masked language models outperform generative model prompting. In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2024*, p. 6829–6852, Miami, Florida, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2024.findings-emnlp.400.

NEVEOL A., GOEURIOT L., KELLY L., COHEN K., GROUIN C., HAMON T., LAVERGNE T., REY G., ROBERT A., TANNIER X. & ZWEIGENBAUM P. (2016). Clinical information extraction at the CLEF eHealth evaluation lab 2016. In *CLEF 2016 (online working notes).*, Evora, Portugal. HAL : hal-01922402.

NÉVÉOL A., GROUIN C., LEIXA J., ROSSET S. & ZWEIGENBAUM P. (2014). The QUAERO French medical corpus : A ressource for medical entity recognition and normalization. In *Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing @LREC2014*, p. 24–30, Reykjavik, Iceland.

NEVEOL A., GROUIN C., TANNIER X., HAMON T., KELLY L., GOEURIOT L. & ZWEIGENBAUM P. (2015). CLEF eHealth Evaluation Lab 2015 Task 1b : Clinical named entity recognition. In *CLEF 2015 Working notes*, Toulouse, France. HAL : hal-01922444.

NICHOLAS I., KUO H., GARCIA F., SÖNNERBORG A., BÖHM M., KAISER R., ZAZZI M., POLIZZOTTO M., JORM L., BARBIERI S. *et al.* (2023). Generating synthetic clinical data that capture class imbalanced distributions with generative adversarial networks : example using antiretroviral therapy for hiv. *Journal of Biomedical Informatics*, **144**, 104436.

NÉVÉOL A., ANDERSON R. N., COHEN K. B., GROUIN C., LAVERGNE T., REY G., ROBERT A., RONDET C. & ZWEIGENBAUM P. (2017). CLEF eHealth 2017 multilingual information extraction task overview : ICD10 coding of death certificates in English and French. In *CLEF 2017 Evaluation Labs and Workshop : Online Working Notes* : CEUR-WS.

NÉVÉOL A., DALIANIS H., VELUPILLAI S., SAVOVA G. & ZWEIGENBAUM P. (2018). Clinical natural language processing in languages other than English : opportunities and challenges. *Journal of biomedical semantics*, **9**(1), 12. DOI : 10.1186/s13326-018-0179-8.

OPENAI *et al.* (2023). GPT-4 Technical Report.

PALAVALLI M., BERTSCH A. & GORMLEY M. R. (2024). A taxonomy for data contamination in large language models. DOI : 10.48550/arXiv.2407.08716.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). BLEU : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, p. 311–318, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073135.

PARAPAR J., MARTÍN-RODILLA P., LOSADA D. E. & CRESTANI F. (2024). Overview of eRisk 2024 : Early risk prediction on the Internet. In L. GOEURIOT, P. MULHEM, G. QUÉNOT, D. SCHWAB, G. M. DI NUNZIO, L. SOULIER, P. GALUŠČÁKOVÁ, A. GARCÍA SECO DE HERRERA, G. FAGGIOLI & N. FERRO, Éds., *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, p. 73–92, Cham : Springer Nature Switzerland.

PATRA B. G., SHARMA M. M., VEKARIA V., ADEKKANATTU P., PATTERSON O. V., GLICKSBERG B., LEPOW L. A., RYU E., BIERNACKA J. M., FURMANCHUK A. *et al.* (2021). Extracting social determinants of health from electronic health records using natural language processing : a systematic review. *Journal of the American Medical Informatics Association*, **28**(12), 2716–2727.

PESTIAN J. P., BREW C., MATYKIEWICZ P., HOVERMALE D., JOHNSON N., COHEN K. B. & DUCH W. (2007). A shared task involving multi-label classification of clinical free text. In K. B. COHEN, D. DEMNER-FUSHMAN, C. FRIEDMAN, L. HIRSCHMAN & J. PESTIAN, Éds., *Biological,*

translational, and clinical language processing, p. 97–104, Prague, Czech Republic : Association for Computational Linguistics.

RAJI D., DENTON E., BENDER E. M., HANNA A. & PAULLADA A. (2021). AI and the everything in the whole wide world benchmark. In J. VANSCHOREN & S. YEUNG, Éds., *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

RAVAUT M., DING B., JIAO F., CHEN H., LI X., ZHAO R., QIN C., XIONG C. & JOTY S. (2025). A comprehensive survey of contamination detection methods in large language models. DOI : 10.48550/arXiv.2404.00699.

RGPD (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (text with EEA relevance).

RODRIGUEZ-ALMEIDA A. J., FABELO H., ORTEGA S., DENIZ A., BALEA-FERNANDEZ F. J., QUEVEDO E., SOGUERO-RUIZ C., WÄGNER A. M. & CALLICO G. M. (2022). Synthetic patient data generation and evaluation in disease prediction using small and imbalanced datasets. *IEEE journal of biomedical and health informatics*, **27**(6), 2670–2680.

SARKER A., ZHANG R., WANG Y., XIAO Y., DAS S., SCHUTTE D., ONIANI D., XIE Q. & XU H. (2024). Natural language processing for digital health in the era of large language models. *Yearb Med Inform*, **33**(1), 229–240. DOI : 10.1055/s-0044-1800750.

SINGH S., KARIMI S., HO-SHON K. & HAMEY L. (2021). Show, tell and summarise : learning to generate and summarise radiology findings from medical images. *Neural Computing and Applications*, **33**(13), 7441–7465.

SONG C., ZHANG S., SADOUGHI N., XIE P. & XING E. (2020). Generalized zero-shot text classification for ICD coding. In C. BESSIERE, Éd., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, p. 4018–4024 : International Joint Conferences on Artificial Intelligence Organization. Main track, DOI : 10.24963/ijcai.2020/556.

TAM Z. R., WU C.-K., TSAI Y.-L., LIN C.-Y., LEE H.-Y. & CHEN Y.-N. (2024). Let me speak freely ? a study on the impact of format restrictions on large language model performance. In F. DERNONCOURT, D. PREOŢIUC-PIETRO & A. SHIMORINA, Éds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing : Industry Track*, p. 1218–1236, Miami, Florida, US : Association for Computational Linguistics. DOI : 10.18653/v1/2024.emnlp-industry.91.

TOUCHENT R. & DE LA CLERGERIE É. (2024). CamemBERT-bio : Leveraging continual pre-training for cost-effective models on French biomedical data. In N. CALZOLARI, M.-Y. KAN, V. HOSTE, A. LENCI, S. SAKTI & N. XUE, Éds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 2692–2701, Torino, Italia : ELRA and ICCL.

TSAI H.-Y., HUANG H.-H., CHANG C.-J., TSAI J.-S. & CHEN H.-H. (2022). Patient history summarization on outpatient conversation. In *2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, p. 364–370 : IEEE.

UZUNER O. (2009). Recognizing obesity and comorbidities in sparse data. *J Am Med Inform Assoc*, **16**(4), 561–570.

UZUNER O., LUO Y. & SZOLOVITS P. (2007). Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, **14**, 550–563. DOI : doi :10.1197/jamia.M2444.

VAKILI T., HENRIKSSON A. & DALIANIS H. (2025). Data-constrained synthesis of training data for de-identification. DOI : 10.48550/arXiv.2502.14677.

XU C., GUAN S., GREENE D. & KECHADI M. T. (2024). Benchmark data contamination of large language models : A survey. *CoRR*, **abs/2406.04244**. DOI : 10.48550/ARXIV.2406.04244.

YU L., ZHANG W., WANG J. & YU Y. (2017). SeqGAN : sequence generative adversarial nets with policy gradient. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, p. 2852–2858 : AAAI Press.

ZAGHIR J., BJELOGRLIC M., GOLDMAN J.-P., AANANOU S., GAUDET-BLAVIGNAC C. & LOVIS C. (2024a). FRASIMED : A clinical French annotated resource produced through crosslingual BERT-based annotation projection. In N. CALZOLARI, M.-Y. KAN, V. HOSTE, A. LENCI, S. SAKTI & N. XUE, Éds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 7450–7460, Torino, Italia : ELRA and ICCL.

ZAGHIR J., BJELOGRLIC M., GOLDMAN J.-P., BENSAHLA A., ZHENG Y. & LOVIS C. (2024b). Beyond tokens : Fair evaluation of French large language models for clinical named entity recognition. In *Digital Health and Informatics Innovations for Sustainable Health Care Systems*, p. 666–670. IOS Press.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2020). BERTScore : Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020* : OpenReview.net.

ZHU Y., LU S., ZHENG L., GUO J., ZHANG W., WANG J. & YU Y. (2018). Texygen : A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, p. 1097–1100.