

UNIVERSITE DE LILLE
FACULTE DE MEDECINE HENRI WAREMBOURG
Année 2024

THESE POUR LE DIPLOME D'ETAT
DE DOCTEUR EN MEDECINE

**Performance d'un grand modèle de langage (LLM) open-source
pour l'extraction d'information depuis des comptes-rendus
d'IRM cérébrale**

Présentée et soutenue publiquement le 17 juin 2024
à 16 : 00 au pôle formation

Par Bastien LE GUELLEC

JURY

Président :

Monsieur le Professeur Jean-Pierre PRUVO

Assesseurs :

Monsieur le Professeur Philippe AMOUYEL

Monsieur le Docteur Aghiles HAMROUN

Directeur de thèse :

Monsieur le Docteur Grégory KUHCINSKI

Avertissement

La Faculté n'entend donner aucune approbation aux opinions émises dans les thèses : celles-ci sont propres à leurs auteurs.

Sigles

EDS	Entrepôt de Données de Santé
GPT	<i>Generative Pre-trained Transformer</i>
HAS	Haute Autorité de Santé
HSA	Hémorragie Sous-Arachnoïdienne
IA	Intelligence Artificielle
ICD	<i>International Classification of Diseases, 10th revision</i>
IRM	Imagerie par Résonance Magnétique
LLM	<i>Large Language Model</i>
MIMIC	<i>Multiparameter Intelligent Monitoring in Intensive Care</i>
SFEMC	Société Française d'Étude des Migraines et des Céphalées
SFN	Société Française de Neurologie
TVC	Thrombose Veineuse Cérébrale

Sommaire

Avertissement	1
Sommaire	3
Introduction	4
L'IRM pour céphalées aux urgences	4
Extraction automatique d'information depuis des comptes-rendus	7
Les modèles de fondation / <i>Large Language Models</i> (LLMs).....	13
Article en Anglais	30
Conclusion	62
Références	72

Introduction

L'IRM pour céphalées aux urgences

Les céphalées sont responsables de 2 à 4% des passages aux urgences, soit environ 450.000 consultations par an en France (1,2). Près de 95% des patients retournent au domicile avec un diagnostic de céphalées dites « primaires », sans lésion sous-jacente (1,2). Dans ces cas, la maladie est la douleur, et le traitement consiste en une antalgie adaptée, les migraines étant une cause significative de handicap dans les activités de la vie quotidienne (3). Plus rarement, la douleur peut être révélatrice d'une pathologie qu'il faut identifier et traiter en urgence. De nombreuses maladies peuvent ainsi se manifester par des maux de tête, couvrant un large spectre de gravité et de manifestations cliniques : causes infectieuses (méningo-encéphalite, sinusite), vasculaires (dissection, hémorragie sous-arachnoïdienne (HSA), thrombose veineuse cérébrale (TVC)), tumorales, inflammatoires, etc. Pour l'urgentiste et le neurologue, il est essentiel de repérer ces céphalées « secondaires », dont la prise en charge peut éventuellement revêtir un caractère d'urgence vitale.

Cependant, cette identification est parfois difficile tant les signes d'alertes peuvent être subtils (4,5), la majorité des signes cliniques et des examens biologiques manquant individuellement de sensibilité et/ou de spécificité (4). Ainsi, le recours à la neuro-imagerie est parfois nécessaire pour éliminer une céphalée secondaire. En France, les recommandations de la Société Française d'Étude des Migraines et des Céphalées (SFEMC) et de la Société Française de Neurologie (SFN) de 2018 (6) indiquent qu'il convient de faire appel à un examen de neuro-imagerie avec injection rapidement au cours de la prise en charge en cas de signe d'alerte. Ces signes d'alerte sont variés : céphalées

brutales ou s'aggravant progressivement, ou au caractère inhabituel, même chez un patient migraineux connu.

L'utilisation de la neuro-imagerie pour les céphalées aux urgences est ainsi en augmentation en Europe et aux États-Unis (7–9). Aux États-Unis, la proportion d'adultes céphalalgiques ayant bénéficié d'un scanner ou d'une IRM lors de leur consultation aux urgences a doublé entre 2006 et 2014, passant de 19 à 37 % (8). À Lille en 2017, d'après les travaux de thèse de Reta Fakih, 16 % des adultes consultant pour céphalées aux urgences ont bénéficié d'un scanner et 37 % d'une IRM (10).

L'examen de référence des céphalées aux urgences est l'IRM, car elle possède sur le scanner des avantages décisifs (11). Premièrement, elle n'utilise pas de rayonnements ionisants et peut donc être utilisée sans contrainte de radioprotection, notamment chez les patients jeunes qui constituent la majorité des patients consultant pour céphalées (11). Deuxièmement, elle est plus sensible pour détecter certains diagnostics : dissection des vaisseaux du cou, thrombose veineuse cérébrale corticale, méningite, etc. (11) Enfin, le développement des séquences par imagerie de perfusion sans injection de produit de contraste (*Arterial Spin Labelling*) permet désormais, au-delà d'écarter des diagnostics graves, de supporter des diagnostics positifs, comme celui de migraine, intégrant le patient dans un parcours de soin adapté (12).

Cependant, l'IRM est une modalité d'imagerie coûteuse : la réalisation d'une IRM cérébrale coûte environ 200 à 300 euros (13,14), contre environ 100 euros pour un scanner cérébral (13,15). L'IRM encéphalique est aussi un examen long : la réalisation de l'ensemble des séquences nécessaires à écarter tous les diagnostics graves peut dépasser les 30 minutes. Enfin, son accès est difficile : seuls de rares centres disposent d'une IRM dédiée aux urgences. Ainsi, le parcours de soin peut être ralenti par la prescription de cet examen et l'attente de ses résultats (16). Ces contraintes ne peuvent donc pas supporter l'augmentation constante du nombre d'examens pour cette indication.

Afin de limiter les coûts de la surutilisation de l'imagerie neurologique de pointe aux urgences, des outils ont été développés pour aider le clinicien et le radiologue à prédire l'apport de la neuro-imagerie au cas par cas [6,17–20]. Cependant, la majorité de ceux-ci ont été développés pour le scanner, sans tenir compte des spécificités liées à l'IRM. Par exemple, la règle d'Ottawa, reprise par les recommandations françaises (6), ne concerne que la suspicion d'hémorragie sous-arachnoïdienne en scanner, et correspond à seulement 6 % des céphalées d'origine secondaire dans l'étude menée à Lille par R. Fakhri [21]. Il est donc crucial de développer des outils prédictifs spécifiques à l'IRM afin de mieux guider la prise en charge des patients aux urgences. Deux études européennes récentes ont tenté de développer de tels scores. Budweg et collègues ont repris les IRM de 82 sujets ayant passé une IRM en urgence pour céphalées [17]. Deux neuroradiologues expérimentés ont relu les images et ont identifié 15 patients (18%) pour lesquels une anomalie à l'origine des symptômes était présente. Leur modèle de prédiction de la positivité de l'examen prenait en compte de nombreuses caractéristiques du patient et de l'épisode (durée, localisation, caractéristiques, intensité, etc.), mais le faible nombre de céphalées secondaires et le caractère monocentrique de l'étude n'ont pas permis d'atteindre la puissance nécessaire au développement d'un score validé. Happonen et collègues ont eux aussi tenté de développer un score prédictif [6]. La revue rétrospective des comptes-rendus d'imagerie a permis d'identifier 136 sujets pour lesquels une anomalie significative expliquant les symptômes avait été décrite, parmi les 696 inclus, soit un taux de positivité des examens de 20 %. Leur modèle prenait en compte seulement des caractéristiques générales sur le patient et sur ses symptômes, sans préciser la nature de l'épisode aigu (pas d'information sur la durée de l'épisode, son intensité, etc.). Les performances du score clinique développé n'étaient que moyennes : sensibilité pour détecter un patient avec une céphalée secondaire à seulement 46%. Si les recommandations de 2018 rappellent la nécessité d'identifier les patients susceptibles de bénéficier d'un bilan en urgence, les urgentistes et les neurologues ne

disposent donc pas d'outil de prédiction fiable de la pertinence des examens d'imagerie prescrits, conduisant à une surutilisation de l'imagerie lourde pour cette indication [8,22]. Il conviendrait donc de développer des outils plus fiables, multimodaux, intégrant des caractéristiques détaillées du patient et de l'épisode.

À Lille, une IRM dédiée aux urgences est installée depuis 2009. Elle assure la permanence de l'expertise neuroradiologique 7 jours sur 7 et 24h sur 24. D'après les travaux de thèse de R. Fakih, 36 % des IRM pour céphalées aux urgences réalisées au CHU de Lille comprenaient au moins une anomalie (10). Cependant, il n'est pas discuté dans cette étude du caractère causal ou non de l'anomalie décrite par le radiologue. Il n'est en effet pas rare que le radiologue décrive dans sa conclusion des anomalies incidentelles, qui ne sont probablement pas la cause des céphalées. De plus, le taux d'examen injecté n'est pas indiqué.

Ainsi, il conviendrait de mesurer précisément le taux de positivité des IRM cérébrales pour céphalées aux urgences à Lille pour discuter de la pertinence de cette indication fréquente et fréquemment négative.

Extraction automatique d'information depuis des comptes-rendus

Les informations nécessaires à estimer le taux de positivité des IRM cérébrales aux urgences se trouvent au sein des comptes-rendus d'imagerie rédigés par le radiologue de garde. Plus généralement, de nombreuses études rétrospectives tirent leurs données des textes générés en routine dans les hôpitaux (17,18). Ces textes (comptes-rendus d'imagerie, de consultation, d'hospitalisation, de réunion de concertation pluridisciplinaire, etc.), ont en commun d'avoir été rédigés initialement pour transmettre une information à un collègue ou à un patient, sans prévoir que le document pourrait être réutilisé un jour à des fins de recherche. Pourtant, ces documents contiennent une mine d'information d'une qualité précieuse. Elles constituent des données dites de vie réelle, en opposition aux

données issues d'essais cliniques. Ce type de données est identifié par la Haute Autorité de Santé (HAS) comme un enjeu majeur (19) car elles présentent plusieurs points forts : elles sont facilement accessibles, elles peuvent représenter des effectifs importants, suivis sur un temps long, et elles sont collectées a priori, sans connaissance de la question de recherche éventuellement posée, ce qui les rend intéressantes pour limiter certains biais de confirmation. Témoin de l'intérêt des pouvoirs publics pour la réutilisation à grande échelle des données de santé de vie réelle, le *Health Data Hub* est une plateforme nationale de promotion d'outils et de stratégies de valorisation des données de vie réelle. Ces données ont pris une importance particulière avec le développement de l'Intelligence Artificielle (IA), reconnue par le rapport Villani de 2018 sur l'IA (20), car elles peuvent servir à entraîner des modèles de détection ou de prédiction de maladie, ou au contraire bénéficier d'un traitement automatique par des modèles innovants pour faciliter leur réutilisation.

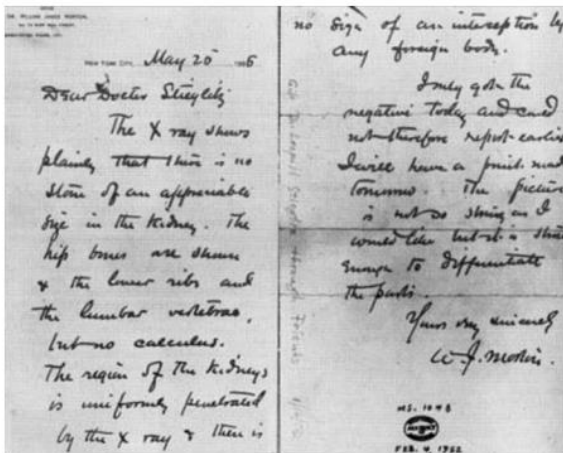
Au CHU de Lille, chaque jour, des centaines de textes sont rédigés à l'occasion d'une hospitalisation ou d'une consultation, dont des dizaines de comptes-rendus d'imagerie. Tous ces textes sont stockés au sein de l'Entrepôt de Données de Santé (EDS) du CHU de Lille, Include. Les données massives ainsi collectées peuvent ensuite être réutilisées à des fins de recherche, conformément aux articles 65 et suivants de la loi « Informatique et Libertés ». Au nombre de 22 en France, dont 17 dans des CHU, les EDS permettent la centralisation de l'information médicale régionale dans une base de données unique, et le développement d'une expertise scientifique sur leur traitement et leur utilisation. À Lille, les équipes d'Include (médecins, *data scientists*) sont chargées du stockage, du traitement et de la valorisation de ces données au travers de leur réutilisation à des fins de qualité et de recherche. Include contient ainsi des résultats de biologie, des données de codage des pathologies ou des actes liés au séjour, et de nombreux documents PDF générés à l'occasion de chaque compte-rendu rédigé en routine par les soignants, parmi lesquels les comptes-rendus de radiologie.

Cependant, malgré leur centralisation au sein d'un même EDS, la réutilisation de ces comptes-rendus pour la recherche est en pratique un défi. En effet, s'agissant de textes générés en routine n'ayant pas initialement vocation à être utilisés pour des analyses secondaires, leur format répond à des contraintes de temps, de simplicité et d'efficacité clinique. Ainsi, en pratique, les comptes-rendus d'imagerie sont rédigés dans un format textuel semi-structuré découpé en paragraphes « indication », « technique », « résultats » et « conclusion », mais dont le contenu est libre. L'essentiel des informations d'intérêt se trouvent donc sous forme textuelle non structurée (figure 1).

Depuis le premier compte-rendu de radiologie rédigé à la plume il y a 128 ans (figure 1), ce format non-structuré est majoritaire (21). Le compte-rendu narratif a ainsi survécu aux innombrables avancées technologiques qui ont fait la force de la radiologie, et son évolution minimale n'a pas su répondre au développement de nouvelles modalités d'imagerie, de l'informatique, et à l'essor de l'IA. S'il possède en effet des avantages pratiques qui expliquent la réticence des radiologues à l'abandonner (21), il se prête néanmoins mal à la réutilisation de données. En effet, une étape d'extraction des informations d'intérêt des textes est nécessaire à leur analyse.

L'extraction d'information est la transformation d'une donnée non structurée en une donnée structurée d'intérêt, qui pourra être intégrée dans des analyses quantitatives ou qualitatives (figure 2). Pour des textes, cette transformation est basée sur un traitement humain (relecture et cotation des textes), ou un traitement automatique selon des règles plus ou moins complexes, de la détection de mot au traitement avancé du langage naturel. Cette extraction est difficile : le format des comptes-rendus ne répondant à aucune règle de vocabulaire, une même anomalie peut être décrite de nombreuses façons par des radiologues différents, voire par un même radiologue. Les mots employés et leurs variations infinies traduisent souvent des subtilités d'interprétation que les radiologues sont parfois en fait les seuls à maîtriser (22). Ainsi, par exemple, si une étude nécessite d'identifier tous les

examens où sont décrits des anévrismes intracrâniens, la recherche simple du mot « anévrisme » dans la base de données risque de repérer de nombreux examens non pertinents : anévrismes thoraciques ou abdominaux, négations (« pas d'anévrisme »), présence d'un anévrisme familial dans les antécédents décrits dans l'indication, etc. À l'inverse, la recherche simple risque de ne pas repérer tous les dossiers où un synonyme ou une périphrase sont utilisés (« image d'addition », « dilatation sacciforme »), ou de ne pas identifier les textes contenant une faute de frappe ou de dictée. Le niveau de compétence pour traiter le texte dans le but d'extraire cette information est pourtant bas, et ne repose que sur une compréhension basique de la langue française. Cette tâche pourrait être confiée à un assistant sans compétence médicale sérieuse mais parlant couramment le français, sous réserve de lui expliquer la tâche et de lui donner quelques exemples. Cependant, les méthodes informatisées disponibles actuellement nécessitent une expertise dans la construction d'une requête optimale voire l'entraînement de modèles spécifiques, et la recherche de comptes-rendus d'intérêt repose finalement en pratique sur une relecture humaine de centaines de textes, longue et coûteuse.



May 20, 1896

Dear Dr. Stieglitz,

The X ray shows plainly that there is no stone of an appreciable size in the kidney. The hip bones are shown & the lower ribs and lumbar vertebrae, but no calculus. The region of the kidneys is uniformly penetrated by the X-ray & there is no sign of an interception by any foreign body. I only got the negative today, and could not therefore report earlier. I will have a print made tomorrow. The picture is not so strong as I would like, but it is strong enough to differentiate the parts.

Yours
William James Morton

sincerely,

Lille, le 17 juin 2024

SCANNER ABDOMINOPELVIEN

Indication:

Bilan de calculs rénaux

Protocole:

Acquisition hélicoïdale sans injection sur l'abdomen et le pelvis

Résultats:

A droite comme à gauche:
Reins de taille et de morphologie normales (110mm à droite, 109mm à gauche).
Bonne trophicité corticale.
Pas de calculs rénaux.
Pas de dilatation des cavités pyélocalicielles.
Pas d'infiltration péri-rénale.

Conclusion:

Pas de calculs rénaux.

Bien confraternellement,
Dr Guillaume Jean-Mouton

Figure 1: 128 ans séparent le compte-rendu de gauche (première trace de compte-rendu pour la jeune discipline de la Roentgenologie) du compte-rendu de droite.

En opposition au compte-rendu non structuré majoritaire, le compte-rendu structuré se présente sous la forme d'une liste d'items que le radiologue renseigne (23) (par exemple : présence de lésion ischémique oui/non, taille des lésions cibles en cm pour le suivi oncologique, etc.). Ce format porte depuis des années l'espoir d'une « révolution de l'information » (24) grâce à la réutilisation massive des données structurées qu'il renferme. En effet, sous ce format, l'étape d'extraction de l'information est minime, les données étant nativement propres à une réutilisation rationnelle. Cependant, depuis sa conception, il peine à être implémenté en pratique (25), en dehors de quelques indications répondant à un cahier des charges précis (mammographie, par exemple). Plusieurs obstacles expliquent ce retard, en particulier la liberté ressentie par les radiologues à nuancer plus en détail leur interprétation dans un compte-rendu narratif (21). De plus, même si le format structuré était

adopté du jour en lendemain, un travail massif de structuration post-hoc des milliers de comptes-rendus déjà présents dans les bases de données serait nécessaire à leur réutilisation.

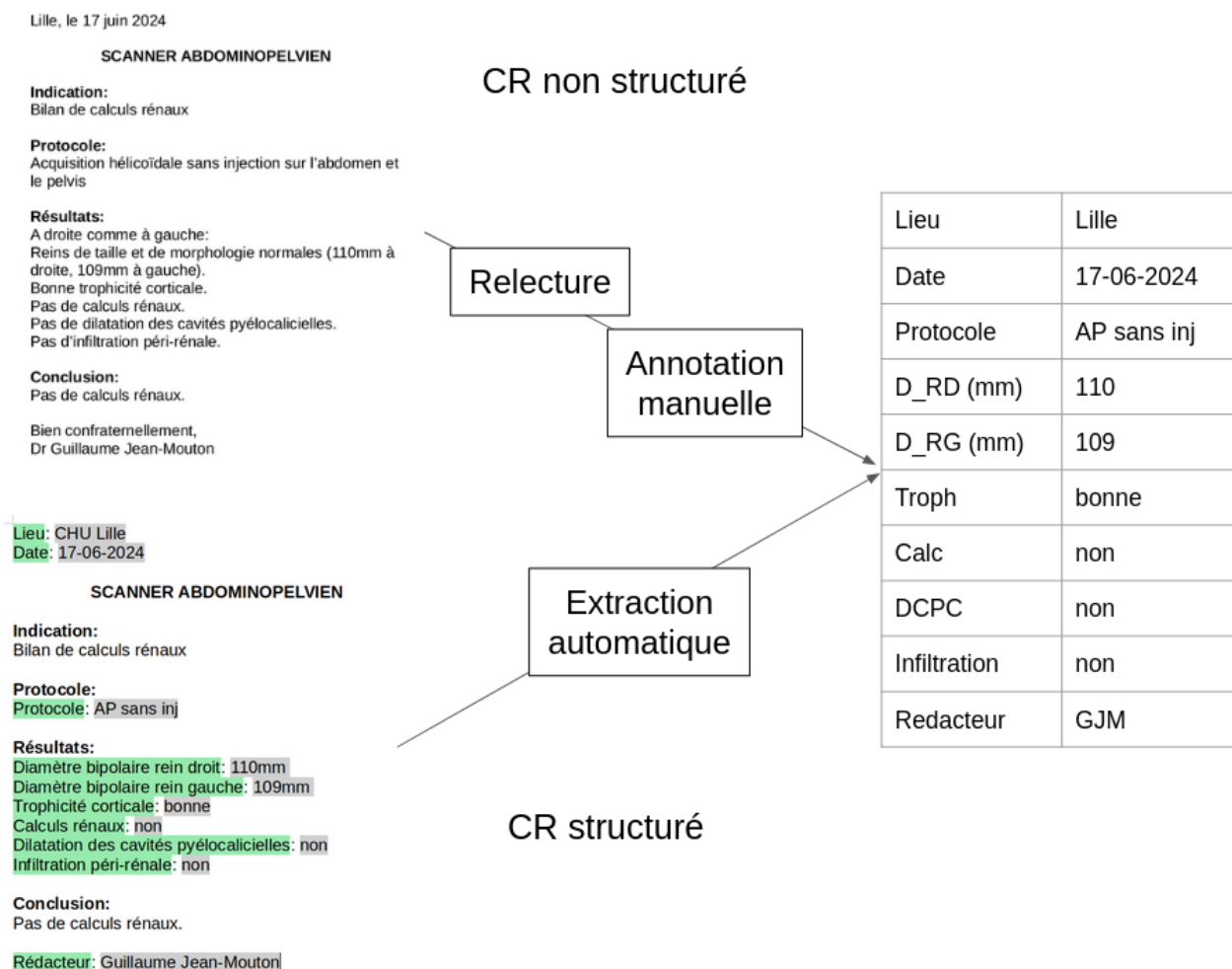


Figure 2: Extraction d'information depuis un compte-rendu non structuré et structuré, convergeant vers la même base de données utilisable.

Ainsi il existe un besoin d'extraire des variables depuis des bases de données de textes médicaux à large échelle à l'aide d'outils informatiques adaptables. Ces outils pourraient se baser sur une excellente compréhension de la langue française pour gérer les négations et les périphrases et synonymes, et un socle de connaissances médicales pour gérer les synonymes et justifier certaines annotations.

Dans ce contexte, le développement des grands modèles de langage de type GPT a récemment bouleversé le champ du traitement du langage naturel.

Les modèles de fondation / *Large Language Models* (LLMs)

Les LLM sont des modèles d'intelligence artificielle reposant sur l'entraînement non supervisé de réseaux de neurones sur de vastes quantités de textes. Les modèles génératifs de texte comme ChatGPT fonctionnent en prédisant prospectivement le prochain mot d'un texte. À partir d'une consigne d'un utilisateur, le modèle prédit le premier mot de la réponse, puis en fonction de la consigne et du premier mot de la réponse, le second mot, etc. Pour prédire efficacement le « prochain mot », ces modèles calculent la probabilité d'apparition de chaque mot en fonction du contexte qui leur est donné (figure 3). La tâche d'un LLM est donc de donner la distribution de probabilité d'apparition de mots en fonction du texte qui précède, créant ainsi un texte pas à pas.

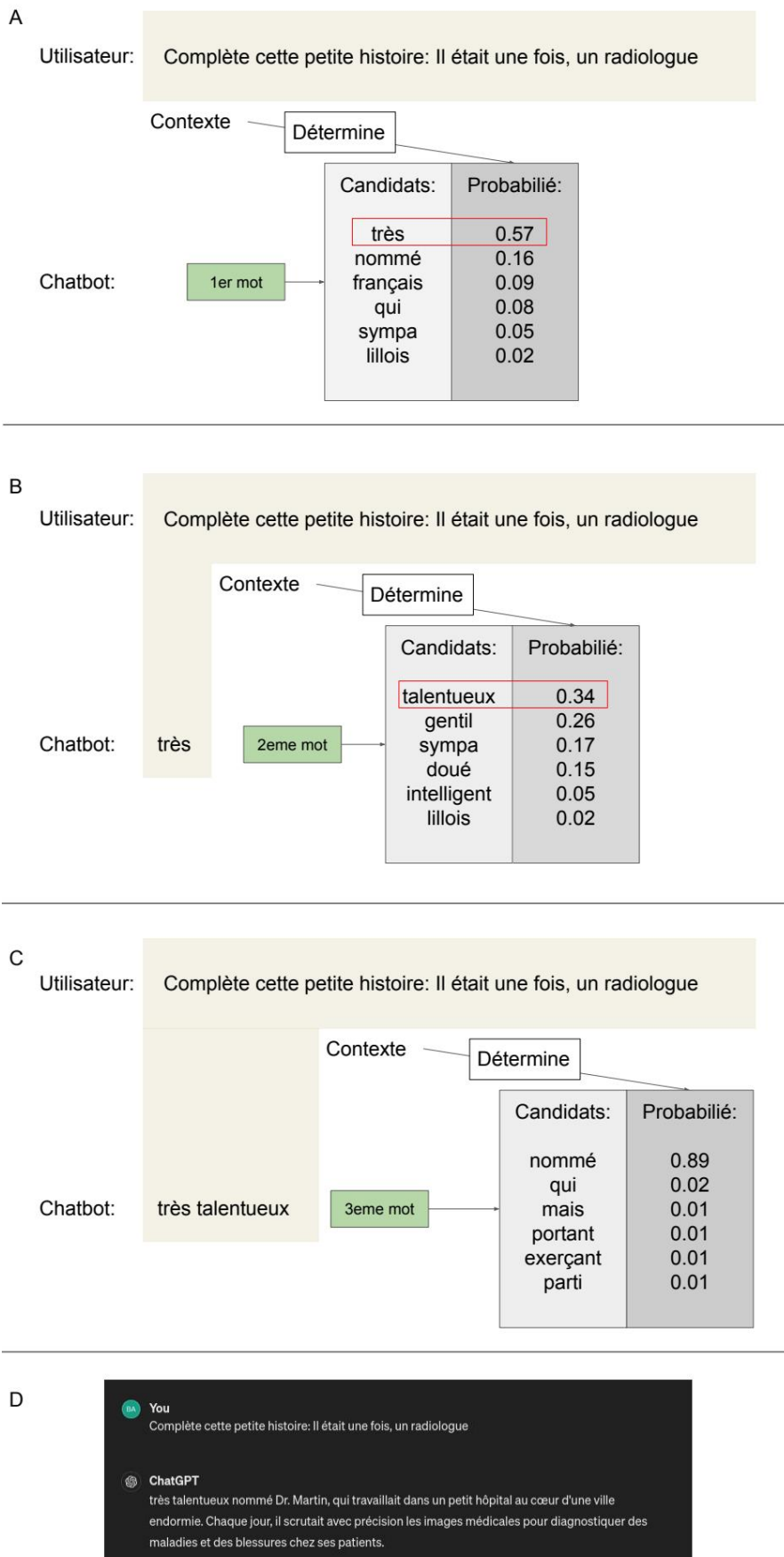


Figure 3: Processus de génération d'un texte par un LLM, mot après mot. D : exemple réel avec GPT3.5.

Les LLM ont reçu une attention toute particulière de la part de la communauté scientifique et des médias, qui s'explique par leur extraordinaire capacité à traiter et à produire du langage. Leur architecture et leur entraînement ressemblent aux modèles d'IA habituels en médecine, avec quelques spécificités.

Réseaux de neurones

Les LLM, comme les modèles d'IA de classification d'images, sont basés sur des réseaux de neurones. Ces réseaux sont des arbres de calcul très vastes, constitués de milliards d'unités de calcul, appelés neurones. Ces neurones effectuent de nombreuses opérations à partir de données d'entrée (la valeur d'intensité de pixels d'une image, par exemple), pour converger sur une sortie d'intérêt (probabilité que l'image soit un chat, par exemple). Initialement, les calculs effectués n'ont aucune raison de donner une réponse pertinente. Les milliards de neurones ne sont qu'un espace potentiel de calcul, chacun pouvant modifier ses paramètres, ou poids, grâce à l'entraînement. L'entraînement, ou apprentissage, de réseaux de neurones consiste à modifier pas à pas les opérations effectuées par chaque neurone. À chaque itération, le système confronte le résultat obtenu par l'arbre de calcul (prédiction de chat ou non) à la vérité connue (chat réel ou non). En fonction de la différence entre la prédiction et la vérité, les calculs effectués sont modifiés pour la prochaine tentative selon des lois, comme par exemple la rétropropagation des gradients. Au terme de l'apprentissage, le réseau de neurones se trouve dans un état qui minimise la différence entre la prédiction et la vérité, c'est-à-dire le nombre d'erreurs. La propriété émergente décisive de ces réseaux est leur capacité à généraliser, c'est-à-dire à être performant sur des données sur lesquelles ils n'ont pas été entraînés. L'apprentissage des réseaux les plus performants est dit « profond » (« *deep learning* ») car les neurones y

sont disposés en plusieurs couches, pouvant être spécialisées et intégrer éventuellement des noyaux convolutifs, les opérations se réalisant en parallèle, mais aussi successivement.

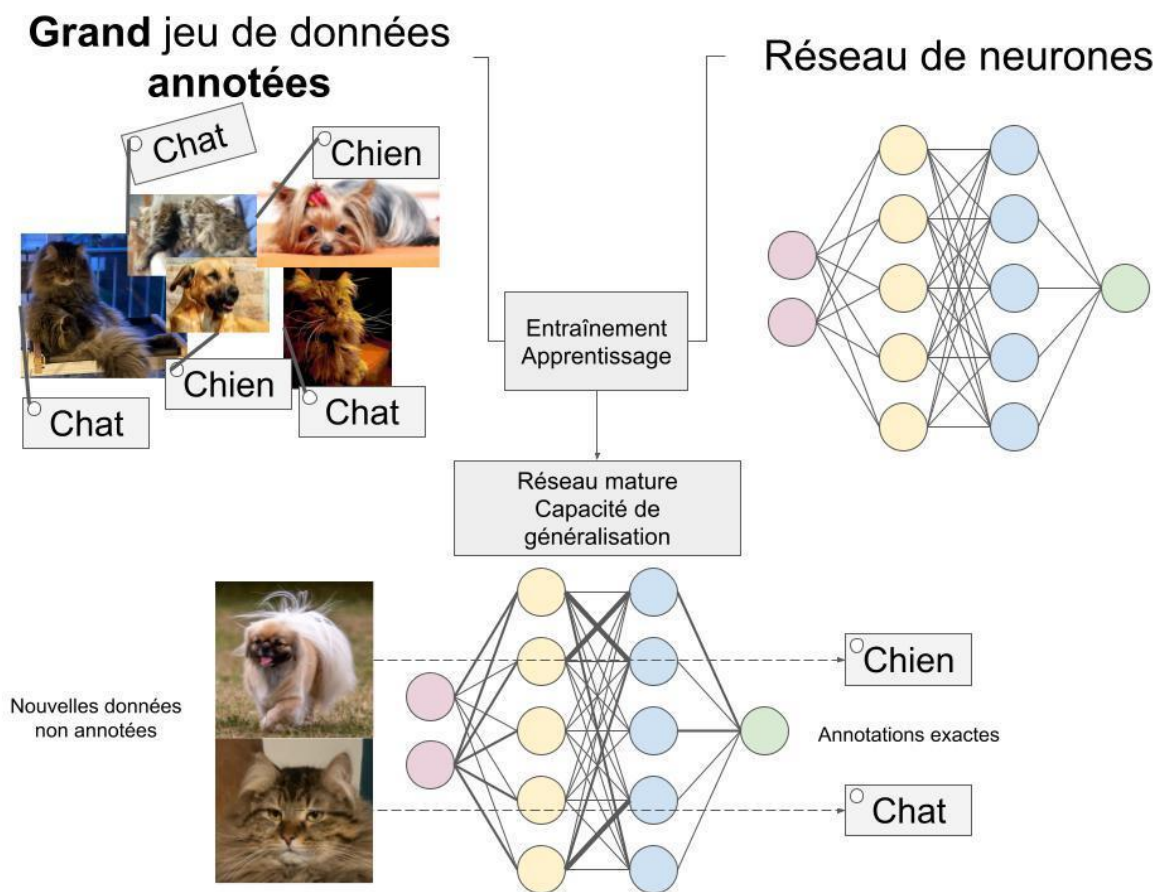


Figure 4: Les capacités de généralisation des modèles de tri d'image reposent sur l'entraînement de réseaux de neurones sur de vastes quantités de données.

Le développement et l'essor de ces réseaux marque le premier boom de l'IA, à partir des années 2000. Ces avancées rapides aboutissant aux premiers algorithmes de classification d'images, et aux premières applications médicales notamment en radiologie (26), ont été permises par la coïncidence de deux phénomènes. Premièrement, l'explosion des capacités de calcul des ordinateurs prédite par la loi de Moore, l'entraînement des réseaux de neurones étant extrêmement coûteux en ressources informatiques (le coût estimé de l'entraînement de GPT en ressources informatiques et en énergie est de plus de 100 millions de dollars (27)). Deuxièmement, le développement et le partage de bases de

données massives, l'entraînement des réseaux de neurones étant extrêmement gourmand en données de qualité.

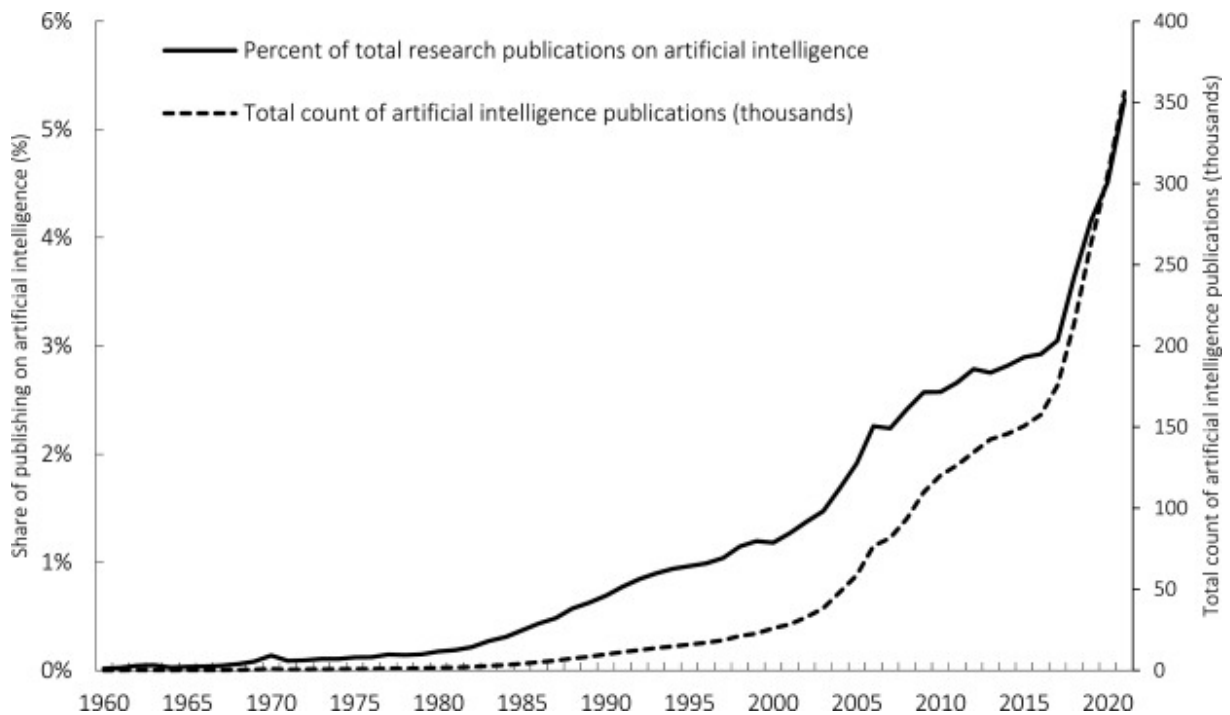


Figure 5: Nombre et part des publications scientifiques traitant d'intelligence artificielle, d'après Hajkowicz et al. 2023.

Entraînement supervisé

Les modèles habituels nécessitent un entraînement supervisé sur des données annotées, qui sont des données à laquelle une étiquette est apposée : on donne à cette image l'étiquette « chat » et à une autre l'étiquette « chien » (figure 4). Au cours de l'entraînement, ce sont à ces étiquettes que sont confrontées les prédictions du modèle pour décider si la prédiction du modèle est correcte ou non, et modifier les paramètres en conséquence. Encore de nos jours, ces étiquettes sont le résultat d'une annotation manuelle très coûteuse. Par exemple, les *Captcha*, ces systèmes de sécurité informatique qu'il est parfois nécessaire de remplir sur le web, consistent parfois en des tâches d'annotation d'image (cliquer sur les images avec un bus, orienter un objet vers une certaine direction, etc.), qui servent à construire d'immenses bases de données annotées manuellement par

des milliers d'utilisateurs-travailleurs (28). C'est cette étape d'annotation qui rend rares les jeux de données de qualité pour l'entraînement des modèles d'IA habituels.

Analyser pas à pas le fonctionnement des LLMs permet de comprendre en quoi ils se distinguent des modèles d'IA habituels, et pourquoi ils sont si performants.

Word embedding

La première étape de la réponse à une consigne pour un LLM est la tokenisation, c'est-à-dire son découpage en mots. En réalité, une unité textuelle, ou « *token* », représente plus souvent un fragment de mot, ou un signe de ponctuation, mais pour la simplicité des explications, nous emploierons le raccourci « *token* = mot ». Afin d'effectuer des calculs aboutissant à une probabilité d'apparition du « prochain mot », principe qui sous-tend le fonctionnement des LLMs, les mots sont transformés en vecteurs, une liste de nombres : c'est le *word embedding*. Sa fonction est en pratique de décrire la position du mot au sein d'un espace d'autres mots. Deux mots dont les significations sont proches, par exemple « chausson » et « pantoufle », auront un vecteur numériquement proche, c'est-à-dire une position dans l'espace proche. Il est possible d'effectuer des opérations sur ces vecteurs : par exemple, les mots « neveu » et « nièce » auront des vecteurs proches, de même que les mots « roi » et « reine ». Au-delà de rapprocher des mots dont le sens est proche, cet espace géométrique témoigne de la relation des mots entre eux. La même relation logique lie les concepts de « neveu »/« nièce » et « roi »/« reine » : le masculin et le féminin. Cette relation logique est aussi traduite dans l'espace géométrique du *word embedding*, où il est remarqué que la soustraction des vecteurs « neveu » et « nièce » donne environ la même valeur que la soustraction de « roi » et « reine ». Les mots de vocabulaire du modèle, au terme de son entraînement, sont donc situés dans espace géométrique à haute dimension, où la position de chaque mot, traduite par un vecteur, informe sur son sens, et ses relations éventuelles avec d'autres mots. Cependant, les mots ne sont pas isolés dans une phrase,

et un « chausson en laine » et un « chausson aux pommes » sont deux « chaussons » sans aucun lien : le « chausson » en laine devra avoir un vecteur proche de « pantoufle » et le « chausson » aux pommes un vecteur proche de « viennoiserie ». C'est le rôle du processus d'attention.

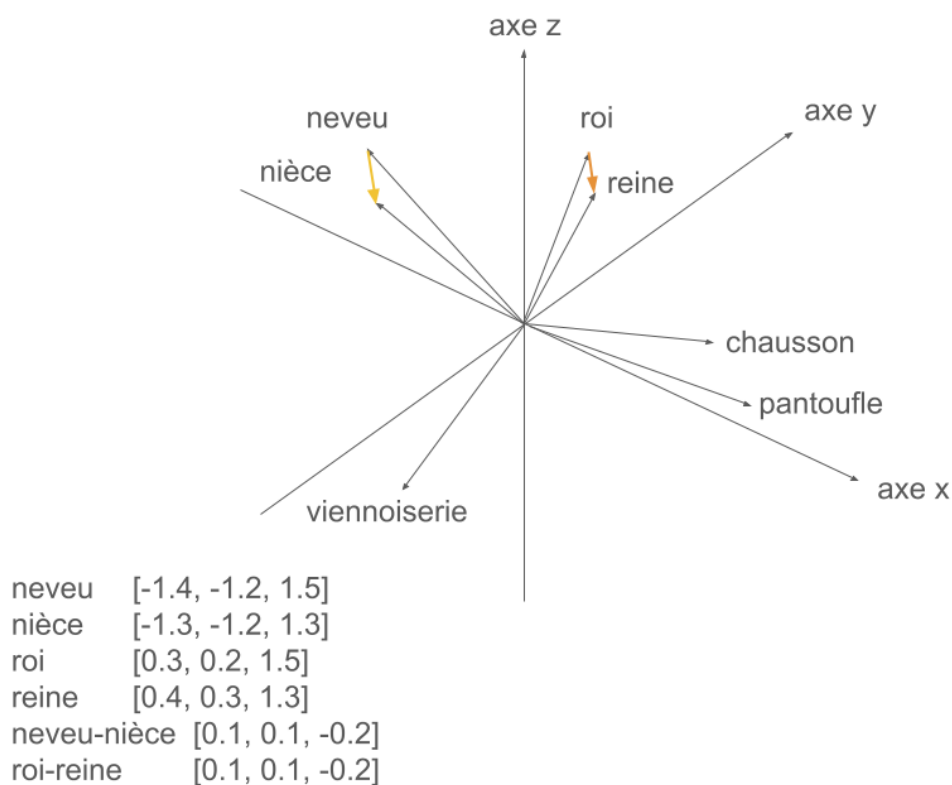


Figure 6: Position de mots au sein d'un espace simplifié en 3 dimensions. Les mots « neveu » et « nièce » ont des vecteurs proches, comme « roi » et « reine » et « chausson » et « pantoufle ». La même relation logique lie « neveu » et « nièce » et « roi » et « reine », ce qui se traduit par la proximité de la valeur des vecteurs « neveu-nièce » et « roi-reine », en jaune et orange, respectivement.

Processus d'attention

« *Attention is all you need* », c'est le titre provocateur de la publication scientifique la plus influente en IA ces 5 dernières années (29). Rédigée en 2017 par une équipe de Google, citée 118.000 fois, elle décrit une méthode qui permet de mettre à jour le vecteur descriptif des mots en fonction de leur contexte dans une phrase. Appelée « attention », elle

permet de rapprocher le « chausson aux pommes » du « pain au chocolat », même si les mots « chausson » et « pain » sont initialement éloignés dans l'espace géométrique. Encore une fois, il s'agit d'une suite d'opérations sur les vecteurs, conditionnées par des vecteurs de « requête » et de « clé » propres à chaque mot et à chaque tête d'attention, qui précisent pas à pas le sens de chaque mot en fonction de son contexte. Le vecteur de requête de chaque mot représente sa disposition à modifier sa position dans l'espace géométrique en fonction d'autres mots. Pour des noms par exemple, il indique leur capacité à supporter des adjectifs placés avant ou après eux, ou après un verbe dont le nom est le sujet. Le vecteur de clé, lui, code justement le rôle des mots modificateurs, et l'intensité des modifications à apporter à la valeur du vecteur initial. Ici, le « chausson » a une requête propre, qui intercepte le vecteur clé des mots « pommes » ou « laine », qui viennent modifier la position finale de « chausson » selon le contexte du mot. Les processus d'attention sont répétés itérativement, de sorte qu'au terme de ceux-ci, le vecteur lié à chaque mot ne représente plus simplement la position théorique du mot dans un espace statique, mais sa place informée du contexte qui l'entoure. Le chausson de la phrase « Le chausson aux pommes sort du four », est probablement proche de « viennoiserie », mais sa relation avec « four » informe probablement du fait qu'il soit encore chaud, ce qui sera intégré dans le vecteur de valeur de « chausson ». Justement, si la phrase se poursuit : « Le chausson aux pommes sort du four. Il est encore », le modèle peut calculer le vecteur de valeur du prochain mot, qui, en raison des processus d'attention, sera informé qu'il s'agit d'un mot en relation avec « chausson », qui est le sujet de « est » au travers du pronom « il ». La position de « chausson », informée du contexte de « four », précisée de « encore », qui témoigne d'un processus amené à s'amenuiser, crée numériquement un vecteur se situant dans l'espace géométrique, et les mots les plus proches de la position théorique du prochain mot sont les candidats à être le dernier mot de la phrase. Sur GPT-3.5, la conversation est visible en figure 8.

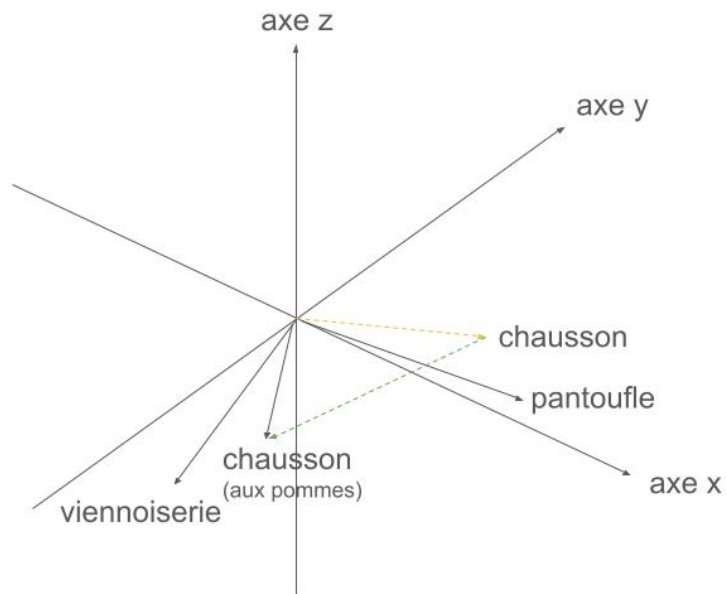


Figure 7: Initialement proche de « pantoufle », la position dans l'espace simplifié de « chausson » (pointillés jaunes) est modifiée par le processus d'attention (en pointillés verts), pour se rapprocher de « viennoiserie ».

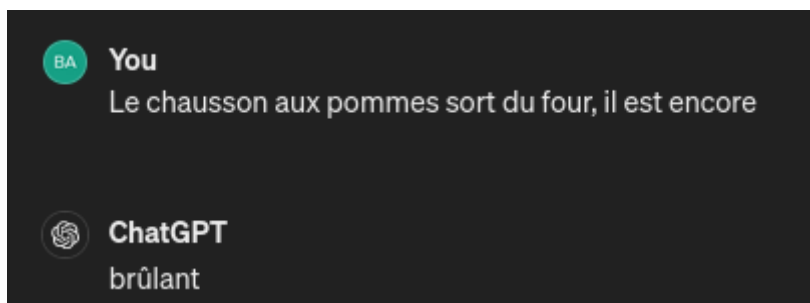


Figure 8: Conversation avec GPT-3.5 illustrant le processus d'attention.

Entraînement non supervisé

Le *word embedding*, comme le processus d'attention, ne sont pas des propriétés innées des LLMs. Ces opérations sont, comme pour les réseaux de neurones habituels, initialement aléatoires. Pour le *word embedding* par exemple, chaque mot est positionné aléatoirement dans l'espace. Puis, phrase après phrase, au cours de l'apprentissage, en modifiant pas à pas des milliards de paramètres, les opérations vont se modifier. Au terme de l'entraînement, le modèle devrait être capable de bien prédire le prochain mot d'un texte. Cet entraînement des LLMs se distingue de l'entraînement des autres modèles habituels en médecine : il n'est pas supervisé, c'est-à-dire que les jeux de données sur lesquels le modèle apprend ne sont pas annotés, ils n'ont pas d'étiquette préalable. Les LLM ont pour but de prédire le « mot suivant », ce qui signifie que la vérité terrain de la tâche d'un LLM est le vrai prochain mot d'un texte tronqué. Il suffit alors pour entraîner un LLM de lui fournir de grandes quantités de textes, dont des mots seront automatiquement dissimulés, et qui devront être retrouvés. En pratique, la phrase « J'entends les ronronnements du chat. » peut servir d'entraînement à un modèle de langage, en cachant le dernier mot, et en demandant au modèle de le retrouver. La vérité établie automatiquement est ici « chat » (c'est le mot dissimulé au modèle). Si le modèle prédit « cheval » dans son état actuel, il devra continuer son entraînement sur d'autres textes. Les LLM bénéficient donc d'une libération de la donnée : puisqu'aucune annotation n'est nécessaire, tous les textes sont des bases d'entraînements potentielles.

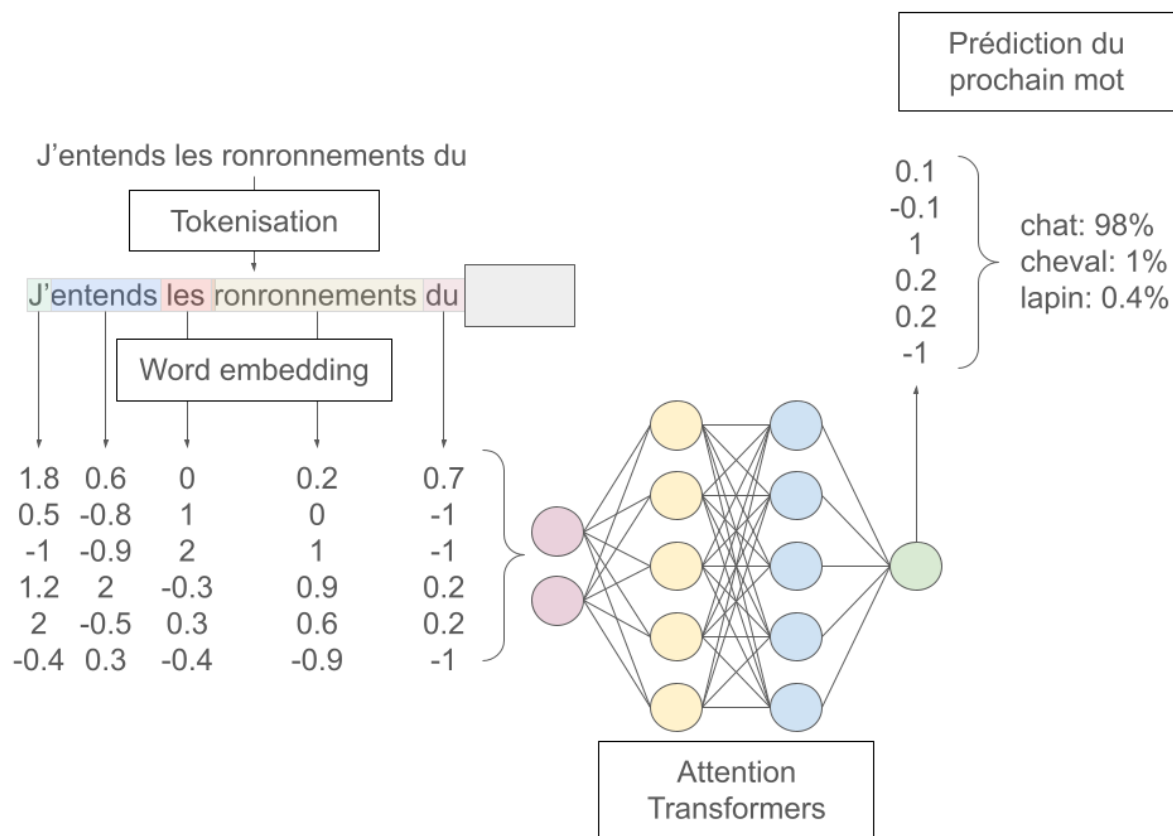


Figure 9: Schéma du fonctionnement d'un LLM. La case grisée est le mot caché du texte, qui est comparé à la sortie du modèle. Si la sortie ne correspond pas au mot caché, il s'agit d'une erreur et le modèle doit modifier ses paramètres.

L'essor des LLM tient donc à des progrès conceptuels (processus d'attention notamment), à des progrès perpétuels en termes de puissance de calculs, et à l'extraordinaire qualité et diversité des bases d'entraînement potentielles. Alors qu'un modèle de tri d'image nécessite des bases de données coûteuses d'images annotées à la main, les LLM tirent leurs données du minage automatique (*scrapping*) du web (30) (Wikipédia, StackOverflow, Reddit, etc.). Ce faisant, ils récoltent un immense trésor de textes multilingues en langage naturel, qui leur donne les clés de la manière dont fonctionne le langage : les négations, les périphrases, le lien complexe entre des mots éloignés au sein d'une même phrase. De plus, et de manière plus inattendue, ils récoltent aussi des connaissances spécialisées (31) (en informatique sur le forum StackOverflow, en science et en médecine sur des forums spécialisés ou sur Wikipédia, etc.).

Ces données sur lesquelles les modèles sont entraînés ont un statut légal incertain (32). Issues pour la grande majorité de données publiques du web, elles ne sont néanmoins pas libres, leur utilisation pouvant être soumise à des licences spécifiques. Par exemple, si des articles de journaux en accès ouvert sont consultables gratuitement, ils appartiennent malgré tout aux journaux en question et ne peuvent pas être réutilisés dans un but commercial sans accord explicite. En 2024, une plainte a été déposée contre OpenAI par huit journaux américains dont le *Chicago Tribune* (33), qui suspecte que les articles publiés sur le site du journal aient été réutilisés illégalement pour entraîner ChatGPT (texte issu de la plainte : « [OpenAI]’s large-scale commercial exploitation of the Publishers’ Works is not licensed, nor have [OpenAI] received permission from the Publishers to copy and use their works to build their GenAI tools. »). De son côté, OpenAI se défend en argumentant que ces données sont accessibles au public, et que ChatGPT y accède au cours de son entraînement comme n’importe quel utilisateur pourrait s’en inspirer pour se construire une opinion.

Il est difficile de déterminer précisément sur quelles données les modèles de langages ont été entraînés (34), en particulier pour les modèles commerciaux, cette information appartenant au secret des affaires. Il n’est pas légalement nécessaire d’apporter la liste des textes sur lesquels le modèle a été entraîné, même pour les modèles open-source. Les concepteurs des LLMs peuvent faire de la spécificité de leur base de données un avantage stratégique sur les concurrents, le marché des IA génératives étant évalué à 100 milliards de dollars d’ici 2030 (35).

En médecine, la principale source de données d’entraînement pour les grands modèles de langage est MIMIC (36), une grande base de données médicales contenant les résultats de biologie, les comptes-rendus cliniques et de radiologie et les données cliniques heure par heure de 430,000 séjours hospitaliers au sein du *Beth Israel Deaconess Medical Center* à Boston, Massachusetts. Des modèles sont déjà entraînés sur ces bases de

données, mais il n'est pas connu si OpenAI utilise ou non ces informations pour entraîner ChatGPT. Ces données dé-identifiées restant considérées comme « sensibles » et soumises à une licence d'utilisation non-commerciale, l'utilisation de MIMIC pour l'entraînement de modèles de langage commerciaux ou diffusés en ligne est dans une zone grise. En effet, d'une part, l'utilisation des données de MIMIC pour améliorer les performances d'un modèle commercial peut constituer une atteinte au point 6 de la licence, puisqu'il ne s'agit pas d'une utilisation dans le cadre de recherche scientifique. D'autre part, le partage d'un modèle entraîné sur MIMIC n'assure pas qu'un utilisateur mal intentionné n'accède aux données d'entraînement par des requêtes pirates, le diffuseur du modèle portant alors atteinte aux points 2 et 3 de la licence.

It is hereby agreed between the data requestor, hereinafter referred to as the "LICENSEE", and MIT-LCP, that:

1. The LICENSEE will not attempt to identify any individual or institution referenced in PhysioNet restricted data.
2. The LICENSEE will exercise all reasonable and prudent care to avoid disclosure of the identity of any individual or institution referenced in PhysioNet restricted data in any publication or other communication.
3. The LICENSEE will not share access to PhysioNet restricted data with anyone else.
4. The LICENSEE will exercise all reasonable and prudent care to maintain the physical and electronic security of PhysioNet restricted data.
5. If the LICENSEE finds information within PhysioNet restricted data that he or she believes might permit identification of any individual or institution, the LICENSEE will report the location of this information promptly by email to PHI-report@physionet.org, citing the location of the specific information in question.
6. The LICENSEE will use the data for the sole purpose of lawful use in scientific research and no other.
7. The LICENSEE will be responsible for ensuring that he or she maintains up to date certification in human research subject protection and HIPAA regulations.
8. The LICENSEE agrees to contribute code associated with publications arising from this data to a repository that is open to the research community.
9. This agreement may be terminated by either party at any time, but the LICENSEE's obligations with respect to PhysioNet data shall continue after termination.

Figure 10: Licence d'utilisation de MIMIC

En effet, si les bases de données sont secrètes et ne sont pas diffusées explicitement avec les modèles commerciaux ou open-source, chaque texte d'entraînement influence le modèle, qui construit une mémoire des textes lus. Cette mémoire n'est pas accessible à l'état brut, elle réside dans le poids de chaque neurone. Cependant, à l'aide de consignes pirates, il est possible de faire révéler des textes d'entraînement (34) au modèle. L'utilisateur mal intentionné peut ainsi accéder à des informations éventuellement confidentielles si la base de données initiale n'est pas parfaitement anonymisée. Si de telles fuites de données

sont en pratique rares, elles sont théoriquement possibles et se sont déjà produites (37), permettant aux utilisateurs pirates de retrouver des données confidentielles.

L'entraînement et l'utilisation des modèles de langage engage donc les industriels et les chercheurs à la prudence, notamment s'ils traitent des données médicales, donc sensibles.

OpenAI vs Open-source

La scène des LLM est dominée depuis sa naissance par ChatGPT, le modèle développé par OpenAI. Sorti en 2018, et amélioré chaque année, il repose sur des technologies propriétaires, et ne peut être utilisé que depuis les serveurs d'OpenAI. Son utilisation est payante. Le modèle ne peut pas être téléchargé sur un ordinateur classique. Même s'il le pouvait, sa taille serait incompatible avec les puissances de calculs courantes, même possédant des cartes graphiques de très haute gamme. Il disposerait de 10^{12} (1,760,000,000,000) paramètres, contre 10^{10} paramètres, 100 fois moins, pour des modèles plus courants qui nécessitent déjà des cartes graphiques avancées pour prédire péniblement quelques mots par seconde.

Cependant, en raison de ses grandes performances et de sa facilité d'accès, des chercheurs ont commencé à utiliser ChatGPT comme assistant pour traiter des textes médicaux. En radiologie plus que dans n'importe quelle autre spécialité médicale, les LLMs ont fait l'objet d'un nombre croissant de publications scientifiques (figure 11). L'équipe de Keno Bressemer a publié en avril 2023 un travail consistant à structurer a posteriori des comptes-rendus de radiologie (38). Si le résultat est étonnant (100 % de réussite), il est néanmoins théorique, puisqu'il a été fait sur des comptes-rendus fictifs, les comptes-rendus réels ne pouvant pas être transférés à des serveurs commerciaux sans accord des autorités locales. Toujours en utilisant ChatGPT, Fink et collègues (39) ont extrait les mesures et la

localisation de lésions secondaires de cancer depuis 424 comptes-rendus de scanner, avec une précision remarquable de près de 99 %. Jethani et collègues (40) ont extrait des scores cliniques de comptes-rendus de consultation avec une précision de 83 %.

En plus de leur capacité à extraire des informations, ces modèles peuvent être utilisés pour améliorer la qualité des comptes-rendus. Amin et collègues (41) ont ainsi utilisé plusieurs modèles de langage pour simplifier des comptes-rendus de radiologie, sans perdre d'information critique. Gertz et collègues ont montré que ChatGPT pouvait relire des comptes-rendus de radiologie pour y repérer des erreurs, avec une efficacité comparable avec celles de radiologues, pour un coût bien plus faible (42). Les LLM possèdent donc des caractéristiques prometteuses pour l'extraction de données depuis des textes médicaux, et l'amélioration de leur qualité.

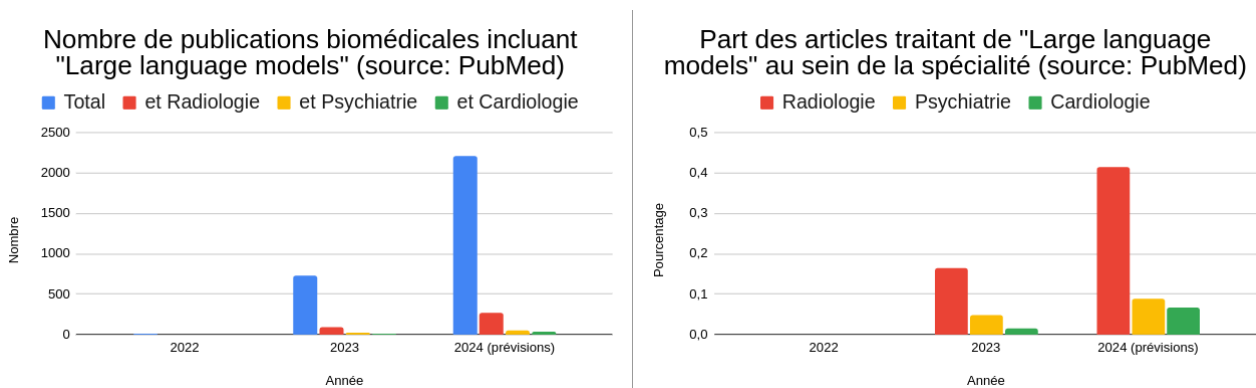


Figure 11: Nombre de publications biomédicales traitant de LLMs au total, et dans trois spécialités médicales (Radiologie, Pscyhiatrie et Cardiologie, choisies en raison d'un volume équivalent de publications indexées dans PubMed), et part de ces publications au sein du nombre total d'articles dans chaque spécialité.

Cependant, les modèles utilisés jusqu'alors dans la plupart des études sont commerciaux, et leur utilisation oblige de transmettre à des serveurs tiers des informations médicales critiques (37). Leur implémentation en pratique courante pour traiter des données médicales critiques n'est donc pas envisageable.

En parallèle de ces solutions commerciales opaques, des modèles open-source, téléchargeables et utilisables à volonté sur des ordinateurs locaux, existent et se développent. Vicuna (43) est un modèle de langage plus léger que GPT (13 milliards de paramètres), pouvant fonctionner sur des serveurs au sein des entrepôts de données. Il a été développé par un groupe de chercheurs d'universités californiennes, et ses données d'entraînement sont transparentes. En Europe, le modèle le plus prometteur est Mistral, un LMM français à visée commerciale mais dont les paramètres ont été initialement diffusés librement, avant de devenir complètement clos en 2024, répétant la transition ouvert-fermé opérée plus tôt par « Open »AI (44). Ce type de modèles possède l'avantage décisif de pouvoir traiter des informations localement sur un ordinateur sur site, sans transmission de données à des tiers. De plus, leur utilisation est gratuite. Alors que l'utilisation des versions les plus récentes de ChatGPT pourrait induire un coût non négligeable pour de grands nombres de comptes rendus, les modèles open-source peuvent fonctionner sur des ordinateurs courants (équipés de cartes graphiques récentes), sans surcoût. Leur installation est simple, et permet une transparence totale sur la version utilisée, à l'inverse de ChatGPT. Utiliser un logiciel dont la version est traçable et communicable est important dans une perspective de coopération sur plusieurs hôpitaux, ce qui est possible avec les modèles Open-Source. Très récemment, Vicuna a été utilisé par Mukherjee et collègues (45) pour extraire des données depuis des comptes-rendus de radiologie standard (radiographie de thorax), avec une précision d'environ 84 %. Les modèles open-source offrent donc des perspectives prometteuses pour l'extraction automatique d'information depuis des textes médicaux, tout en offrant des garanties dans la protection des données des patients.

Problématique

Les céphalées sont un motif fréquent de demande d'IRM aux urgences. Une meilleure connaissance du taux de positivité de cet examen et des caractéristiques des patients permettrait de mieux identifier les patients qui pourraient bénéficier d'une IRM en urgence. Ces données sont difficiles à extraire des comptes-rendus textuels libres rédigés par les soignants au cours des soins courants. Les LLMs offrent des perspectives nouvelles dans le traitement automatique du langage naturel. Cependant, la majorité des études publiées à ce jour traitent des comptes-rendus en anglais à l'aide de modèles commerciaux dont l'utilisation pose des problèmes de confidentialité des données. Les performances d'un modèle open-source sur des comptes-rendus de vie réelle en français ne sont donc pas connues.







Nous nous sommes donc proposé d'utiliser un LLMs open-source pour extraire des données de comptes-rendus d'IRM cérébrales réalisées en urgence, dans le but de calculer le taux de positivité des examens réalisés pour céphalées.

Article en Anglais

Original Research

 Open Access

Performance of an Open-Source Large Language Model in Extracting Information from Free-Text Radiology Reports

 Bastien Le Guellec , Alexandre Lefèvre, Charlotte Geay,  Lucas Shorten, Cyril Bruge, Lotfi Hacein-Bey,  Philippe Amouyel, Jean-Pierre Pruvo,  Grégory Kuchcinski,  Aghiles Hamroun

▼ **Author Affiliations**

Published Online: May 8 2024 | <https://doi.org/10.1148/ryai.230364>

Radiology: Artificial Intelligence

Reproduit ici selon les termes de la licence *Creative Commons BY 4.0*

Keywords: Large Language Model (LLM), Generative Pre-trained Transformers (GPT), open-source, information extraction, report

Key points:

- Vicuna, an on-premises, open-source Large Language Model reviewed 2398 emergency brain MRI free-text reports and achieved near-perfect performance metrics for detecting the presence of headache in the clinical context, the use of contrast medium injection in the protocol, and study categorization as normal or abnormal (respective F1 scores (95%CI): 98.0 (97-98.7), 99.6 (99.0-99.9) and 97.3 (95.4-98.6))
- It achieved etiological inference between a radiological finding and a symptom with 82% accuracy, a task challenging even radiologists (inter-rater kappa <0.80)

Summary statement: A local open-source Large Language Model can extract information from real-life free-text radiological reports with an excellent accuracy without specific training.

Abbreviations:

GPT: Generative Pre-trained Transformers

LLM: Large Language Model

GPU: Graphics Processing Unit

BERT: Bidirectional Encoder Representation from Transformers

Abstract

Background: The need for reliable and scalable solutions to extract clinically relevant data from unstructured radiological reports is increasing. While recent results on the use of GPT-4 and ChatGPT for this task seem promising, their commercial cloud-based nature is a serious limitation for their widespread utilization.

Purpose: To assess the performance of a local open-source Large Language Model (LLM) on various information extraction tasks from real-life emergency brain MRI reports.

Methods and Materials: All consecutive emergency brain MRI reports written in 2022 from a French quaternary center were retrospectively reviewed. Two radiologists identified MRIs that were done for headaches. Four radiologists scored reports' conclusions as normal or abnormal. Abnormalities were labeled as either headache-generating or incidental. In parallel, Vicuna, an open-source LLM, performed the same tasks. Vicuna's performance metrics were evaluated using the radiologists' consensus as the gold standard.

Results: Among the 2398 reports of the study period, radiologists identified 595 that included headaches in their indication (median age 35, 68% female). A positive finding was reported in 227/595 (38%) cases, 135 of which could explain the headache. Our LLM-based method had sensitivity(Se)/specificity(Sp)(95%CI) of respectively Se=0.98 (583/595)(0.97-0.99)/Sp=0.99 (1791/1803)(0.99-1.00) for detecting the presence of headache in the clinical context, Se=0.99 (514/517)(0.98-1.00)/Sp=0.99 (68/69)(0.92-1.00) for the use of contrast medium injection, Se=0.97 (219/227)(0.93-0.99)/Sp=0.99 (364/368)(0.97-1.00) for study categorization as normal or abnormal and Se=0.88 (120/136)(0.82-0.93)/Sp=0.73 (66/91)(0.62-0.81) for etiological inference between MRI findings and headache.

Conclusion: An open-source LLM can extract information from radiological reports with excellent performances without further training.

Introduction:

Scalable and flexible solutions to harness the potential of large unstructured radiology reports databases are urgently needed (1). The expected “information revolution” (2) secondary to reports structuring has yet to materialize, as large-scale retrospective review of cases of interest still often relies on manual analysis of thousands of free-text reports.

Previous studies have leveraged information contained in the radiology reports to estimate the frequency of positive findings to assess the appropriate use of imaging for various indications (3-6). Assessing the relevance of imaging is particularly pertinent for advanced and expansive exams like MRI and CT in already high-intensity settings, such as on-call shifts. For instance, the use of MRI for emergency patients with headache is frequent in Europe, and increasing in the US (6). The ongoing debates over the reported overuse of imaging for this indication (7) and the potential severity of subtle cases highlight the need for large volume of data to address the actual appropriateness of MRI in this indication. However, the post-hoc identification of cohorts of patients, as well as the extraction of information from the free-text reports still rely on time-consuming human reviews of hundreds of reports.

Readily available methods to automatize information extraction from imaging reports to address this question are lacking. Existing solutions rely either on rule-based methods requiring language- and institution-specific dictionaries, or Bidirectional Encoder Representation from Transformers (BERT)-based models dependent on subsequent fine-tuning (8). While well-suited for pretargeted tasks, these methods usually lack the expected flexibility to match the diversity of radiology reports and intended aims, as tedious specific training or adaptation to the institution data are still necessary (1). Large Language Models (LLMs), owing to their training on extensive natural language databases, exhibit unique adaptability to new tasks in different languages (9). Larger models like GPT4, ChatGPT (OpenAI), and Bard (Google) have demonstrated excellent performances in various medical tasks without requiring fine-tuning (9). However, their cloud-based nature conflicts with the necessity for data confidentiality, leading to debates regarding their compatibility with real-life medical data (10). A recent study on radiology reports had to create fictional reports to circumvent this issue (11), and Fink and colleagues showed that measurements from previous exams could be extracted from CT reports with GPT-4, while calling for replication with privacy-preserving models

(12). In this context, the use of LLMs that are accessible on-premises, such as Vicuna (13), could represent a promising alternative. A recent study has assessed the ability of Vicuna to extract findings from chest X-ray reports from a publicly available database (14). However, to the best of our knowledge, no study has investigated the use of open-source LLMs for information extraction from real-life radiology reports yet. We hypothesize that a task-agnostic on-premises LLM can automate information extraction tasks from real-life radiology reports without further training. Our objective is to assess the performance of such a model, using human review as the gold standard.

Methods:

This retrospective study aimed to evaluate the feasibility of automated information extraction from French free-text radiology reports with an open-source LLM. The data warehouse from which the reports were extracted was approved by the French data protection authority (reference [#2019-103](#)). Use of the data for this specific study was approved by the institutional review board in June 2023 (EDS2307251350).

Data: Pseudonymized reports were obtained from the health data warehouse of our institution. The place of residence of the patient, their name, and the name of the prescribing physician were detected and removed from the reports using the eHOP software (Université de Rennes). Eligible reports were MRI brain scans from patients from the emergency department performed from January to December 2022. Reports were segmented into the following sections 'Clinical context', 'Protocol', 'Results' and 'Conclusion', with Python 3.10. The report text itself was left unaltered. Reports were initially written in French and were not translated for this study. For the reader's comprehension, texts shown in Table 2 and 3 and Supplementary Table 6 were translated in English by the authors and manually modified for de-identification. Reports were initially written by 22 different trainees and 21 different board-certified radiologists, unaware of this study at the time of reporting and free of their reporting style.

Variables extracted: Four clinically pertinent information extraction tasks were defined as follows: 1) presence of headache as a symptom from the clinical context; 2) presence of contrast medium injection from the protocol; 3) classification of the exam as normal or abnormal based on its conclusion; 4) inference of etiology between the findings from the conclusion and the headache presented by the patient. All information was initially reported by the radiologist as free text. Contrast medium injection could not be inferred directly from the text for 7 reports which were excluded for task 2 analysis. Age and sex of the patients were automatically inserted in the reports by the reporting software (IMPAX Ris Qdoc 6.1), subsequently collected with regular expression with Python 3.1.

Ground truth: Ground truth by task was defined as follows: for tasks 1 and 2, rating from two in-training radiologists (AA and BB), with an up to 3 year-experience in neuroradiology. Headache had to be explicitly mentioned in report to be considered positive. Symptoms had to be acute, recently exacerbated, persistent despite analgesia, or atypical for the patient. Tasks 3 and 4, rating from two board-certified neuroradiologists (CC and DD) with respectively 7 and 13 years of experience and the two previously involved trainees (AA and BB). The objective of the study as initially declared to the raters was to determine the frequency of findings responsible for headache in emergency patients who underwent a brain MRI in our center. Raters had only access to the section of the report related to the task. All ratings were done independently, blinded to the rating of their colleagues and from Vicuna's. Discrepancies were settled by consensus of the two senior radiologists. For task 4, rating was as follows: *responsible of headache or not responsible of headache*. Then, the rater could determine whether if they were *certain* or *not certain* about their rating. The causal nature of a finding was left to the appreciation and experience of the raters, as done in previous studies (4,5). No predefined list of pathologies was established in order to account for the specificity of each report, as findings from a single diagnostic group (for example, sinus pathologies), can be considered as headache causing or not, depending on their size, shape, signal, enhancement, adjectives used by the reporting neuroradiologist, and potential changes from previous examinations. *Consensual* exams were defined as reports for which at least 3 radiologists rated the association between the main finding and the headache as *certain*.

Sample size: Sample size was calculated using a previously developed formula (15). To estimate sensitivity and specificity for task 4 of 0.7 (16) with a 95% confidence interval, a precision of 10% and an expected prevalence of etiological findings of 20% (4), the number of reports should be at least 400. With approximately 20% of emergency MRIs performed for headache in our institution, at least 2000 reports had to be screened.

Model: We used the Vicuna 1.3 13B model as available from <https://huggingface.co/lmsys/vicuna-13b-v1.3> (retrieved June 27th). Vicuna is based on the LLAMA model by MetaAI, and fine-tuned on conversations shared by users on ShareGPT (12). We made no alteration to the model. To mitigate variability, temperature was set to 0. Interaction with the model was made through FastChat <https://github.com/lm-sys/FastChat> (retrieved June 27th) (17). We developed a Python script to automate interactions with Vicuna and to control the quality of its output (<https://github.com/manuscript-submission/RadioVicuna>). The model had only access to the same section of the report as the human raters. Final script took free-text reports as input and provided a table as output (**Figure 1**). Model was run on two Quadro RTX 6000 GPUs (Graphics Processing Unit).

Prompting: Prompts were written in English. They were designed to be short and simple to adapt them intuitively for new tasks, and align with empirical best practice for prompting LLMs (18). Vicuna was specifically prompted to segment its answers to improve the transparency of error analysis (**Table 3**). We used few-shot in-context learning: fake contextual examples created manually were provided to the model (19). Increasing numbers of examples were used until saturation of diagnostic performance. Examples were engineered to depict the variability of the sample (positive and negative examples, frequent phrasing used by the radiologists). In a sensitivity analysis, we repeated the same tasks with prompts translated in French (**Supplementary table 3**). For each error displayed in Table 3, the following prompt was given to the model following its initial rating: “Reflect on your answer and detail it”.

Performance evaluation and statistical analysis: We estimated sensitivity, specificity, accuracy and F1-scores against the human-established ground truth with 95% binomial confidence intervals. Inter-rater agreement was measured with Fleiss kappa statistics, and their respective 95% CI. We tested the robustness of the results with different numbers of contextual examples. For task 4, comparison of the model accuracy between the consensual reports and the rest of the reports was performed with a Chi² test. Comparison of performance between prompts in English and French was tested with a McNemar test (R version 4.3.1). A p-value<0.05 was considered statistically significant. All statistical analyses were performed by a statistician not involved in the rating of the reports.

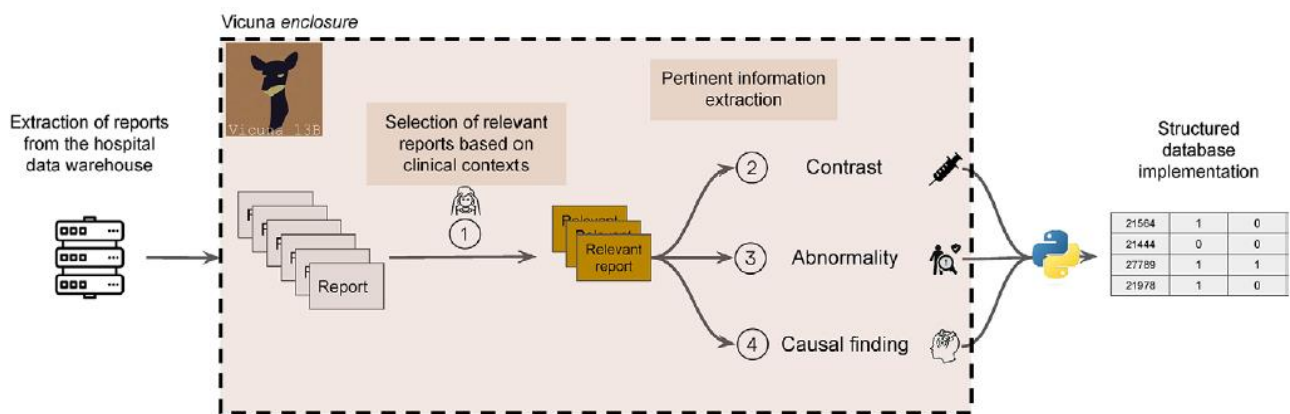


Figure 1: Diagram shows workflow for automated information extraction from pseudonymized radiology reports with Vicuna 13-B (LMSYS Org), an open-source large language model. Four tasks are defined as follows: 1, reports' triaging based on the presence of headache; 2, extraction of contrast medium injection from the protocol; 3, classification of the examination as either normal or abnormal based on the conclusion; and 4, causal inference between the main finding and headache.

Results:

Radiologists' rating

Among the 2398 brain MRIs included in the study, radiologists identified 595 studies performed for headache. The median age of those patients was 35 years (Q1-Q3: 26-51), 68% were women. Of those exams, 517 were performed with contrast medium injection (517/588, 88%). Thirty-eight percent (227/595) included an abnormal finding in the conclusion (inter-rater Fleiss kappa: 0.96 (95%CI:0.95-0.99)). Twenty three percent (136/595) included at least one abnormality which could certainly or probably explain the patient's headaches (inter-rater Fleiss kappa: 0.77 (95%CI:0.72-0.83), **Figure 2** and **Supplementary Figure 1**). Sixteen percent (95/595) reported unrelated findings that were not the cause of headache. Diagnoses are available in **Supplementary Table 1**. The

range of the estimated time spent by each radiologist to review and rate the exams was 3h-3h30 for the 2398 exams of task 1, 1h30-1h45 for the 588 exams of task 2, 1h40-2h20 for task 3 and 4.

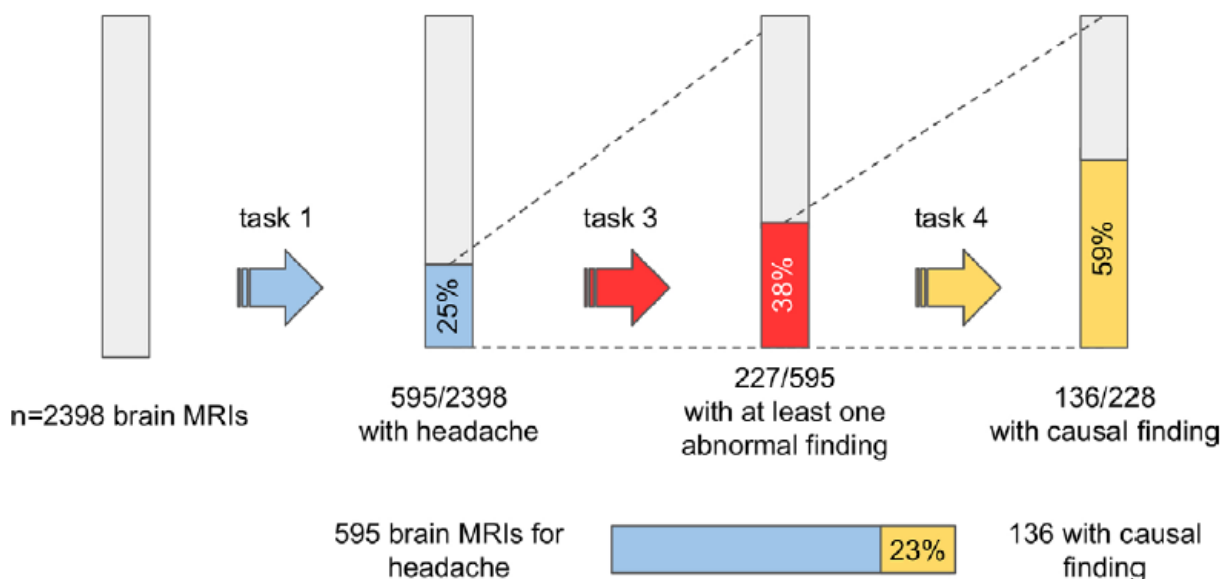


Figure 2: Graph shows diagnostic yield of brain MRI scans in patients in the emergency department with headache.

Vicuna's performance

All performance metrics were >95% for the assessment of the presence of headache as a symptom in the indication (Sensitivity: 98.0%(96.5-99.0), Specificity: 99.3%(98.8-99.7), Accuracy: 99.0%(98.5-99.4)), contrast medium injection assessment (Sensitivity: 99.4%(98.3-99.9), Specificity: 98.6%(92.2-100), Accuracy: 99.3%(98.3-99.8)), and classification of the exam as normal or abnormal based on the report's conclusion (Sensitivity 96.0%(92.5-98.2), Specificity: 98.9%(97.2-99.7), Accuracy: 97.8%(96.3-98.8)). Inference of etiology between the exam main findings and the presence of headache as presenting symptom was accurate in 81.9%(76.4-86.7) of cases (189/227)(Sensitivity 88.2%(81.5-93.1), Specificity 72.5%(62.2-81.4)) (**Table 1** and **Figure 3**). For this task, Vicuna performed significantly better on consensual exams (89.7% accuracy (82.6-84.5) on 116 consensual exams, 73.9 (64.7-81.8) on the remaining 111 reports, p=0,002) (**Supplementary Table 2**). No change in performance metrics was observed when prompting Vicuna in French (**Supplementary Table 3**).

Prompts were engineered in 30min for task 1 and 2 and 1h for task 3 and 4. Calculation time on our machine ranged from 30 min for task 4 (227 reports) to 3h for task 1 (2398 reports)(**Table 1**). Prompts used and examples of answers are provided in Table 2. Contingency matrices are available in **Supplementary Table 4**.

Providing fake contextual examples (few-shot) systematically improved performance over zero-shot. Saturation of accuracy was obtained with 4 to 6 examples, depending on the task (**Supplementary Table 5**). Contextual examples are provided in **Supplementary table 6**.

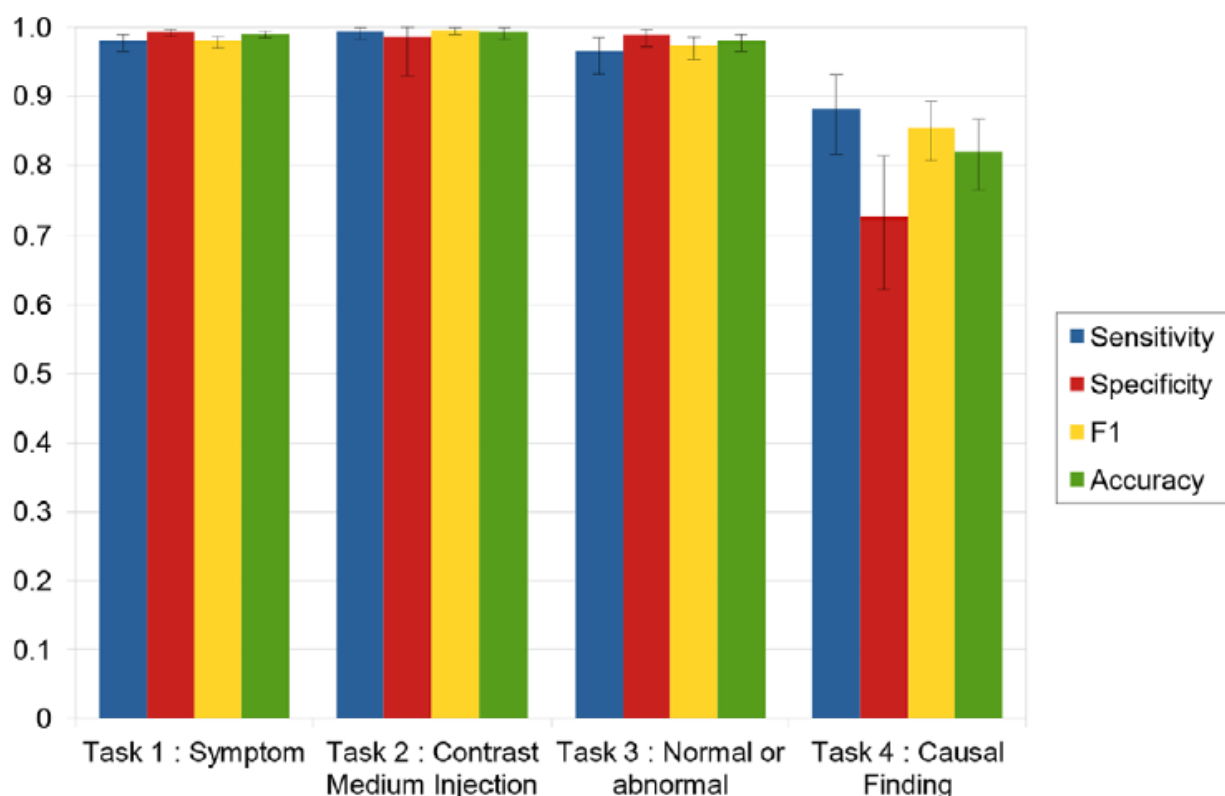


Figure 3: Graph shows performance of Vicuna 13-B (LMSYS Org) for tasks 1, 2, 3, and 4. Error bars are 95% CIs.

Vicuna’s errors analysis

For all task, no systematic error pattern could be identified (**Table 3**). False positives and false negatives included various phrasings, contexts and findings. For task 5, aneurysms were frequently marked as responsible for headache for Vicuna (4 false positives) and none of the two cytotoxic

corpus callosum lesions were marked as positive (false negatives). No other finding was frequently associated with errors.

When prompted to self-analyze its responses, Vicuna provided the user with additional information, uncertainties, or self-corrections, displayed in **Table 3**.

Discussion:

Our study demonstrates that an open-source task-agnostic LLM run on-premises can efficiently extract medical information from real-life radiology reports.

Human manual review of 2398 brain MRI reports for emergency patients with headache revealed that 25% of emergency MRIs were performed for patients with headache, and among those, 23% had significant findings that could explain the headache. A recent meta-analysis reported a rate of causal finding of 9%, significantly lower than ours (20). That review pooled studies with different imaging modalities (MRI and CT), regions, and center types (secondary, tertiary). It is expected that MRI is more sensitive than CT scan to detect subtle abnormalities which could explain the symptoms (cortical vein thrombosis, carotid or vertebral artery dissection, meningitis...)(4). Additionally, our center being the regional reference center for oncology and neurosurgery, it is expected to recruit severe cases, as shown by the listed diagnoses (tumors, infarcts, etc). Our results are in line with a recent study on brain MRIs for non-traumatic headache conducted in a Finnish tertiary center, yielding a 20% rate of causal findings (4).

By implementing an innovative solution, we provided practical insights on the use of LLM for biomedical information extraction. Performances of commercially available LLMs for feature extraction tasks are still largely debated (21,22). Initial studies using GPT-3 on benchmark data reported accuracy below 60% for various name-entity recognition and relation extraction tasks (21). More recent works on ChatGPT and GPT-4 reported higher accuracy on simpler tasks, notably more than 90% precision to retrieve medication names from public datasets, improving with the use of self-verification methods (22). Adams and colleagues used GPT-4 to structure fictional radiology reports with a 100% accuracy (11), and Fink and colleagues successfully extracted measurements from cancer lesions in 96% of reports with the same model (12). Along with these results, we found

that a lesser-powered open-source LLM chatbot can extract a symptom from a clinical context and the presence of contrast medium injection from a MRI protocol with near-perfect accuracy. Recent studies put emphasis on the innate medical knowledge of large language models and their logical abilities (23,24), improved with a low number of contextual examples (19). In line with these results, we show that lesser-powered LLM is able to relate radiological findings and symptoms with an 82% accuracy. When prompted to develop its answers, Vicuna was able to ground its rating in plausible clinical explanations. Even though the causal inference task herein described relies on the initial inference process of listing findings from images, our results offer insights into a new approach for inferring causality in radiology reports (25).

The ability of open-source models to run on-premises has multiple advantages. First, the absence of communication of medical data to private third-party servers ensures critically needed data privacy (10). Second, as opposed to cloud-based methods relying on ever-changing algorithms, reproducibility over time (version control) can be ensured (26). This allows for replicability of results and monitoring model drift during continuous optimization, with the possibility to modify parameters, update or install new models at any time. Third, the training of task-agnostic models such as Vicuna on diverse natural texts distinctively separates their intended use as chatbots from their emerging application on specialized medical texts. This separation significantly mitigates the risk of over-fitting, as the model's training data does not overlap with the specific data it encounters in clinical settings. Fourth, only open source models can allow the users to access their attention process and compute saliency scores, which synthesize to what amount each word of the context supported the model's answer, a critical aspect of LLMs explainability. Future studies addressing LLM explainability for clinical information extraction could greatly benefit from the open nature of on-premises models. Finally, there is no additional cost associated with the use of local models. As recently stated (27,28), the cost per token generated for commercial method scaling with the volume of text to analyze may hamper their large-scale diffusion. Local models can run on commercially available machines (20GB VRAM for vicuna 13B) with no scaling cost.

As the method we propose is based on a publicly available, task-agnostic model that was not altered for this study, including no training or fine-tuning, it is readily available for new tasks, in different centers and languages. New tasks can be defined by writing a prompt in natural language alongside

a few illustrative examples. Synonyms, negations and circumlocutions are handled directly by the model via its initial training, with no need for additional post-processing, as opposed to text-searching methods. The code for our scripts is made available and a practical guide is provided (<https://github.com/manuscript-submission/RadioVicuna>). As the model we used is inherently task-agnostic, we hypothesize that our method could also be applied to other medical texts beyond radiology reports (29,30). Similar work, limited to ChatGPT, has shown promising results on information extraction from clinical texts (31)

With the advent of LLMs in radiology, a new and intuitive approach for deciphering the complex information contained in radiological reports is emerging, paving the way for applications in clinical settings. In the future, LLMs will likely serve as powerful tools to (i) radiologists, by extracting information from previous exams to facilitate follow-up (12), drafting reports to improve workflow (32) or checking for readability and completeness of their report, (ii) clinicians, by annotating or structuring the report to improve its readability, and (iii) patient, by providing an assistant able to digest and explain the exam and answer related questions real time (33). While recent studies have highlighted the diagnostic abilities of LLMs in the field of radiology (16), LLMs may be integrated sooner in the workflow of radiologists as assistants for low-expertise tasks. This integration is expected to enhance and promote the radiologists' expertise rather than replace it, relieve them of time-consuming tasks, and facilitate more effective communication of information in their reports. Moreover, from a research perspective, LLMs could significantly contribute to addressing the scarcity of high-quality labeled data on which to train computer vision models. By leveraging the expert information already present in the report, LLMs could enable a substantial increase in the volume of labeled data for training computer vision models, with a limited cost.

AI explainability is a critical aspect of its practical implementation (33). LLMs being conversational agents, it is possible to ask for additional explanations for every rating, ensuring that their answers are based on plausible medical knowledge and good understanding of the task. Other methods exist to assess the level of certainty for ratings LLMs, including displaying the probability of the output token, with a higher probability proxying for a higher confidence level. Accessing this confidence level is possible with our proposed method, and is explained in the practical guide we provide. Future work will be needed to implement other aspects of LLMs explainability to our workflow (33).

This study has some limitations. First, due to its single center characteristic, we could not test for variability in center writing styles and languages. However, our center is particularly large (165.000 emergency visits annually) and our dataset is diverse with 43 different writers. As our prompts were in English and reports were in French, we anticipate high robustness to language variations, as also evidenced by similar performances after translating the prompts from English to French. The dominant language on which the model has been trained being English, one should expect even higher performance for reports written in English, and consistent results across various languages (34). Second, the available clinical information was restricted to brief contexts from the report itself. Further studies should extend the scope of this method to data extraction from unsegmented clinical reports. Third, the gold-standard for the most complex task of etiological inference between a finding and a symptom is in itself subjective, based on the experience from the radiologists. Third, the most complex task of inferring the etiological nature of findings from the reports' impressions is a difficult and subjective task. As in previous studies, ground truth was defined by expert consensus. The difficulty of this task, as shown by the high rate of non-consensual exams in our studies (49%) may originate from the frequent lack of explicit comment by the reporting neuroradiologist on the significance of findings. LLMs have already been employed to detail radiology reports for patients (33), and future work is needed to improve the conveyance of radiology reports in an emergency setting. Finally, we used a model that will potentially be outperformed by new developments, especially with the recent release of LLAMA 2 (35). However, this workflow is not limited to a specific model or set of models, allowing for flexibility with newer technologies. Further work is needed to study the respective strengths and limits of newer available models.

This proof-of-concept study shows the potential of open-source LLMs to perform information extraction tasks from real-life free-text radiology reports, without the need for additional training. Given the rising social and scientific need for transparency in AI, the open-source framework we provide creates a stepping stone for replication studies and may open a new horizon for automated analysis of unstructured medical texts.

References:

1. Pons E, Braun LMM, Hunink MGM, Kors JA. Natural Language Processing in Radiology: A Systematic Review. *Radiology*. Radiological Society of North America; 2016;279(2):329–343. doi: [10.1148/radiol.16142770](https://doi.org/10.1148/radiol.16142770).
2. Langlotz CP. Automatic Structuring of Radiology Reports: Harbinger of a Second Information Revolution in Radiology. *Radiology*. Radiological Society of North America; 2002;224(1):5–7. doi: [10.1148/radiol.2241020415](https://doi.org/10.1148/radiol.2241020415).
3. Lungren MP, Amrhein TJ, Paxton BE, et al. Physician Self-Referral: Frequency of Negative Findings at MR Imaging of the Knee as a Marker of Appropriate Utilization. *Radiology*. Radiological Society of North America; 2013;269(3):810–815. doi: [10.1148/radiol.13130281](https://doi.org/10.1148/radiol.13130281).
4. Happonen T, Nyman M, Ylikotila P, Merisaari H, Mattila K, Hirvonen J. Diagnostic yield of emergency MRI in non-traumatic headache. *Neuroradiology*. 2023;65(1):89–96. doi: [10.1007/s00234-022-03044-2](https://doi.org/10.1007/s00234-022-03044-2).
5. Budweg J, Sprenger T, De Vere-Tyndall A, Hagenkord A, Stippich C, Berger CT. Factors associated with significant MRI findings in medical walk-in patients with acute headache. *Swiss Med Wkly*. 2016;146:w14349. doi: [10.4414/smw.2016.14349](https://doi.org/10.4414/smw.2016.14349).
6. Trofimova AV, Duszak Jr. R, Kadom N, Sadigh G. Increasing and disparate use of neuroimaging for adults and children with non-traumatic headaches in the US emergency departments: Opportunities for improvement. *Headache: The Journal of Head and Face Pain*. 2021;61(1):179–189. doi: [10.1111/head.14020](https://doi.org/10.1111/head.14020).
7. Dunne CL, Elzinga JL, Vorobeichik A, et al. A Systematic Review of Interventions to Reduce Computed Tomography Usage in the Emergency Department. *Annals of Emergency Medicine*. 2022;80(6):548–560. doi: [10.1016/j.annemergmed.2022.06.001](https://doi.org/10.1016/j.annemergmed.2022.06.001).
8. Jupin-Delevaux É, Djahnine A, Talbot F, et al. BERT-based natural language processing analysis of French CT reports: Application to the measurement of the positivity rate for pulmonary embolism. *Research in Diagnostic and Interventional Imaging*. 2023;6:100027. doi: [10.1016/j.redii.2023.100027](https://doi.org/10.1016/j.redii.2023.100027).
9. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *New England Journal of Medicine*. Massachusetts Medical Society; 2023; doi: [10.1056/NEJMSr2214184](https://doi.org/10.1056/NEJMSr2214184).
10. Marks M, Haupt CE. AI Chatbots, Health Privacy, and Challenges to HIPAA Compliance. *JAMA*. 2023; doi: [10.1001/jama.2023.9458](https://doi.org/10.1001/jama.2023.9458).
11. Adams LC, Truhn D, Busch F, et al. Leveraging GPT-4 for Post Hoc Transformation of Free-Text Radiology Reports into Structured Reporting: A Multilingual Feasibility Study. *Radiology*. 2023;230725. doi: [10.1148/radiol.230725](https://doi.org/10.1148/radiol.230725).
12. Fink MA, Bischoff A, Fink CA, et al. Potential of ChatGPT and GPT-4 for Data Mining of Free-Text CT Reports on Lung Cancer. *Radiology*. Radiological Society of North America; 2023;308(3):e231362. doi: [10.1148/radiol.231362](https://doi.org/10.1148/radiol.231362).
13. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality | LMSYS Org. . <https://lmsys.org/blog/2023-03-30-vicuna>. Accessed July 9, 2023.
14. Mukherjee P, Hou B, Lanfredi RB, Summers RM. Feasibility of Using the Privacy-preserving Large Language Model Vicuna for Labeling Radiology Reports. *Radiology*. Radiological Society of North America; 2023;309(1):e231147. doi: [10.1148/radiol.231147](https://doi.org/10.1148/radiol.231147).

15. Buderer NMF. Statistical Methodology: I. Incorporating the Prevalence of Disease into the Sample Size Calculation for Sensitivity and Specificity. *Academic Emergency Medicine*. 1996;3(9):895–900. doi: 10.1111/j.1553-2712.1996.tb03538.x.
16. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations. *Radiology*. Radiological Society of North America; 2023;307(5):e230582. doi: 10.1148/radiol.230582.
17. Zheng L, Chiang W-L, Sheng Y, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv*; 2023. doi: 10.48550/arXiv.2306.05685.
18. Weng L. Prompt Engineering. 2023. <https://lilianweng.github.io/posts/2023-03-15-prompt-engineering/>. Accessed November 21, 2023.
19. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large Language Models are Few-Shot Clinical Information Extractors. *arXiv*; 2022. doi:10.48550/arXiv.2205.12689
20. Jang YE, Cho EY, Choi HY, Kim SM, Park HY. Diagnostic Neuroimaging in Headache Patients: A Systematic Review and Meta-Analysis. *Psychiatry Investig*. 2019;16(6):407–417. doi: 10.30773/pi.2019.04.11.
21. Gutiérrez BJ, McNeal N, Washington C, et al. Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again. *arXiv*; 2022. doi: 10.48550/arXiv.2203.08410
22. Gero Z, Singh C, Cheng H, et al. Self-Verification Improves Few-Shot Clinical Information Extraction. *arXiv*; 2023. doi: 10.48550/arXiv.2306.00024
23. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. Nature Publishing Group; 2023;1–9. doi: 10.1038/s41586-023-06291-2.
24. Webb T, Holyoak KJ, Lu H. Emergent analogical reasoning in large language models. *Nat Hum Behav*. Nature Publishing Group; 2023;1–16. doi: 10.1038/s41562-023-01659-w.
25. Sebro RA, Kahn CE Jr. Automated detection of causal relationships among diseases and imaging findings in textual radiology reports. *Journal of the American Medical Informatics Association*. 2023;30(10):1701–1706. doi: [10.1093/jamia/ocad119](https://doi.org/10.1093/jamia/ocad119).
26. Spirling A. Why open-source generative AI models are an ethical way forward for science. *Nature*. 2023;616(7957):413–413. doi: 10.1038/d41586-023-01295-4.
27. Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. *The Lancet Digital Health*. Elsevier; 2023;5(6):e333–e335. doi: 10.1016/S2589-7500(23)00083-3.
28. López-Úbeda P, Martín-Noguerol T, Luna A. Radiology in the era of large language models: the near and the dark side of the moon. *Eur Radiol*. 2023; doi: 10.1007/s00330-023-09901-9.
29. Kulkarni PA, Singh H. Artificial Intelligence in Clinical Diagnosis: Opportunities, Challenges, and Hype. *JAMA*. 2023;330(4):317. doi: 10.1001/jama.2023.11440.
30. Uddin Y, Nair A, Shariq S, Hannan SH. Transforming primary healthcare through natural language processing and big data analytics. *BMJ*. British Medical Journal Publishing Group; 2023;381:p948. doi: 10.1136/bmj.p948.
31. Jethani N, Jones S, Genes N, et al. Evaluating ChatGPT in Information Extraction: A Case Study of Extracting Cognitive Exam Dates and Scores. *medRxiv*; 2023. p. 2023.07.10.23292373. doi: 10.1101/2023.07.10.23292373.

32. Langlotz CP. The Future of AI and Informatics in Radiology: 10 Predictions. Radiology. Radiological Society of North America; 2023;309(1):e231114. doi: [10.1148/radiol.231114](https://doi.org/10.1148/radiol.231114).
33. Amin KS, Davis MA, Doshi R, Haims AH, Khosla P, Forman HP. Accuracy of ChatGPT, Google Bard, and Microsoft Bing for Simplifying Radiology Reports. Radiology. Radiological Society of North America; 2023;309(2):e232561. doi: [10.1148/radiol.232561](https://doi.org/10.1148/radiol.232561).
34. Viet Dac Lai, Nghia Trung Ngo, Ami Pourn Ben Veyseh, et al. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning, arXiv 2023, doi:10.48550/arXiv.2304.05613
35. Llama 2 - Meta AI. . <https://ai.meta.com/llama/>. Accessed July 31, 2023.

Tables

Table 1: Detailed Vicuna performances and calculation time for each extraction task.

	Task 1: Symptom n=2398	Task 2: contrast medium injection n=588	Task 3: Normal or abnormal classification n= 595	Task 4: Finding- symptom etiological inference n= 227
Vicuna Sensitivity	98.0 (96.5-99.0) 583/595	99.4 (98.3-99.9) 514/517	96.5 (93.2-98.5) 219/227	88.2 (81.6-93.1) 120/136
Vicuna Specificity	99.3 (98.8-99.7) 1791/1803	98.6 (92.2-100) 68/69	98.9 (97.2-99.7) 364/368	72.5 (62.2-81.4) 66/91
Vicuna F1	98.0 (97.0-98.7) 1166/1190	99.6 (99.0-99.9) 1028/1032	97.3 (95.4-98.6) 438/450	85.4 (80.7-89.3) 240/281
Vicuna Accuracy	99.0 (98.5-99.4) 2374/2398	99.3 (98.3-99.8) 584/588	98.0 (96.5-99.0) 583/595	81.9 (76.4-86.7) 186/227
Calculation time for all reports (on two Quadro RTX 6000 GPUs)	3h	45min	1h	30min
Prompt engineering	30min	30min	45min	1h

Table 2: Extracts from dialogues with Vicuna

Task 1: Symptom of interest (headache)		
Prompt	Doctor	<p>I will present you short indications for radiology exams. Your task is to list all the symptoms from them and tell me if they correspond to headache or to another symptom. You will answer only according to this template: Symptoms from INDICATION:</p> <ul style="list-style-type: none"> - List all the symptoms /is it headache or another symptom ? (If you find no symptom, answer "N/A")
Example 1	Doctor	INDICATION: Headaches and progressively worsening vertigo for the past 2 days, associated with a sensation of hypoesthesia on the right side of the body. Suspicion of cerebral venous thrombosis.
	Robot	<p>Symptoms from INDICATION:</p> <ul style="list-style-type: none"> - Headaches /headache - Progressively worsening vertigo /other - Sensation of hypoesthesia on the right side of the body /other
Example 2	Doctor	INDICATION: Suspicion of cerebral venous thrombosis.
	Robot	Symptoms from INDICATION: - N/A
Example 3	Doctor	INDICATION: Investigation for intracranial hypertension in a 27-year-old female patient who presented with intermittent visual blur and sudden loss of visual acuity. Associated with bilateral and symmetric papilledema revealed on fundoscopy.
	Robot	<p>Symptoms from INDICATION:</p> <ul style="list-style-type: none"> - Intermittent visual blur /other - Sudden loss of visual acuity /other - Bilateral and symmetric papilledema revealed on fundoscopy /other
Task 2: Contrast medium injection		
Prompt	Doctor	<p>I will present you protocolos from radiology reports. I want to you to tell me if contrast medium injection was performed (often with gadolinium or gado). Ignore all other information, they are insignificant. One crucial rule is that respond only according to this template:</p> <ul style="list-style-type: none"> - Words mentioning injection: [retrieve the words mentioning injection or gadolinium or gado in the text] - Injected ?: [Yes/No]

Case 1	Doctor	ID:98678062.0 Axial FLAIR, T2*, SWI, and Diffusion sequences. ADC mapping. Axial 3D-SET1 BV sequence on the cervical carotid and vertebral arteries. Gado-MR with 3D reconstructions. / Ms - born: [JJ]/[MM]/1970 (Age: 51)
	Robot	- Words mentioning injection: Gado-MR with 3D reconstructions . - Injected ?: Yes
Case 2	Doctor	ID:13009397.0 Axial Diffusion and FLAIR sequences.Volumetric SWI and 3D TOF sequences. / Mr born: [JJ]/[MM]/1988 (Age: 33)
	Robot	- Words mentioning injection: None - Injected ?: No
Task 3: Classification of the exam as normal or abnormal		
Prompt	Doctor	I will present you conclusions from radiology exams. Your task is to list all the findings from them and tell me if they correspond to normal or abnormal findings. Keep in mind, an abnormal finding may be described as stable or regressing, so please be sure not to miss any abnormal finding. Each conclusion is independant and should not influence your answer. You will answer only according to this template: List of findings: - List all the findings /normal or /abnormal finding ?
Case 1	Doctor	Conclusion: No evidence of recent hemorrhagic or ischemic lesions. No arterial occlusion on the Willis polygon. Known right deep hematoma sequela without significant changes.
	Robot	List of findings: - No evidence of recent hemorrhagic or ischemic lesions /normal - No arterial occlusion on the Willis polygon /normal - Known right deep hematoma sequela without significant changes /abnormal
Case 2	Doctor	Conclusion: No evidence supporting cervical arterial dissection, RCVS, or cerebral venous thrombosis.
	Robot	List of findings: - No evidence supporting cervical arterial dissection /normal - No evidence supporting RCVS /normal - No evidence of cerebral venous thrombosis /normal

Task 4: Inference of etiology between a finding and a symptom (headache)

Prompt	Doctor	<p>I will present you abnormal findings from radiology exams. Your task is to list all the findings from them and tell me if they usually cause headache or not. Each conclusion is independent and should not influence your answer. You will answer only according to this template:</p> <p>List of findings:</p> <ul style="list-style-type: none"> - List all the findings /can cause headache or not ?
Case 1	Doctor	<p>Conclusion: No evidence of recent hemorrhagic or ischemic lesions. No arterial occlusion on the Willis polygon. Known right deep hematoma sequela without significant changes.</p>
	Robot	<p>List of findings:</p> <ul style="list-style-type: none"> - No evidence of recent hemorrhagic or ischemic lesions / does not usually cause headache - No arterial occlusion on the Willis polygon / does not usually cause headache - Known right deep hematoma sequela without significant changes / does not usually cause headache
Case 2	Doctor	<p>Conclusion: Presence of 4 nodular cerebral lesions with ring enhancement suggestive of secondary lesions. A left parieto-occipital parenchymal hematoma probably secondary to a hemorrhagic metastatic lesion.</p>
	Robot	<p>List of findings:</p> <ul style="list-style-type: none"> - Presence of 4 nodular cerebral lesions with ring enhancement suggestive of secondary lesions /can cause headache - A left parieto-occipital parenchymal hematoma probably secondary to a hemorrhagic metastatic lesion /can cause headache
Case 3	Doctor	<p>Conclusion: No evidence of recent ischemic or hemorrhagic lesions. No MRI findings supporting cervical arterial dissection. Incidental discovery of a 2.5mm intracranial arterial aneurysm image on the left sylvian region, requiring further evaluation by a specialist.</p>
	Robot	<p>List of findings:</p> <ul style="list-style-type: none"> - No evidence of recent ischemic or hemorrhagic lesions / does not usually cause headache - No MRI findings supporting cervical arterial dissection / does not usually cause headache - Incidental discovery of a 2.5mm intracranial arterial aneurysm image on the right sylvian region, requiring further evaluation by a specialist

		/ does not usually cause headache
Case 4	Doctor	Conclusion: Infiltrating lesion with contrast enhancement in the left parietal region, extending to the left insula and crossing the midline to the right occipital region, spanning approximately 6 cm, compatible with an intra-cerebral lymphoma. Mass effect with early anterior subfalcial herniation observed. Two recent ischemic spots detected (left insular and left cerebellar regions). No proximal occlusion of the arteries in the Willis polygon. A neurosurgical evaluation is recommended.
	Robot	List of findings: - Infiltrating lesion with contrast enhancement in the left parietal region, extending to the left insula and crossing the midline to the right occipital region, spanning approximately 6 cm, compatible with an intra-cerebral lymphoma /can cause headache - Mass effect with early anterior subfalcial herniation observed /can cause headache - Two recent ischemic spots detected (left insular and left cerebellar regions) /can cause headache - No proximal occlusion of the arteries in the Willis polygon / does not usually cause headache - A neurosurgical evaluation is recommended / does not usually cause headache
Case 5	Doctor	Conclusion: Vascular leukoencephalopathy, chronic microbleeds, and a right thalamic lacuna suggestive of small vessel disease.
	Robot	List of findings: - Vascular leukoencephalopathy / does not usually cause headache - Chronic microbleeds / does not usually cause headache - Right thalamic lacuna / does not usually cause headache

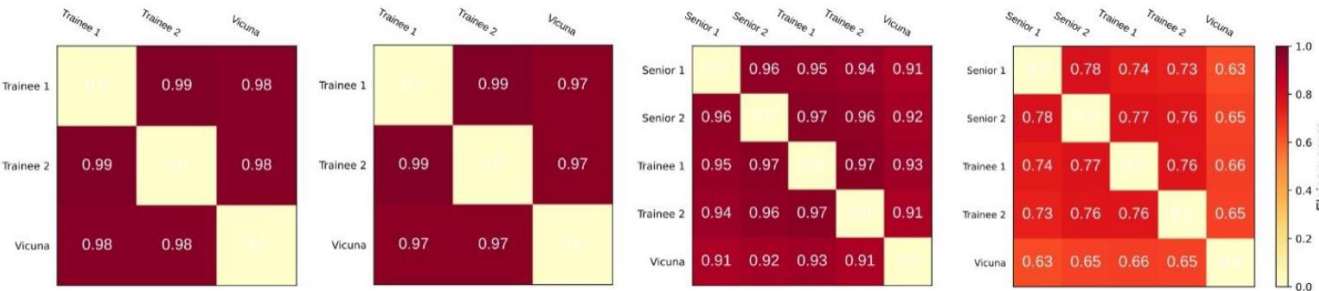
Table 3: Examples of errors by task and self-explanations from Vicuna (*extraits*)

Task	Error	Input	Output	Vicuna self-explanation
1	False Positives	Episode of visual blurring with fluctuations in blood pressure (as observed by ambulatory blood pressure monitoring) in a patient with a history of migraine. Suspicion of TIA.	Symptoms from INDICATION: - Visual blurring /other - Fluctuations in blood pressure /other - Migraine /headache - Suspicion of TIA /other	The patient has an episode of visual blurring with fluctuations in blood pressure, which could be a symptom of a transient ischemic attack (TIA). However, the patient also has a history of migraine, which could also explain the visual blurring and fluctuations in blood pressure.
	False Negatives	Progressive onset of decreased visual acuity, orbital pain, and photophobia in a 68-year-old female patient with an evolving glioblastoma that was operated on in February 2017 and currently undergoing treatment with radiotherapy and chemotherapy.	Symptoms from INDICATION: - Decreased visual acuity /other - Orbital pain /other - Photophobia /other	I apologize for the confusion in my previous response. Upon reflection, I realize that I did not accurately identify the symptoms in the given indication. Here is the corrected response: Symptoms from Indication : * Decreased visual acuity /other * Orbital pain /other * Photophobia /other * Glioblastoma (evolving) /other * Radiotherapy and chemotherapy (current treatment) /other Please note that the symptoms listed are not headaches, but rather other symptoms related to the patient's condition.
2	False Positives	Axial T1, SWI, FLAIR, and Diffusion sequences. Coronal T2 Phase-Contrast Venous sequence.	- Words mentioning injection: Phase-Contrast Venous sequence - Injected: Yes	My apologies, I misunderstood the task. I thought the task was to identify if contrast media injection was mentioned in the text, not if it was performed. I will provide a revised response using the template provided: 'Doctor', "PROTOCOLE : Axial T1, SWI, FLAIR, and Diffusion sequences. Coronal T2 Phase-Contrast Venous sequence.", 'Robot', - Mention of injection : Phase-Contrast Venous sequence - Injected? : No I hope this revised response is more accurate. Please let me know if you have any further questions or if there is anything else I can assist you with.
	False Negatives	Cerebral exploration with axial Diffusion, T1, T2*, and SWI sequences, followed by 3D FLAIR. After Gadolinium injection, Venous MR Angiography (MRA), 3D T1-weighted Enhanced Gradient (EG), and axial T1 sequences are performed. Coronal T2 Dixon slices are obtained on the anterior visual pathways.	- Words mentioning injection: None - Injected: No	My apologies, I made an error in my previous response. Upon further analysis, there is no mention of injection in the provided protocol. Therefore, it can be concluded that no contrast media injection was performed.
3	False Positives	No evidence of recent hemorrhagic or ischemic lesions. No intracranial mass effect. No cerebral venous thrombosis. Thin appearance of the right transverse sinus.	List of findings: - No evidence of recent hemorrhagic or ischemic lesions / normal - No intracranial mass effect / normal - No cerebral venous thrombosis / normal - Thin appearance of the right transverse sinus / abnormal	The conclusion states that there is no evidence of recent hemorrhagic or ischemic lesions, no intracranial mass effect, and no cerebral venous thrombosis. Additionally, the right transverse sinus appears to be thin, which could be a cause for concern. It is important to note that the thin appearance of the sinus may be a normal variation or may be related to other factors, so further investigation may be necessary
	False	No recent ischemic or hemorrhagic lesions.	List of findings:	In the last response, the conclusion was that there were no

	Positives	Permeability of the right temporosylvian bypass.	- No recent ischemic or hemorrhagic lesions / normal - Permeability of the right temporosylvian bypass / normal	recent ischemic or hemorrhagic lesions, and that the right temporosylvian bypass showed permeability. This means that there were no recent instances of tissue death or bleeding in the brain, which is a normal finding. Additionally, the right temporosylvian bypass, which is a surgical procedure used to improve blood flow to the brain, is functioning properly and allowing for adequate blood flow.
	False Positives	Vascular leukoencephalopathy. Chronic microbleeding. Right thalamic ischemic sequelae.	List of findings: - Vascular leukoencephalopathy / can cause headache - Chronic microbleeding / can cause headache - Right thalamic lacune / does not usually cause headache	* Vascular leukoencephalopathy: This is a condition characterized by damage to the blood-brain barrier, which can lead to the accumulation of water and white matter changes in the brain. It can be caused by a variety of factors, including hypertension, diabetes, and infections. Vascular leukoencephalopathy can cause headache, but it is not a common symptom. * Chronic microbleeding: This refers to small areas of bleeding in the brain that occur over a long period of time. It can be caused by a variety of factors, including hypertension, anticoagulation, and amyloid angiopathy. Chronic microbleeding can cause headache, but it is not a common symptom. * Right thalamic lacune: This is a small area of damage or degeneration in the thalamus, which is a part of the brain that is involved in a variety of functions, including sensory processing, motor control, and cognition. Thalamic lacunae can be caused by a variety of factors, including aging, hypertension, and trauma. Right thalamic lacunae do not usually cause headache.
	False Positives	No visible intramural hematoma within cervical vessels suggestive of dissection. No cerebral venous thrombosis or pericerebral collection. Appearance of diffuse pachymeningeal enhancement and discreetly convex aspect of cervical venous sinuses and epidural venous plexuses. May be consistent with signs of cerebrospinal fluid hypotension.	List of findings: - No visible intramural hematoma within cervical vessels suggestive of dissection / does not usually cause headache - No cerebral venous thrombosis or pericerebral collection / does not usually cause headache - Appearance of diffuse pachymeningeal enhancement and discreetly convex aspect of cervical venous sinuses and epidural venous plexuses / does not usually cause headache - May be consistent with signs of cerebrospinal fluid hypotension / does not usually cause headache	* No visible intramural hematoma within cervical vessels suggestive of dissection: This finding indicates that there is no visible blood clot within the walls of the cervical arteries, which could suggest a dissection. A dissection occurs when there is a tear in the inner lining of an artery, which can cause bleeding and can be life-threatening. This finding is usually a good sign, as it means that there is no active dissection happening. * No cerebral venous thrombosis or pericerebral collection: This finding means that there is no evidence of blood clots in the veins that drain blood from the brain or any collection of blood around the brain. Cerebral venous thrombosis is a rare but serious condition that can cause stroke-like symptoms and can be life-threatening. * Appearance of diffuse pachymeningeal enhancement and discreetly convex aspect of cervical venous sinuses and epidural venous plexuses: This finding indicates that there is increased thickness of the membranes surrounding the brain and spinal cord, as well as a slight bulging of the veins that drain blood from the brain and the epidural veins. This could be a sign of cerebrospinal fluid hypotension, which occurs when there is a decrease in the pressure of the cerebrospinal fluid that surrounds the brain and spinal cord. This can be caused by a variety of factors, including dehydration, blood loss, or certain medications. * May be consistent with signs of cerebrospinal fluid hypotension: This conclusion means that the findings mentioned above could be consistent with a decrease in the pressure of the cerebrospinal fluid. This is not a definitive diagnosis, but it is something to consider given the findings. Cerebrospinal fluid hypotension can cause symptoms such as headache, nausea, and vomiting, and it is important to address this condition if it is present.

Supplementary material

Supplementary figure



Supplementary Figure 1: Binary agreement matrix between pairs of raters for tasks 1, 2, 3 and 4. Numbers are Fleiss kappa coefficients.

Supplementary tables

Supplementary Table 1: Detailed MRI abnormalities and etiological inference, according to radiologists' consensus and Vicuna

Anomaly	Number of cases
Vascular	
Cerebral Infarct	18
Hematoma	10
Dissection	13
Subarachnoid hemorrhage	6
Aneurysm	9
Cerebral venous thrombosis	4
Internal carotid artery stenosis	4
Reversible cerebral vasoconstriction syndrome	3
Stroke sequelae	2
Cerebral venous thrombosis sequelae	3
Vascular leukoencephalopathy	4
Microbleeds	1
Chronic lobar hematoma sequela	1
Infectious/Inflammatory	
Inflammatory sinus	20
Neuritis	4
Inflammatory white matter lesions	11
Meningitis	3
Granulomatosis	1
Central nervous system infection	1
Tolosa-Hunt syndrome	1
Tumor	
Central nervous system tumor	25
Tumor progression	4
Meningioma	4
Arachnoid cyst	6
Pituitary gland cyst	2
Radionecrosis	1
Pituitary adenoma necrosis	1
Parotid gland tumor	1
Pineal gland cyst	1

Lipoma	1
Malformative	
Chiari malformation	3
Cavernoma	3
Developmental venous anomaly	2
Arterio-venous malformation	1
Non-specified abnormalities in size/shape of venous sinuses	6
Heterotopia	2
Other	
Idiopathic intracranial hypertension signs	16
Migraine aura	6
Epilepsy	2
Intracranial hypotension signs	8
Non-specific white matter hyperintensities	3
Posterior reversible encephalopathy syndrome	2
Fahr syndrome	1
Gayet-Wernicke encephalopathy	1
Cytotoxic lesion of the corpus callosum	2
Degenerated cervical disc	1
Meningeal calcification	1
Non-specified post-operative changes	2
Total	227

Supplementary Table 2: Performance of the model for task 4 on easy-to-asses exams.

	Consensual (n=116)	Non-consensual (n=111)	p-value*
Sensitivity	94.9 (97.4-98.6)	79.3 (66.7-88.8)	<0.001
Specificity	79.0 (62.7-90.5)	67.9 (53.7-80.1)	0.25
F1	90.8 (84.4-95.1)	76.0 (67.4-83.3)	<0.001
Accuracy	89.7 (82.6-94.5)	73.9 (64.7-81.8)	0.002

*: using a Chi² test

Supplementary Table 3: Extraction performances of Vicuna, according to the task and prompting language.

	Task 1:		Task 2:		Task 3:		Task 4:	
	En	Fr	En	Fr	En	Fr	En	Fr
Sensitivity	98.0 (96.5-99.0) 583/595	98.0 (96.5-99.0) 583/595	99.4 (98.3-99.9) 514/517	100 (99.2-100) 517/517	96.5 (93.2-98.5) 219/227	96.5 (93.2-98.5) 219/227	88.2 (81.6-93.1) 120/136	86.8 (79.9-92.0) 118/136
Specificity	99.3 (98.8-99.7) 1791/1803	98.8 (98.2-99.3) 1782/1803	98.6 (92.2-100) 68/69	95.7 (87.2-99.1) 66/69	98.9 (97.2-99.7) 364/368	98.1 (96.1-99.2) 361/368	72.5 (62.2-81.4) 66/91	76.9 (66.9-85.11) 70/91
F1	98.0 (97.0-98.7) 1166/1190	97.3 (96.2-98.1) 1166/1199	99.6 (99.0-99.9) 1028/1032	99.8 (99.3-100) 1034/1036	97.3 (95.4-98.6) 438/450	96.7 (94.6-98.1) 438/453	85.4 (80.7-89.3) 240/281	85.8 (81.1-89.7) 236/275
Accuracy	99.0 (98.5-99.4) 2374/2398	98.6 (99.1-99.1) 2365/2398	99.3 (98.3-99.8) 584/588	99.5 (98.5-99.9) 585/588	98.0 (96.5-99.0) 583/595	97.5 (95.9-98.6) 580/595	81.9 (76.4-86.7) 186/227	82.8 (77.3-87.5) 188(227
P-value*	0.64		0.65		0.18		0.71	

*p-value of the McNemar test performed between the contingency tables for prompts in English (En) and in French (Fr)

Supplementary Table 4: Contingency tables of Vicuna’s rating according to the consensus of the radiologists.

Task 1: symptom extraction (n=2398)		Task 2: extraction of contrast agent injection (n=588)	
TP = 584	FP = 11	TP = 514	FP = 1
FN = 13	TN = 1790	FN = 3	TN = 68
Task 3: classification of the exam as normal or abnormal (n=595)		Task 4: inference between findings and a symptom (n=227)	
TP = 219	FP = 4	TP = 120	FP = 25
FN = 8	TN = 364	FN = 16	TN = 66

TP: True Positives ; FP = False Positives ; FN: False Negatives;TN: True Negatives

Supplementary Table 5: Performance of the model depending on the number of contextual examples provided

Task 1: Symptoms					
Number of examples	0	2	4	6	8
Sensitivity	93.3 (91.0-95.2)	95.1 (93.1-96.7)	97.3 (95.7-98.5)	98.0 (96.5-99.0)	98.3 (96.9-99.2)
Specificity	58.0 (55.7-60.3)	84.8 (83.1-86.4)	96.8 (95.9-97.6)	99.3 (98.8-99.7)	98.9 (98.3-99.3)
F1	58.2 (56.0-60.4)	78.9 (76.7-81.0)	94.0 (82.5-95.3)	98.0 (97.0-98.7)	97.5 (96.5-98.3)
Accuracy	66.8 (64.8-68.7)	87.4 (86.0-88.7)	96.9 (96.1-97.6)	99.0 (98.5-99.4)	98.8 (98.2-99.2)
Task 2: Contrast medium injection					
Sensitivity	90.2 (87.3-92.6)	98.3 (96.7-99.2)	99.4 (98.3-99.9)	99.0 (97.8-99.7)	99.2 (98.0-99.8)
Specificity	100 (94.8-100)	98.6 (92.2-100)	98.6 (92.2-100)	98.6 (92.2-100)	91.6 (82.5-96.8)
F1	94.8 (93.3-96.1)	99.0 (98.2-99.5)	99.6 (99.0-99.9)	99.4 (98.7-99.8)	99.0 (98.2-99.5)
Accuracy	91.3 (88.8-93.5)	98.3 (96.9-99.2)	99.3 (98.3-99.8)	98.8 (97.8-99.6)	98.3 (96.9-99.2)
Task 3: Normal or abnormal					
Sensitivity	73.1 (66.9-78.8)	95.2 (91.5-97.6)	96.0 (92.6-98.2)	96.5 (93.2-98.5)	94.0 (90.2-96.7)
Specificity	95.4 (92.7-97.3)	96.7 (94.4-98.3)	98.9 (97.2-99.7)	98.9 (97.2-99.1)	98.3 (96.4-99.4)
F1	81.0 (76.8-84.7)	94.6 (92.5-96.8)	97.1 (95.1-98.5)	97.3 (95.4-98.6)	95.7 (93.4-97.3)
Accuracy	86.9 (83.9-89.5)	96.1 (94.3-97.5)	97.8 (96.3-98.8)	98.0 (96.5-99.0)	96.6 (94.9-97.9)
Task 4: Etiological relation between a finding and a symptom					
Sensitivity	85.3 (78.2-90.8)	88.2 (81.6-93.1)	92.7 (87.0-96.2)	88.2 (81.6-93.1)	84.6 (77.4-90.2)
Specificity	52.8 (42-63.3)	66.7 (56.0-76.3)	61.5 (50.8-71.6)	72.5 (62.2-81.4)	73.6 (63.4-82.3)
F1	78.6 (73.5-83.2)	83.9 (79.1-88.0)	84.9 (80.3-88.7)	85.4 (80.7-89.2)	83.6 (78.7-87.8)
Accuracy	72.3 (65.9-78.0)	79.7 (73.8-84.7)	80.2 (74.4-85.2)	81.9 (76.3-86.7)	80.2 (74.4-85.2)

Supplementary Table 6: Examples of few-shot prompting for task 4. All provided examples are fake and provided to the model as a conversation.

Task 4: etiologic inference between a finding and a symptom		
Prompt	Doctor	I will present you abnormal findings from radiology exams. Your task is to list all the findings from them and tell me if they usually cause headache or not. Each conclusion is independant and should not influence your answer. You will answer only according to this template: List of findings: - List all the findings /can cause headache or not ?
Example 1	Doctor	No sign of cerebral venous thrombosis. Stable signs of intracranial hypertension.
	Robot	List of findings: - No signs of cerebral venous thrombosis /no headache - Stability of signs of intracranial hypertension /can cause headache
Example 2	Doctor	Left parietal periventricular heterotopia. Brain MRI otherwise normal.
	Robot	List of findings: - Left parietal periventricular heterotopy /no headache - otherwise normal encephalic MRI /no headache
Example 3	Doctor	Hypersignal B1000 lesion visible on FLAIR of the left precentral cortex, compatible with an ischemic origin in the first hypothesis. Permeability of the polygon of Willis.
	Robot	List of findings: - FLAIR-visible B1000 hypersignal lesion of the left precentral cortex, compatible with ischemic origin in first hypothesis /can cause headache - Permeability of the polygon of Willis /no headache
Example 4	Doctor	Voluminous intracranial mass responsible for subcentral involvement.
	Robot	List of findings: - Voluminous intracranial mass responsible for subcentral involvement /can cause headache
Example 5	Doctor	FLAIR hypersignals in the white matter, the radial nature of which may suggest an inflammatory origin. Leptomeningeal contrast enhancement (lumbar puncture? IIH?).
	Robot	List of findings: - FLAIR hypersignals in white matter whose radial nature may suggest an inflammatory origin /can cause headache - Leptomeningeal contrast enhancement: lumbar puncture or IIH /can cause headache
Example 6	Doctor	Filling of maxillary sinuses compatible with sinusitis. Multiple intracranial aneurysms as described: 8mm from basilar termination, 5mm from left sylvian bifurcation
	Robot	List of findings: - Framed filling of maxillary sinuses compatible with sinusitis /can cause headache - 8mm aneurysm of basilar termination /no headache - 5mm aneurysm of sylvian bifurcation /no headache
Real input	Doctor	<i>Real report</i>

Conclusion

Nous avons montré qu'un LLM open-source pouvait extraire des informations simples de comptes-rendus d'IRM cérébrales avec une précision de plus de 95%, et inférer une relation causale entre des anomalies décrites dans la conclusion et les symptômes avec une précision de 82%, sans avoir nécessité d'entraînement spécifique. Nous avons montré que l'utilisation d'un nombre restreint d'exemples contextuels permettait d'améliorer les performances initiales du modèle.

Notre étude est à ce jour la première à étudier les performances d'un LLM léger utilisé sur site pour extraire des informations depuis des comptes-rendus de vie réelle. Elle complète les données connues sur ce sujet, notamment l'étude de Mukherjee et collègues (45) dans laquelle Vicuna 13b a aussi été utilisé pour extraire des informations depuis des comptes-rendus de radiographie de thorax issus de la base de données monocentrique MIMIC. Leur travail consistait à extraire un nombre restreint de variables binaires (présence ou non de 13 anomalies). Nous complétons ce résultat en poussant Vicuna à réaliser une interprétation simple du compte-rendu pour en extraire sa normalité ou non et le caractère causal vis-à-vis des céphalées de l'anomalie éventuellement décrite. De plus, en travaillant sur des comptes-rendus d'IRM cérébrale issus de soins courants, notre étude explore donc l'applicabilité des modèles de langage open-source à des comptes-rendus réels et complexes, moins standardisés que ceux de radiographie de thorax.

Notre étude est à ce jour également la première à étudier les performances d'un LLM léger utilisé sur site sur des comptes-rendus non-anglophones. La plupart des études traitent jusqu'à présent de comptes-rendus issus de la base de données MIMIC

(39,46), en anglais. Les performances des modèles open-source pour des tâches appliquées à des textes non-anglophones, représentant la majorité des comptes-rendus rédigés chaque jour dans le monde, n'était donc pas connues.

Notre étude possède cependant des limites. Son caractère monocentrique, le manque de contexte clinique pour chaque compte-rendu, le caractère subjectif de la tâche la plus complexe, et l'amélioration continue des modèles disponibles, amenés à dépasser Vicuna, ont été détaillés dans l'article. Nous aimerions ici détailler trois autres limites, qui sont autant de perspectives pour les travaux à venir.

Premièrement, si nous avons montré une progression des performances avec le nombre d'exemples fournis au modèle, avec un plateau à 4-6 exemples, la nature des exemples n'a pas été développée. Des travaux théoriques montrent l'effet des exemples choisis et de leur ordre sur les performances des LLMs (47). Ici, nous n'avons utilisé qu'un seul jeu d'exemple par tâche, sans mesurer l'impact de ses variations sur les performances du modèle. Il conviendra donc d'étudier ces effets dans nos travaux à venir afin d'espérer implémenter en pratique la méthode que nous présentons.

Deuxièmement, l'influence de la langue n'a été analysée dans l'article que concernant la consigne donnée au modèle. Nous montrons en effet que donner la consigne en anglais ou en français n'impacte pas significativement les performances du modèle. Cependant, il se peut que Vicuna soit moins performant pour réaliser des tâches complexes sur des textes en français par rapport à des textes en anglais. En effet, l'anglais représente 89.7% du corpus sur lequel Vicuna a été entraîné, contre 0.16% pour le français, ce qui peut avoir des conséquences sur ses performances pour des tâches d'annotation (48). Si tel était le cas, il pourrait être intéressant de mesurer les performances de plusieurs modèles dans plusieurs langues pour

permettre aux centres de choisir le modèle le plus adapté à leur langue et à la tâche envisagée.

Enfin, la base de données sur laquelle nous avons travaillé n'est pas partageable car elle contient des données potentiellement identifiantes. Pour assurer la répliquabilité de nos résultats, il pourrait être intéressant de tester notre méthode sur un corpus de comptes-rendus de radiologie non-anglophones partageable, qui n'existe pas à l'heure actuelle.

Les LLMs sont susceptibles de modifier les pratiques des radiologues dans un avenir proche. De nombreuses tâches pourraient être automatisées, libérant du temps aux radiologues (49,50). Des travaux émergent actuellement pour saisir le potentiel de ces modèles de fondation.

Gertz et collègues (42) ont utilisé ChatGPT pour détecter des incohérences entre la section « résultats » et la section « conclusion » de comptes-rendus de radiologie écrits en allemand. Sur un corpus de 200 comptes-rendus, parmi lesquels ont été introduites 150 incohérences de différents types, ChatGPT était aussi bon qu'une relecture humaine modérément experte (interne en radiologie ou jeune radiologue titulaire), en détectant environ 80 % des erreurs. De plus, ils montrent que les coûts d'utilisation de ChatGPT sont largement amortis par le temps qu'il libère aux radiologues. Avec l'augmentation de la charge de travail en radiologie, cette étude offre une perspective réaliste sur l'utilisation des modèles de fondation en radiologie clinique. En restreignant son analyse à ChatGPT, elle n'apporte cependant pas d'alternative aux centres ne souhaitant pas transférer de comptes-rendus à OpenAI, ou disposant de modèles open-source performants qui pourraient réaliser ces tâches. De plus, une répliquabilité de ces travaux dans d'autres langues que l'allemand est nécessaire.

Les patients ayant de plus en plus accès aux comptes-rendus de leurs examens, et ces examens étant de plus en plus complexes, la bonne compréhension des comptes-rendus par tous leurs lecteurs (patients, médecins spécialisés ou non), est aussi un enjeu. Amin et collègues (41,46) ont mesuré les capacités de simplifier des comptes-rendus de radiologie de différents modèles de langage, tous commerciaux (Bard et différentes versions de GPT). Ils montrent que les modèles utilisés parviennent à abaisser le niveau de langue requis pour comprendre le texte, tout en conservant les informations du compte-rendu. Ce type d'étude pionnière ouvre la porte à des utilisations pratiques fascinantes, où plusieurs comptes-rendus de différents niveaux de compréhension pourraient être rédigés et validés par le radiologue à partir de la rédaction d'un compte-rendu unique. Pour le moment, aucune étude n'a répliqué ces résultats dans d'autres langues que l'anglais, ou avec d'autres modèles que ChatGPT ou Bard. De plus, la compréhension du compte-rendu n'est testée qu'automatiquement par des logiciels. Il pourrait être intéressant de présenter différents types de comptes-rendus (natifs, simplifiés à la main, simplifiés par des modèles de langage) à des patients ou à des médecins non-experts pour tester leurs préférences.

Au-delà de ces travaux sur le texte du compte-rendu, les modèles génératifs de fondation ont désormais la capacité d'intégrer des informations multimodales, c'est-à-dire du texte, mais aussi des images, du son, de la vidéo, etc. Toujours entraînés de manière non supervisée, ils ne sont pas destinés à accomplir une tâche en particulier, mais à transformer un contexte (une image, une information, une consigne) en texte ou en image. Ainsi, leur représentation du monde est élargie par rapport à celle d'un modèle d'IA habituel en médecine, qui n'a « vu » que des examens correspondant à une tâche spécifique (des AVC en IRM axiale, par exemple), alors que le modèle

multimodal a vu différents types d'examens contextualisés par des informations cliniques, de biologie, mais aussi éventuellement d'autres images ou d'autres textes issus du monde naturel. Le modèle GPT-4V (« V » pour vision) peut analyser des radiographies de thorax et en proposer une interprétation. Zhou et collègues (51) ont mesuré les capacités de GPT-4V à donner le code *International Classification of Diseases* (ICD) des pathologies présentes sur des radiographies de thorax issues de bases de données. Ils montrent que la valeur prédictive positive du modèle n'est que de 25 % (Si un diagnostic est avancé par GPT-4V, il a 25 % de chance d'être présent en réalité.). Ces performances basses ne doivent pas faire croire que ces modèles ne seront jamais capables d'interpréter des examens de radiologie. Déjà, un modèle plus récent développé par Google spécifiquement pour la médecine (alors que GPT-4V peut autant décrire l'espèce précise d'une photo de sauterelle que tenter d'interpréter une radiographie), Med-Gemini 3D, semble dépasser les performances du modèle d'OpenAI (52). De plus, ce modèle peut traiter des examens en 3D (des scanners ou des IRM), ce que ne pouvait pas faire GPT-4V, sorti il y a seulement quelques mois. Ces modèles de fondation se distinguent des modèles habituels d'IA parce qu'ils ne sont pas entraînés pour une tâche spécifique, mais sur des corpus divers d'images et de textes. Leur capacité à interpréter des examens de radiologie émerge donc d'une connaissance plus holistique de la diversité des examens de radiologie.

En intégrant des informations multimodales, ces modèles de fondation pourraient marquer un tournant dans l'IA en radiologie. Comme le radiologue, ils basent leur réponse sur une intégration de données multimodales (biologie, cliniques et images), et des informations sur le monde naturel (Qu'est-ce qu'un signe du colibri pour un modèle qui n'a jamais vu de colibri ?). Ces nouveaux modèles qui ne dépendent pas de données annotées pour une tâche précise émergent et capitalisent

sur les progrès à l'origine des LLM. Il sera passionnant de suivre leur maturation et leur commercialisation afin de vérifier si leurs immenses promesses se concrétisent cette fois par une adoption quotidienne dans le flux de travail des radiologues.

Avant même d'être des outils disponibles pour les radiologues, les LLMs seront probablement intégrés dans les services de santé publique et au sein des entrepôts de données de santé. La capacité des LLMs à assimiler une grande quantité d'informations multimodales répondra en effet aux attentes de ces professionnels des données. Les médecins de santé publique seront des acteurs privilégiés de cette « révolution de l'information », et seront sollicités pour faire émerger des projets autour des données nouvelles libérées par ces assistants infatigables. Aux registres et aux cohortes s'ajouteront des données de vie réelles structurées issues de la valorisation des textes générés en soins courants, ou des documents produits par les patients (questionnaires libres, photographies, enregistrements). Les médecins de santé publique, experts en données et en méthodologie, seront au premier plan de cette révolution. Ils auront à faire des choix stratégiques déterminants : Quels modèles utiliser ? Issus de quel entraînement ou réentraînement ? Comment intégrer ces modèles aux EDS et aux services de santé publique en assurant la protection des données critiques ? Et inversement, comment partager les données des EDS et des services de santé publique avec les acteurs de l'IA? Ils auront aussi pour rôle d'informer les patients des forces et des faiblesses de ces modèles, et de former les étudiants à leurs applications cliniques.

L'adoption de ces nouveaux outils aura d'autant plus de chances d'être effective que leurs utilisateurs en connaîtront les limites. Ces limites sont actuellement difficiles à cerner, en raison de leurs capacités à maintenir l'illusion d'une facilité à manipuler les sujets avec autant de confiance. Ces modèles sont systématiquement trop

confiants : quand un utilisateur leur demande leur niveau de certitude sur la réponse apportée, elle est pratiquement toujours maximale. Récemment, Krishna et collègues (53) ont montré que GPT-3.5 et GPT-4 rapportaient une confiance de 80% ou plus dans leur réponse à des questions de radiologie pour 100% et 94% des questions posées respectivement. Dans leur étude, la confiance exprimée n'apportait pas ou peu d'informations sur l'exactitude de la réponse. Les modèles utilisent aussi abondamment des marqueurs linguistiques de confiance ("assurément", "je sais", etc.), et arrivent mal à témoigner explicitement de leurs incertitudes. Il est donc difficile pour un utilisateur de mesurer la confiance du modèle dans sa réponse, ce qui peut être la source de fausses croyances en cas d'erreur ou d'hallucination du modèle. Des travaux devront être menés pour apporter aux utilisateurs un indice de confiance dans la réponse apportée, notamment en cas d'utilisation du modèle dans un contexte médical. En pratique, il s'agit donc de faire preuve d'esprit critique vis-à-vis des réponses des modèles de langage. Ces assistants ont l'air d'avoir réponse à tout et de ne refuser aucune tâche, mais leur certitude déconcertante peut cacher des zones grises de savoir, que seul un utilisateur informé peut identifier.

Par exemple, seul un radiologue francophone peut identifier que cette traduction de compte-rendu de MIMIC (fig. 12) n'est pas parfaite, et que partager ce compte-rendu avec des collègues avisés pourrait trahir l'utilisation d'une IA. En particulier, certains idiomes ne sont pas respectés (On ne dirait pas « Histoire clinique » mais « Indication », on ne dirait pas « Les cellules mastoïdiennes gauches sont réduites et opacifiées, similaires au scanner précédent. Cela est probablement chronique. » mais « Stabilité de la raréfaction de la trame osseuse des cellules mastoïdiennes gauches, comblées, probablement chronique ».). Ces détails traduisent probablement d'un faible niveau d'exposition de ces modèles à des

comptes-rendus de radiologie en langue française, ce qui ne les rend pas capables de reconnaître une tournure idiomatique d'une tournure issue d'une traduction mot à mot. Autre exemple, pour critiquer l'interprétation des radios de thorax par GPT-4V, il est nécessaire d'être médecin (fig. 13: On ne peut voir aucun implant urologique sur le champ d'acquisition de la radiographie de thorax.). En l'état, même si les modèles de langage sont bluffants, ils ne sont pas parfaits et, tant qu'ils ne pourront pas eux-mêmes témoigner de leurs incertitudes, les médecins devront être des interlocuteurs privilégiés pour interpréter des données à l'aide d'IA générative.

CLINICAL HISTORY:left-sided weakness. History of CNS lymphoma. Evaluate for mass or other acute abnormality.

TECHNIQUE: Contiguous axial images of the brain were obtained without the administration of IV contrast. Coronal and sagittal as well as thin bone-algorithm reconstructed images were obtained.

FINDINGS:

There is no evidence of hemorrhage or large vascular territory infarction. The ventricles are enlarged, similar to the prior exam. There is no evidence of worsening hydrocephalus. The basal cisterns are patent. Periventricular confluent white matter hypodensities are noted, and similar to the FLAIR abnormalities on the prior MRI from the end of _____. The hypodensities are most marked along the right occipital lobe, and in the area of the previously known lymphoma and biopsy tract. Again, these are not significantly changed from the prior MRI. The 8 mm mass seen on the MRI is not appreciated on today's CT, though likely because of technique. No new mass or focal edema is identified.

No fracture is identified. Postsurgical changes from a biopsy are noted in the right occipital bone. There is near-complete opacification of the left maxillary sinus with thickening of the surrounding bone. This suggests chronic sinusitis. Mild mucosal thickening is noted in the remainder of the paranasal sinuses. No air-fluid level is identified. The right mastoid air cells are clear. The left mastoid air cells are diminutive and opacified, similar to the prior CT. This is likely chronic.

IMPRESSION:

No acute intracranial abnormality. Enlargement of the ventricles and periventricular hypodensities appear similar to the prior MRI. The known small mass in the right occipital lobe is not well appreciated on today's exam. If further evaluation is needed to evaluate for change in the extent of disease, an MRI is recommended.

HISTOIRE CLINIQUE : Nouvelle faiblesse du côté gauche. Antécédents de lymphome du système nerveux central. Évaluation à la recherche de masse ou d'autres anomalies aiguës.

TECHNIQUE : Des images axiales continues du cerveau ont été obtenues sans administration de contraste IV. Des images reconstruites coronales et sagittales ainsi que des images fines avec un algorithme osseux ont été obtenues.

RÉSULTATS :

Il n'y a pas de preuve d'hémorragie ou d'infarctus de grand territoire vasculaire. Les ventricules sont élargis, similaires à l'examen précédent. Il n'y a pas de preuve d'aggravation de l'hydrocéphalie. Les citernes basales sont perméables. Des hypodensités de la substance blanche péri-ventriculaire confluentes sont notées, similaires aux anomalies FLAIR de l'IRM précédente datant de la fin de _____. Les hypodensités sont les plus marquées le long du lobe occipital droit et dans la région du lymphome précédemment connu et du trajet de la biopsie. Encore une fois, celles-ci n'ont pas significativement changé par rapport à l'IRM précédente. La masse de 8 mm vue sur l'IRM n'est pas visible sur l'actuel scanner, probablement en raison de la technique. Aucune nouvelle masse ou œdème focal n'est identifiée.

Aucune fracture n'est identifiée. Des changements postchirurgicaux dus à une biopsie sont notés dans l'os occipital droit. Il y a une opacification presque complète du sinus maxillaire gauche avec un épaississement de l'os environnant. Cela suggère une sinusite chronique. Un épaississement muqueux léger est noté dans le reste des sinus paranasaux. Aucun niveau air-liquide n'est identifié. Les cellules mastoïdiennes droites sont claires. Les cellules mastoïdiennes gauches sont réduites et opacifiées, similaires au scanner précédent. Cela est probablement chronique.

IMPRESSION :

Aucune anomalie intracrânienne aiguë. L'élargissement des ventricules et les hypodensités péri-ventriculaires semblent similaires à l'IRM précédente. La petite masse connue dans le lobe occipital droit n'est pas bien visualisée lors de l'examen d'aujourd'hui. Si une évaluation supplémentaire est nécessaire pour évaluer un changement dans l'étendue de la maladie, une IRM est recommandée

Figure 12: À gauche, compte-rendu de MIMIC, à droite, traduction par GPT-4, en précisant que le compte-rendu devait être formulé comme un vrai compte-rendu écrit initialement en français.

GPT-V4 (GPT-4 with Vision) for Detection of Findings on Chest Radiographs

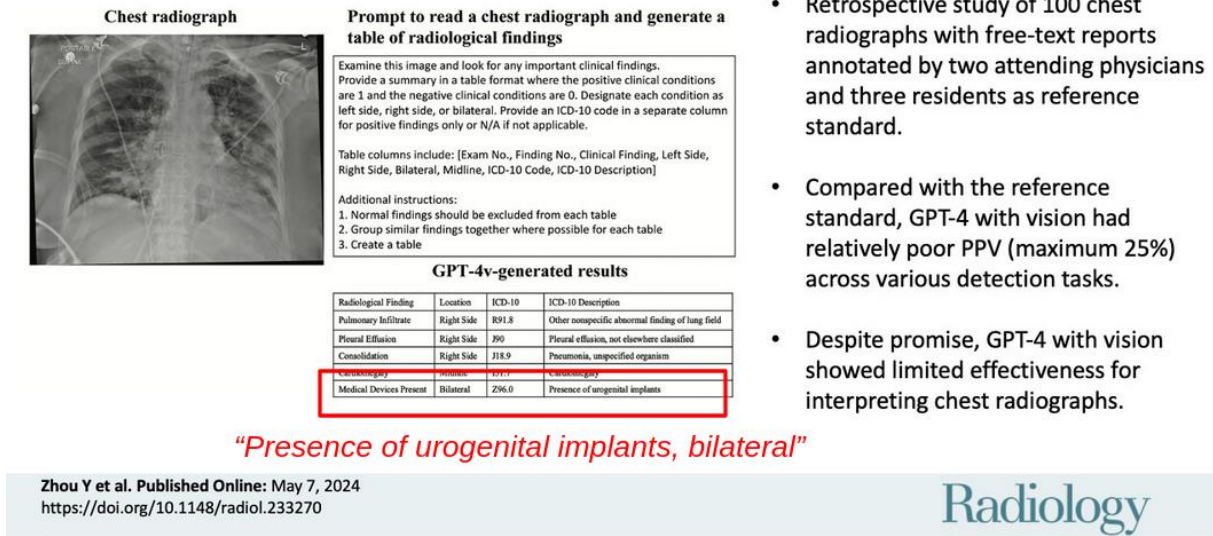


Figure 13: Graphical abstract de l'article GPT-4V for Detection of Findings on Chest Radiographs. L'interprétation de GPT fait état d'un matériel uro-génital bilatéral. Le diagnostic est absurde pour un radiologue. D'après Woojin Kim.

Tant que ces développements continuent, il devra donc exister un dialogue entre les développeurs de ces modèles et les médecins, pour identifier les besoins cliniques réels, les forces des modèles existants, mais aussi leurs faiblesses, parfois subtiles pour des utilisateurs non médecins. En particulier, la radiologie semblant pionnière dans ce domaine, des experts à la fois de la radiologie et de ces modèles de langage devront faire le lien entre ces deux communautés, pour faire profiter aux médecins et aux patients des avancées formidables des IA génératives, et discuter avec leurs concepteurs des attentes et des contraintes du métier de radiologue.

À l'aube de cette ère nouvelle, où l'intelligence artificielle et les modèles de langage ouvrent des horizons inédits en radiologie, notre étude préfigure les progrès et les défis qui marqueront la transformation de cette spécialité, faisant ainsi le pont

entre innovation technologique et pratique clinique pour une médecine toujours plus précise et personnalisée.

(La dernière phrase de ce texte a été écrite par ChatGPT. Le reste a été écrit par l'auteur.)

Références

1. Pari E, Rinaldi F, Gipponi S, et al. Management of headache disorders in the Emergency Department setting. *Neurol Sci.* 2015;36(7):1153–1160. doi: 10.1007/s10072-015-2148-7.
2. Goldstein JN, Camargo CA, Pelletier AJ, Edlow JA. Headache in United States emergency departments: demographics, work-up and frequency of pathological diagnoses. *Cephalalgia.* 2006;26(6):684–690. doi: 10.1111/j.1468-2982.2006.01093.x.
3. Henry P, Auray JP, Gaudin AF, et al. Prevalence and clinical characteristics of migraine in France. *Neurology.* Wolters Kluwer; 2002;59(2):232–237. doi: 10.1212/WNL.59.2.232.
4. Do TP, Remmers A, Schytz HW, et al. Red and orange flags for secondary headaches in clinical practice. *Neurology.* 2019;92(3):134–144. doi: 10.1212/WNL.0000000000006697.
5. Bø SH, Davidsen EM, Gulbrandsen P, Dietrichs E. Acute headache: a prospective diagnostic work-up of patients admitted to a general hospital. *Eur J Neurol.* 2008;15(12):1293–1299. doi: 10.1111/j.1468-1331.2008.02279.x.
6. Moisset X, Mawet J, Guegan-Massardier E, et al. Recommandations pour la prise en charge d'une céphalée en urgence. *Douleurs : Évaluation - Diagnostic - Traitement.* 2018;19(1):4–16. doi: 10.1016/j.douler.2018.01.001.
7. ElHabr A, Merdan S, Ayer T, et al. Increasing Utilization of Emergency Department Neuroimaging From 2007 Through 2017. *AJR Am J Roentgenol.* 2022;218(1):165–173. doi: 10.2214/AJR.21.25864.

8. Trofimova AV, Duszak Jr. R, Kadom N, Sadigh G. Increasing and disparate use of neuroimaging for adults and children with non-traumatic headaches in the US emergency departments: Opportunities for improvement. *Headache: The Journal of Head and Face Pain*. 2021;61(1):179–189. doi: 10.1111/head.14020.
9. Jang YE, Cho EY, Choi HY, Kim SM, Park HY. Diagnostic Neuroimaging in Headache Patients: A Systematic Review and Meta-Analysis. *Psychiatry Investig*. 2019;16(6):407–417. doi: 10.30773/pi.2019.04.11.
10. Fakih R. Céphalées aux urgences: évaluation de la prise en charge et prévalence des céphalées secondaires dans un centre tertiaire. Université de Lille, 2018.
11. Happonen T, Nyman M, Ylikotila P, Merisaari H, Mattila K, Hirvonen J. Diagnostic yield of emergency MRI in non-traumatic headache. *Neuroradiology*. 2023;65(1):89–96. doi: 10.1007/s00234-022-03044-2.
12. Russo A, Silvestro M, Tessitore A, et al. Arterial spin labeling MRI applied to migraine: current insights and future perspectives. *J Headache Pain*. 2023;24(1):71. doi: 10.1186/s10194-023-01597-y.
13. Décision du 28 mai 2018 de l'Union nationale des caisses d'assurance maladie relative à la liste des actes et prestations pris en charge par l'assurance maladie.
14. IRM. ccam-radiologie. <https://www.ccam-radiologie.fr/irm/>. Accessed December 19, 2023.
15. scanner. ccam-radiologie. <https://www.ccam-radiologie.fr/scanner/>. Accessed December 19, 2023.

16. Redd V, Levin S, Toerper M, Creel A, Peterson S. Effects of Fully Accessible Magnetic Resonance Imaging in the Emergency Department. *Academic Emergency Medicine*. 2015;22(6):741–749. doi: 10.1111/acem.12686.
17. Hassanpour S, Langlotz CP. Information extraction from multi-institutional radiology reports. *Artificial Intelligence in Medicine*. 2016;66:29–39. doi: 10.1016/j.artmed.2015.09.007.
18. Ficheur G, Schaffar A, Caron A, Balcaen T, Beuscart J-B, Chazard E. Elderly Surgical Patients: Automated Computation of Healthcare Quality Indicators by Data Reuse of EHR. *Stud Health Technol Inform*. 2016;221:92–96.
19. Judith F. Études en vie réelle pour l'évaluation des médicaments et dispositifs médicaux. 2021;
20. Villani, Cedric. Rapport de Cédric Villani: donner un sens à l'intelligence artificielle (IA). https://fichiers.acteurspublics.com/redac/pdf/2018/2018-03-28_Rapport-Villani.pdf. Accessed May 9, 2024.
21. Bosmans JML, Peremans L, Menni M, De Schepper AM, Duyck PO, Parizel PM. Structured reporting: if, why, when, how-and at what expense? Results of a focus group meeting of radiology professionals from eight countries. *Insights Imaging*. 2012;3(3):295–302. doi: 10.1007/s13244-012-0148-1.
22. Lee B, Whitehead MT. Radiology Reports: What YOU Think You're Saying and What THEY Think You're Saying. *Current Problems in Diagnostic Radiology*. 2017;46(3):186–195. doi: 10.1067/j.cpradiol.2016.11.005.
23. Larson DB, Towbin AJ, Pryor RM, Donnelly LF. Improving Consistency in Radiology Reporting through the Use of Department-wide Standardized Structured

Reporting. *Radiology*. Radiological Society of North America; 2013;267(1):240–250. doi: 10.1148/radiol.12121502.

24. Langlotz CP. Automatic Structuring of Radiology Reports: Harbinger of a Second Information Revolution in Radiology. *Radiology*. Radiological Society of North America; 2002;224(1):5–7. doi: 10.1148/radiol.2241020415.

25. Pinto dos Santos D, Cuocolo R, Huisman M. O structured reporting, where art thou? *Eur Radiol*. 2023; doi: 10.1007/s00330-023-10465-x.

26. Lo S-CB, Lou S-LA, Lin J-S, Freedman MT, Chien MV, Mun SK. Artificial convolution neural network techniques and applications for lung nodule detection. *IEEE Transactions on Medical Imaging*. 1995;14(4):711–718. doi: 10.1109/42.476112.

27. Knight W. OpenAI's CEO Says the Age of Giant AI Models Is Already Over. *Wired*. <https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/>. Accessed May 9, 2024.

28. Poignon V. Dès que vous validez un Captcha, vous entraînez en réalité les intelligences artificielles. Ça m'intéresse. 2023. <https://www.caminteresse.fr/insolite/des-que-vous-validez-un-captcha-vous-entraenez-en-realite-les-intelligences-artificielles-11188430/>. Accessed May 9, 2024.

29. Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. *arXiv*; 2023. doi: 10.48550/arXiv.1706.03762.

30. Papers with Code - C4 Dataset. <https://paperswithcode.com/dataset/c4>. Accessed May 9, 2024.

31. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. Nature Publishing Group; 2023;620(7972):172–180. doi: 10.1038/s41586-023-06291-2.
32. Ryan E. What Are the Legal Implications of ChatGPT? Scribbr. 2023. <https://www.scribbr.com/ai-tools/legal-implications-chatgpt/>. Accessed May 9, 2024.
33. IA : huit journaux américains poursuivent OpenAI et Microsoft en justice. *Le Monde.fr*. May 3, 2024; https://www.lemonde.fr/pixels/article/2024/05/03/ia-huit-journaux-americains-poursuivent-openai-et-microsoft-en-justice_6231347_4408996.html. Accessed May 8, 2024.
34. Extracting Training Data from ChatGPT. 2023. <https://not-just-memorization.github.io/extracting-training-data-from-chatgpt.html>. Accessed May 9, 2024.
35. IA générative – Un marché à 100 mds de dollars dès 2028 selon Sopra Steria Next. Sopra Steria FR. 2024. <https://www.soprasteria.fr/espace-media/publications/details/ia-generative-un-marche-a-100mds-de-dollars-des-2028-selon-sopra-steria-next>. Accessed May 9, 2024.
36. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*. Nature Publishing Group; 2023;10(1):1. doi: 10.1038/s41597-022-01899-x.
37. Marks M, Haupt CE. AI Chatbots, Health Privacy, and Challenges to HIPAA Compliance. *JAMA*. 2023; doi: 10.1001/jama.2023.9458.
38. Adams LC, Truhn D, Busch F, et al. Leveraging GPT-4 for Post Hoc Transformation of Free-Text Radiology Reports into Structured Reporting: A Multilingual Feasibility Study. *Radiology*. 2023;230725. doi: 10.1148/radiol.230725.

39. Fink MA, Bischoff A, Fink CA, et al. Potential of ChatGPT and GPT-4 for Data Mining of Free-Text CT Reports on Lung Cancer. *Radiology*. Radiological Society of North America; 2023;308(3):e231362. doi: 10.1148/radiol.231362.
40. Jethani N, Jones S, Genes N, et al. Evaluating ChatGPT in Information Extraction: A Case Study of Extracting Cognitive Exam Dates and Scores. *medRxiv*; 2023. p. 2023.07.10.23292373. doi: 10.1101/2023.07.10.23292373.
41. Amin KS, Davis MA, Doshi R, Haims AH, Khosla P, Forman HP. Accuracy of ChatGPT, Google Bard, and Microsoft Bing for Simplifying Radiology Reports. *Radiology*. Radiological Society of North America; 2023;309(2):e232561. doi: 10.1148/radiol.232561.
42. Gertz RJ, Dratsch T, Bunck AC, et al. Potential of GPT-4 for Detecting Errors in Radiology Reports: Implications for Reporting Accuracy. *Radiology*. Radiological Society of North America; 2024;311(1):e232714. doi: 10.1148/radiol.232714.
43. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality | LMSYS Org. . <https://lmsys.org/blog/2023-03-30-vicuna>. Accessed July 9, 2023.
44. Jiang AQ, Sablayrolles A, Mensch A, et al. Mistral 7B. *arXiv*; 2023. doi: 10.48550/arXiv.2310.06825.
45. Mukherjee P, Hou B, Lanfredi RB, Summers RM. Feasibility of Using the Privacy-preserving Large Language Model Vicuna for Labeling Radiology Reports. *Radiology*. Radiological Society of North America; 2023;309(1):e231147. doi: 10.1148/radiol.231147.
46. Doshi R, Amin KS, Khosla P, Bajaj SS, Chheang S, Forman HP. Quantitative Evaluation of Large Language Models to Streamline Radiology Report Impressions:

A Multimodal Retrospective Analysis. Radiology. Radiological Society of North America; 2024;310(3):e231593. doi: 10.1148/radiol.231593.

47. Zhao TZ, Wallace E, Feng S, Klein D, Singh S. Calibrate Before Use: Improving Few-Shot Performance of Language Models. arXiv; 2021. doi: 10.48550/arXiv.2102.09690.

48. Mohta J, Ak KE, Xu Y, Shen M. Are large language models good annotators? NeurIPS 2023 Workshop on I Can't Believe It's Not Better (ICBINB), 2023

49. Large Language Models as an Inexpensive and Effective Extra Set of Eyes in Radiology Reporting Radiology. <https://pubs.rsna.org/doi/10.1148/radiol.240844>. Accessed May 17, 2024.

50. Rahsepar AA. Large Language Models for Enhancing Radiology Report Impressions: Improve Readability While Decreasing Burnout. Radiology. Radiological Society of North America; 2024;310(3):e240498. doi: 10.1148/radiol.240498.

51. Zhou Y, Ong H, Kennedy P, et al. Evaluating GPT-4V (GPT-4 with Vision) on Detection of Radiologic Findings on Chest Radiographs. Radiology. Radiological Society of North America; 2024;311(2):e233270. doi: 10.1148/radiol.233270.

52. Advancing medical AI with Med-Gemini. <http://research.google/blog/advancing-medical-ai-with-med-gemini/>. Accessed May 17, 2024.

53. Krishna S, Bhambra N, Bleakney R and Bhayana R, Evaluation of Reliability, Repeatability, Robustness and Confidence of GPT-3.5 and GPT-4 on a Radiology Board-style Examination. Radiology. Radiological Society of North America; 2024; 311:2. Doi: 10.1148/radiol.232715

AUTEUR : Nom : Le Guellec **Prénom :** Bastien

Date de Soutenance : 17/06/2024

Titre de la Thèse : Performance d'un grand modèle de langage (LLM) open-source pour l'extraction d'information depuis des comptes-rendus d'IRM cérébrale

Thèse - Médecine - Lille 2024

Cadre de classement : Radiologie

DES + FST ou option : Radiologie

Mots-clés : céphalées, IRM, compte-rendu, intelligence artificielle, GPT, modèle de langage, NLP

Résumé :

Introduction : L'utilisation de l'IRM en urgence pour céphalées est en augmentation en Europe et aux États-Unis. Cependant, le rendement diagnostique de cet examen est débattu, avec un taux de positivité probablement faible, autour de 15%. Il n'existe pas d'outil pour automatiser le traitement des comptes-rendus d'IRM dans le but de calculer automatiquement cet indicateur. L'émergence des grands modèles de langage de type GPT a révolutionné le traitement du langage naturel, mais leur caractère fermé est un frein à leur déploiement en vie réelle. Notre objectif était d'utiliser Vicuna, un grand modèle de langage open-source, dans le but d'extraire automatiquement des données depuis des comptes-rendus textuels non structurés.

Matériel et méthodes : Tous les comptes-rendus consécutifs d'IRM cérébrales réalisées en urgence du CHU de Lille en 2022 ont été examinés rétrospectivement par quatre radiologues pour en extraire quatre variables : l'indication, l'injection de produit de contraste, le caractère normal ou non de l'examen, et le caractère causal ou non de l'anomalie décrite. Parallèlement, Vicuna, un grand modèle de langage open-source, a effectué les mêmes tâches.

Résultats : Un total de 2398 comptes rendus d'IRM cérébrales réalisées en urgence ont été identifiées, dont 595 incluait une céphalée dans leur indication. Notre méthode basée sur un modèle de langage avait une excellente précision (> 95 %) pour des tâches simples d'extraction d'informations depuis le compte-rendu telles que le sexe et l'âge du patient, l'injection de produit de contraste et la catégorisation de l'examen comme normal ou anormal d'après sa conclusion. La précision de Vicuna était également très bonne (82 %) pour la tâche la plus complexe d'inférence de causalité entre une anomalie décrite à l'IRM et la céphalée.

Conclusion : Nous avons montré qu'un modèle de langage de type GPT open-source peut extraire des informations de comptes-rendus radiologiques avec une excellente précision sans nécessiter d'entraînement spécifique. Cette méthode ne nécessite pas d'entraînement supplémentaire, peut être utilisée et partagée gratuitement, et respecte la protection des données de santé.

Composition du Jury :

Président : Monsieur le Professeur Jean-Pierre PRUVO

Assesseurs : Monsieur le Professeur Philippe AMOUYEL

Monsieur le Docteur Aghiles HAMROUN

Directeur : Monsieur le Docteur Grégory KUCHCINSKI