UNIVERSITÉ DE LILLE

UFR3S-MÉDECINE

Année : 2026

THÈSE POUR LE DIPLÔME D'ÉTAT
DE DOCTEUR EN MÉDECINE

Le centrisme américain dans la capacité des grands modèles de langage à suivre des recommandations de neuroimagerie

Présentée et soutenue publiquement le 30/01/2026 à 16:00
au Pôle Recherche
par Naël Bazerbachi

_____

JURY
Président :
Monsieur le Professeur Grégory KUCHCINSKI
Assesseurs :
Monsieur le Professeur Philippe AMOUYEL
Monsieur le Docteur Aghiles HAMROUN
Directeur de thèse :
Monsieur le Docteur Bastien LE GUELLEC

_____

Avertissement

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

Sigles

ACR	American College of Radiology
ADERIM	Aide à la Demande d'Examen de Radiologie et d'Imagerie Médicale
ASNR	American Society of Neuroradiology
AVC	Accident Vasculaire Cérébral
CMSC	Consortium of Multiple Sclerosis Centers
ESNR	European Society of Neuroradiology
FNMR	Fédération Nationale des Médecins Radiologues
GPT	Generative Pre-trained Transformer
HAS	Haute Autorité de Santé
IA	Intelligence Artificielle
IRM	Imagerie par Résonnance Magnétique
LLM	Large Language Model
MAI-DxO	Microsoft AI Diagnostic Orchestrator
MAGNIMS	Magnetic Resonance Imaging in Multiple Sclerosis
NAIMS	North American Imaging in Multiple Sclerosis Cooperative
OFSEP	Observatoire Français de la Sclérose en Plaque
PACS	Picture Archiving and Communication System
RLHF	Reinforcement Learning from Human Feedback
USMLE	United States Medical Licensing Examination

Sommaire

Avertissement 2
Sigles 3
Sommaire 4
Introduction 6
1 L'imagerie médicale de nos jours 6
1.1 Histoire et épidémiologie 6
1.2 Spécificités en neuroimagerie 9
2 Les difficultés de prescription et les recommandations de bonnes pratiques 12
3 Les grands modèles de langage (LLM) : fonctionnement et concepts 16
3.1 Le machine learning 17
3.1.1 L'apprentissage supervisé 18
3.1.2 L'apprentissage non supervisé 20
3.2 Neurones artificiels 22
3.3 Production de texte selon un modèle de probabilité 26
3.3.1 La tokenisation 27
3.3.2 Plongement lexical 27
3.3.3 Le processus d'attention 29
4 Potentiel et limites des LLMs pour l'aide à la décision médicale 31
4.1 Aide au diagnostic et à la pratique médicale courante 32
4.2 Aide à la prescription d'examens d'imagerie 34
4.3 Les limites des LLMs 37
5 Objectif 40
Article en Anglais 41
1 Abstract 41
1.1 Objectives 41
1.2 Materials and Methods 41
1.3 Results 41
1.4 Conclusion 42
1.5 Keywords 42
1.6 List of abbreviations 42
1.7 Key points 43
2 Introduction 43
3 Materials and methods 44
3.1 Guidelines 45

3.2 Adversarial cases 45
3.3 Large language models 46
3.4 Prompting 46
3.5 Responses review 46
3.6 Statistical analyses 47
4 Results 47
4.1 Implicit preference for U.S. guidelines 47
4.2 Effect of language on the implicit preference 47
4.3 Ability to follow an explicit guideline 48
4.4 Impact of web searches on explicit guideline following 48
4.5 Impact of providing the guideline as PDF on explicit guideline following 49
5 Discussion 49
6 References 53
7 Figure Legends 57
8 Annexes 60
Conclusion en Français 64
Références 67

Introduction

1 L’imagerie médicale de nos jours

1.1 Histoire et épidémiologie

Du fait des évolutions et progrès technologiques continus, l’imagerie médicale s’est considérablement développée ces dernières décennies pour devenir aujourd’hui un élément au cœur de la prise en charge des patients. Depuis la découverte des rayons X et la première radiographie de la main par Wilhelm C. Röntgen en 1895, les progrès se sont succédés. Les techniques invasives d’étude du cerveau et de la moelle épinière émergent aux environs de l’entre-deux-guerres avec la pneumoencéphalographie en 1918, la myélographie en 1921 ou encore les débuts de l’angiographie cérébrale en 1927. Les années 1950 voient le début de l’utilisation clinique de l'échographie, qui s'enrichit par la suite de nouvelles techniques comme le doppler, puis des microbulles de contraste et de l’élastographie. En 1972, Godfrey Hounsfield réalise le premier scanner cérébral, ouvrant la voie aux techniques d'acquisition hélicoïdales, aux multidétecteurs, aux reconstructions itératives, puis à la double énergie pour enfin arriver aujourd’hui à l'ère du comptage photonique. L’imagerie par résonance magnétique (IRM), introduite cliniquement dans les années 1980 à bas champs magnétique (1,5 T), n’a cessé de gagner en résolution et en fonctionnalités. Au début des années 1990, la séquence de diffusion a révolutionné le diagnostic précoce de l’accident vasculaire cérébral ischémique et la séquence de

susceptibilité magnétique est devenue routinière dans la détection des hémorragies cérébrales. Puis, au milieu des années 2010, les séquences de perfusion sans injection de produit de contraste ont trouvé une place importante dans la prise en charge de l'épilepsie, des pathologies tumorales ou encore des anomalies vasculaires cérébrales. De nos jours, l'IRM à haut champ magnétique (7T) explore finement le cortex et les petites structures anatomiques, ouvrant la voie vers de nouvelles innovations. Enfin, la numérisation des flux sur système d'archivage et de transmission d'images (PACS), la téléradiologie et le partage sécurisé des données ont amélioré la qualité de travail du personnel d'imagerie médicale, permettant ainsi de fluidifier la prise en charge des patients. Ces avancées technologiques perpétuelles ont inscrit l'imagerie comme un élément clé de la prise en soin des patients, et expliquent que son utilisation est en augmentation constante [1].

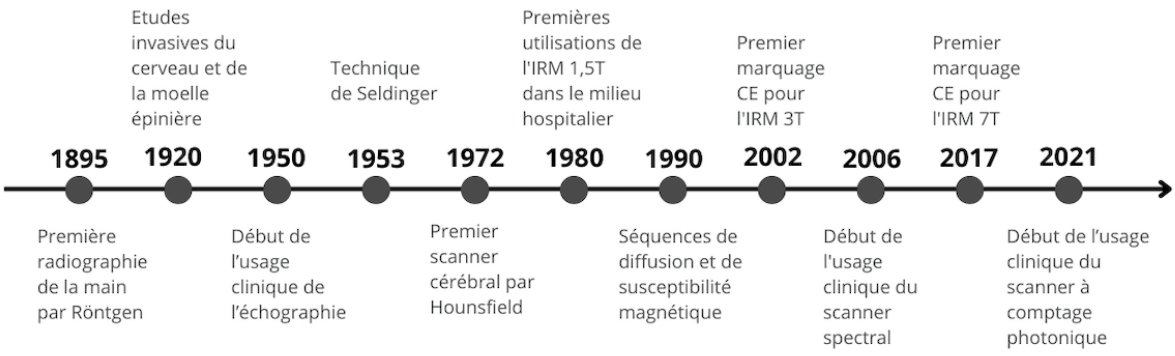
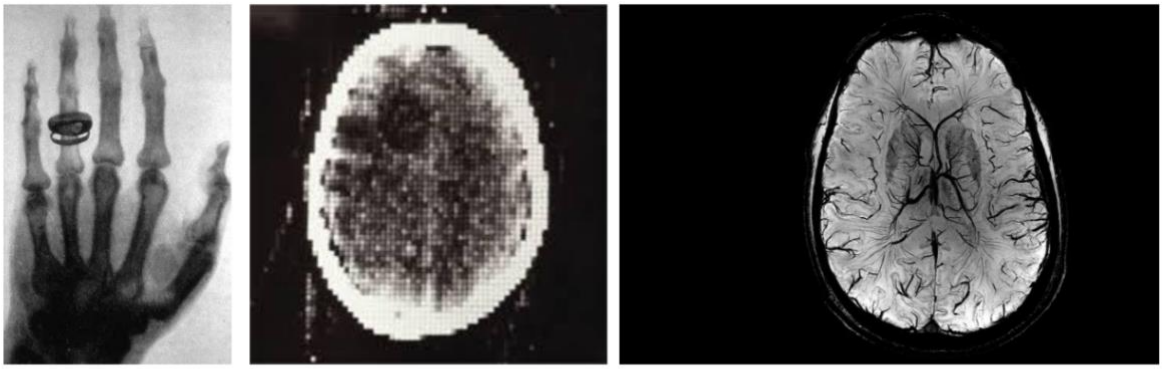


Figure 1 Frise chronologique des innovations technologiques de la radiologie.



a b c

Figure 2 Quelques exemples d'évolution des techniques d'imagerie

a. Première radiographie de la main d'Anna Bertha Ludwig, l'épouse de Wilhelm Röntgen (1895).

b. Premier scanner cérébral par Hounsfield (1972).

c. IRM cérébrale à 7T (2024).

En 2022, environ 96,5 millions d'actes d'imagerie ont été réalisés en France, seulement dans le secteur libéral qui représente à lui seul environ 70% de l'activité totale [2] .

Selon les différentes modalités techniques, la radiologie conventionnelle représente encore plus de la moitié des actes avec 53,7 millions d'examens, l'échographie vient en seconde position avec 30,8 millions d'examens et enfin l'imagerie dite lourde avec le scanner et l'IRM pour 12 millions d'examens. Ces chiffres sont en constante augmentation et on note une croissance d'environ 3% par an pour la radiographie conventionnelle, 2,5 % pour l'échographie et jusqu'à 13,5 % pour le scanner et l'IRM entre 2015 et 2022 selon la Fédération Nationale des Médecins Radiologues (FNMR) [2,3].

Aux Etats-Unis, une étude menée de 2013 à 2018 sur environ 2,8 milliards de visites a montré un recours à l'imagerie dans 12,5% des cas [4].
Un impact économique considérable découle de cette utilisation croissante des moyens d'imagerie avec un coût de 4,6 milliards d'euros par an en France en 2021 selon un rapport de la Sécurité Sociale, uniquement pour la radiologie libérale [5].
L'ensemble de ces données reflètent une augmentation croissante de l'utilisation des modalités techniques les plus sophistiquées en imagerie.

1.2 Spécificités en neuroimagerie

La neuroimagerie ne fait pas exception à ce développement continu et émerge aux environs de l'entre-deux-guerres avec l'arrivée de méthodes invasives et spécifiques d'étude du cerveau et de la moelle épinière. Il faut cependant attendre les années 1960 pour voir arriver le début de sa structuration administrative avec la société Américaine de Neuroradiologie (ASNR) fondée en 1962 et la Société Européenne de Neuroradiologie (ESNR) de 1969. Cette discipline devient alors reconnue officiellement comme sur-spécialité de la radiologie avec notamment la création de diplômes universitaires et d'agréments spécifiques en Europe dans les années 1990.

De par les caractéristiques anatomiques et physiologiques de l'encéphale et de la boîte crânienne, le développement de nouvelles technologies comme le scanner et l'IRM ont apporté une réelle plus-value dans ce domaine, offrant alors des outils de choix pour l'exploration de ces structures et devenant les deux modalités prépondérantes.

Le scanner est un examen rapide et accessible, particulièrement adapté aux situations d'urgence, dont le coût reste modéré en comparaison de l'IRM (environ 100€ pour un examen sans injection) [6].
Cependant, cet examen est irradiant, inconvénient inhérent au fait de sa technologie utilisant des rayonnements ionisants, d'une part responsable de la majeure partie de l'irradiation médicale à laquelle est exposée la population [7] mais pouvant également être source d'effets secondaires graves et imprévisibles comme l'apparition de cancer à distance [8].
De plus, l'utilisation d'un produit de contraste iodé comme agent opacifiant peut engendrer des effets secondaires comme l'apparition d'insuffisance rénale aiguë, ou être à risque d'allergie.

L'IRM quant à elle est une modalité d'imagerie moins accessible, dont la durée excède parfois une trentaine de minutes et dont le coût est nettement supérieur à celui du scanner (environ 250-300€) [9]. Elle est aussi bruyante et peut être gênante pour les patients sujets à la claustrophobie.
Cependant, l'IRM est un examen d'imagerie non irradiant, dont les performances dépassent celles du scanner en terme de sensibilité et de spécificité dans de nombreuses situations, notamment urgentes, comme par exemple la recherche de lésions ischémiques encéphaliques [10] ou la prise en charge de céphalées non traumatiques [11].
L'IRM peut nécessiter l'utilisation d'un agent opacifiant gadoliné, qui engendrera un surcoût mais aussi d'autres potentiels effets indésirables, à moindre mesure en comparaison du scanner [12].

La neuroradiologie se caractérise également par son activité dédiée aux urgences. La demande est telle que dans certains centres comme celui du Centre Hospitalier Universitaire de Lille, elle mobilise un neuroradiologue et un interne 24/24h et 7/7j, notamment en raison de certains agréments comme celui de centre référent dans le cadre de l'alerte thrombolyse pour les accidents vasculaires cérébraux (AVC). Une pathologie fréquente, puisqu'elle touche environ 150 000 patients par an en France [13], dans laquelle l'IRM joue un rôle crucial et pour laquelle chaque seconde de prise en charge compte afin d'espérer une récupération la plus optimale possible [10].

D'autres symptômes, comme les céphalées, font également partie du corpus de prise en charge avides d'imagerie cérébrale, justifiant cette activité d'urgences dédiée. On dénombre par exemple aux Etats-Unis 2,1 millions de consultations aux urgences par an pour ce motif, dont 14% bénéficient d'une imagerie cérébrale [14].

En complément de cette activité d'urgences, la neuro-imagerie trouve également sa place dans son activité d'examen programmé.

Un grand nombre de pathologies neurologiques nécessitent un recours à l'imagerie, autant dans le diagnostic que le suivi, notamment dans l'évaluation de la réponse aux traitements. C'est le cas par exemple du suivi oncologique des tumeurs cérébrales, des maladies neurodégénératives, de l'épilepsie ou des maladies démyélinisantes comme la sclérose en plaque, dont l'arsenal thérapeutique repose aujourd'hui de plus en plus sur l'imagerie pour savoir qui et quand traiter, à quelle dose, et détecter les effets indésirables des traitements.

Le développement des techniques de chirurgie et le suivi post-opératoire par imagerie, tout comme la multiplicité des techniques d'exploration, notamment des

séquences d'IRM (diffusion, temps de vol, perfusion, spectroscopie,...), impliquent l'expertise d'un radiologue qualifié en neuro-imagerie. A Lille par exemple, une IRM est installée au bloc de neurochirurgie et permet notamment de réaliser des imageries peropératoires, c'est-à-dire en cours de chirurgie, pour s'assurer de la qualité de l'exérèse de tumeurs et l'absence de complication. Les examens sont interprétés par un neuroradiologue et les résultats sont discutés en direct avec le chirurgien.

Enfin, le neuroradiologue s'inscrit dans la prise en charge multidisciplinaire du patient en devenant un correspondant de choix pour permettre une communication optimale avec d'autres spécialistes comme les neurologues, neurochirurgiens, neurooncologues, anesthésistes-réanimateurs,...

2 Les difficultés de prescription et les recommandations de bonnes pratiques

Avec ce recours croissant à l'imagerie médicale, les innovations technologiques constantes, la sur-spécialisation de la radiologie et l'avènement de l'imagerie multi-modale, il devient de plus en plus difficile pour le médecin prescripteur d'être constamment pertinent quant à la prescription et au choix de la modalité d'imagerie à employer.

De ce fait, on note une sur-utilisation de l'imagerie médicale, avec un surcoût pour les examens dont l'indication est parfois jugée peu pertinente dans la prise en charge

des patients. Une enquête réalisée en 2020 a montré qu'environ une demande sur sept était correctement formulée de la part du prescripteur [15], pouvant parfois entraîner un surcroît d'effets secondaires comme l'augmentation d'effets indésirables liés aux rayonnements ionisants dans le cas du scanner [7].

Afin de pallier cette sur-utilisation et dans le but de guider le médecin prescripteur, des sociétés savantes émettent des recommandations de bonne pratique pour la prescription d'examen d'imagerie [16] [17] [18].
Il peut s'agir de documents traitant spécifiquement de la prise en charge d'une pathologie en général avec une rubrique concernant l'imagerie [19] ou encore de documents traitant exclusivement des différents types de modalité d'imagerie pertinents pour le diagnostic ou la prise en charge d'une pathologie [20]. Il existe également des algorithmes de prise en charge précisant l'utilité ou non du recours à l'imagerie médicale dans certaines situations [21].

Ces sociétés savantes qui rédigent ces documents sont cependant elles-aussi multiples. Il en existe de nombreuses au sein d'un même pays, mais aussi à l'échelle internationale ; des sociétés nationales et continentales, traitant parfois des mêmes pathologies.
De plus, elles ne sont pas toujours d'accord entre elles et se contredisent parfois dans la prise en charge d'une même situation ; par exemple, lors de la suspicion d'un AVC ischémique, la Haute Autorité en Santé (HAS) française recommande l'utilisation d'une IRM cérébrale, examen possédant la meilleure sensibilité [10], tandis que le Collège Américain de Radiologie (ACR) recommande un scanner cérébral sans injection, examen plus facilement disponible [22]. Ces

recommandations s'inscrivent dans des pratiques cliniques qui varient selon les pays: en France, le diagnostic positif d'AVC doit la plupart du temps être posé en IRM pour amorcer un traitement, alors qu'aux Etats-Unis, c'est plutôt l'élimination d'un diagnostic alternatif par scanner qui motive la réalisation rapide de l'examen le plus disponible. On voit donc que les recommandations en neuro-imagerie témoignent d'habitudes et de raisonnements différents selon les contextes géographiques et économiques et ne sont pas interchangeables.

Ces recommandations divergentes peuvent traiter de diagnostics graves, comme celui d'état de mort encéphalique, où la loi française stipule qu'il est obligatoire, même si le diagnostic est cliniquement posé, de réaliser un examen complémentaire pour confirmation. Aux Etats-Unis, la loi n'indique un examen complémentaire de confirmation que si le diagnostic ne peut pas être posé cliniquement. De plus, les modalités d'examen ne sont pas les mêmes ; angioscanner cérébral ou électro-encéphalogramme en France / angiographie cérébrale, doppler transcrânien ou scintigraphie de perfusion encéphalique aux Etats-Unis [23,24].

Les recommandations peuvent également diverger au sujet des protocoles d'examen à réaliser dans certaines indications, découlant de la multiplicité des séquences accessibles en IRM. C'est le cas du protocole recommandé dans le suivi de la sclérose en plaque. L'Observatoire Français de la Sclérose en Plaque (OFSEP) recommande l'utilisation systématique d'une séquence de diffusion encéphalique [25], contrairement à l'association des sociétés américaines de Magnetic Resonance Imaging in Multiple Sclerosis (MAGNIMS), the Consortium of Multiple Sclerosis Centers (CMSC) et North American Imaging in Multiple Sclerosis Cooperative (NAIMS), qui rend cette séquence optionnelle [26].

Avec cette multiplication croissante de recommandations, intrinsèquement liées à l'augmentation de la prescription d'examens d'imagerie, et malgré leur but initial d'aide à la prescription, il devient de plus en plus difficile pour le prescripteur de se tenir informé de leurs évolutions et de les appliquer.

Quinze ans après leur création, une étude à montrer que les critères de bonnes pratiques de l'ACR n'étaient que peu utilisés par les médecins prescripteurs qui se basaient plus souvent sur leurs propres connaissances, malgré le fait que de nombreuses ressources aient été mobilisées pour les mettre en place [27].

De cette sous-utilisation des recommandations découle également des difficultés relationnelles entre médecin prescripteur et radiologue, générant des incompréhensions et parfois même des tensions [15].

Dans ce contexte, il serait pertinent d'avoir un outil rapide, facile d'accès et d'utilisation, afin de garantir une aide optimale au prescripteur d'examen, dont le temps de consultation est malheureusement souvent trop court (10,7 minutes en moyenne en Europe) [28].

Compte-tenu de la multiplicité infinie des situations cliniques, un algorithme d'aide à la décision devrait s'adapter précisément au contexte de présentation (informations cliniques, lieu de date de la consultation, accès à l'imagerie). Des outils d'aide à la prescription sont en développement en France, notamment Aide à la Demande d'Examens de Radiologie et d'Imagerie Médicale (ADERIM). Cependant, ceux-ci fonctionnent comme des listes de situations de départ et ne peuvent contenir toute la variabilité des consultations de vie réelle.

Les grands modèles de langage (Large Language Models, ou LMM) sont aujourd'hui des outils de plus en plus performants et prometteurs pour générer, extraire mais aussi pour rechercher et mettre en relation différents éléments de textes. De ce fait, ils pourraient constituer un bon outil d'aide à la prescription d'examen d'imagerie [29].

3 Les grands modèles de langage (LLM) : fonctionnement et concepts

En 1959, Arthur Samuel, l'un des pionniers dans le domaine de l'intelligence artificielle (IA), l'a définie comme : "la branche d'étude qui confère aux ordinateurs la capacité d'apprendre sans être explicitement programmé pour le faire". Il faut cependant attendre les années 2010 pour arriver à l'heure de la montée en puissance du "deep learning", pourtant bien connu depuis les années 1990 et dont le français Yann Le Cun est l'un des précurseurs, pour ouvrir la voie aux LLMs. Puis, la parution en 2017 de l'article "Attention is all you need" [30], a constitué le concept socle des LLMs modernes, popularisé en 2020 par OpenAI, avec le modèle GPT-3 (Generative Pre-trained Transformer 3).

Les LLM sont des outils dont la fonction première est d'analyser et de générer du texte. Ils permettent par exemple de résumer des textes, écrire des histoires, traduire des langues différentes, rechercher et restituer des informations dans des textes,... Leurs performances impressionnantes, à l'origine d'un engouement croissant auprès du grand public, sont telles qu'on pourrait croire qu'ils « pensent » et qu'ils sont capables d'une réflexion proche de celle de l'homme. En réalité, ce sont des

modèles algorithmiques entrainés à prédire une réponse selon un mode probabiliste à partir d'une grande quantité d'informations engrangée au préalable. Ils reposent pour cela sur des algorithmes d'intelligence artificielle, ou algorithmique probabiliste. Les LLM, sont donc des programmes informatiques qui traitent le langage humain. Le terme "Large" fait référence à la taille du programme, celui-ci étant composé de milliards de paramètres. Ils sont entraînés avec de gigantesques quantités de textes, afin de leur permettre par la suite d'en générer eux-mêmes, dans différents contextes, via ces multiples paramètres, qui sont des variables d'ajustement des calculs réalisés sur le texte. Le tout permet ensuite aux modèles de traiter les nuances les plus subtiles et complexes de notre langage, afin de prédire la réponse la plus probable à la question qui leur est posée.

3.1 Le machine learning

Le machine learning ou apprentissage automatique, correspond au fait "d'apprendre sans être explicitement programmé pour le faire". C'est à dire qu'un programme informatique utilisant un algorithme ne sera pas spécifiquement programmé à réaliser une tâche particulière, mais qu'il apprendra à le faire de lui-même par le biais de l'apprentissage.

C'est par exemple ce qui a été démontré en 2012, lorsque le projet Google Brain de la firme Google a découvert par lui-même le concept de "chat". Pendant trois jours, 10 millions de captures d'écran issues de YouTube ont été analysées par l'algorithme. Par la suite, le programme était capable de reconnaître un chat sur de nouvelles images, sans que l'on ne lui ait jamais donné la définition propre de cet animal.

Théoriquement, le programme engendre un certain nombre d'informations via une banque de données initiale, constituant son socle de connaissance et lui permettant par la suite de prédire une réponse adaptée à la question posée.
Il existe classiquement deux manières de faire du machine learning.

3.1.1 L'apprentissage supervisé

Celui-ci repose sur trois piliers :

- Une banque de données, comme par exemple une banque d'images.
- Un label, c'est-à-dire une description des données. Pour le cas d'une banque d'images il peut s'agir de leur description et leur séparation en différentes catégories.
- Un algorithme à qui ces données seront soumises.

Si la variable d'intérêt est l'appartenance à une classe (présence ou non d'une pathologie, stade tumoral, etc...), comme dans le cas de la reconnaissance d'images, on fera de la classification.
Par exemple, le modèle sera entraîné en visualisant des images de singes étiquetées "singe" et des images de dragons étiquetées "dragon". Après quoi, il devra prédire, en classifiant de nouvelles images encore inconnues, s'il s'agit de singes ou de dragons.

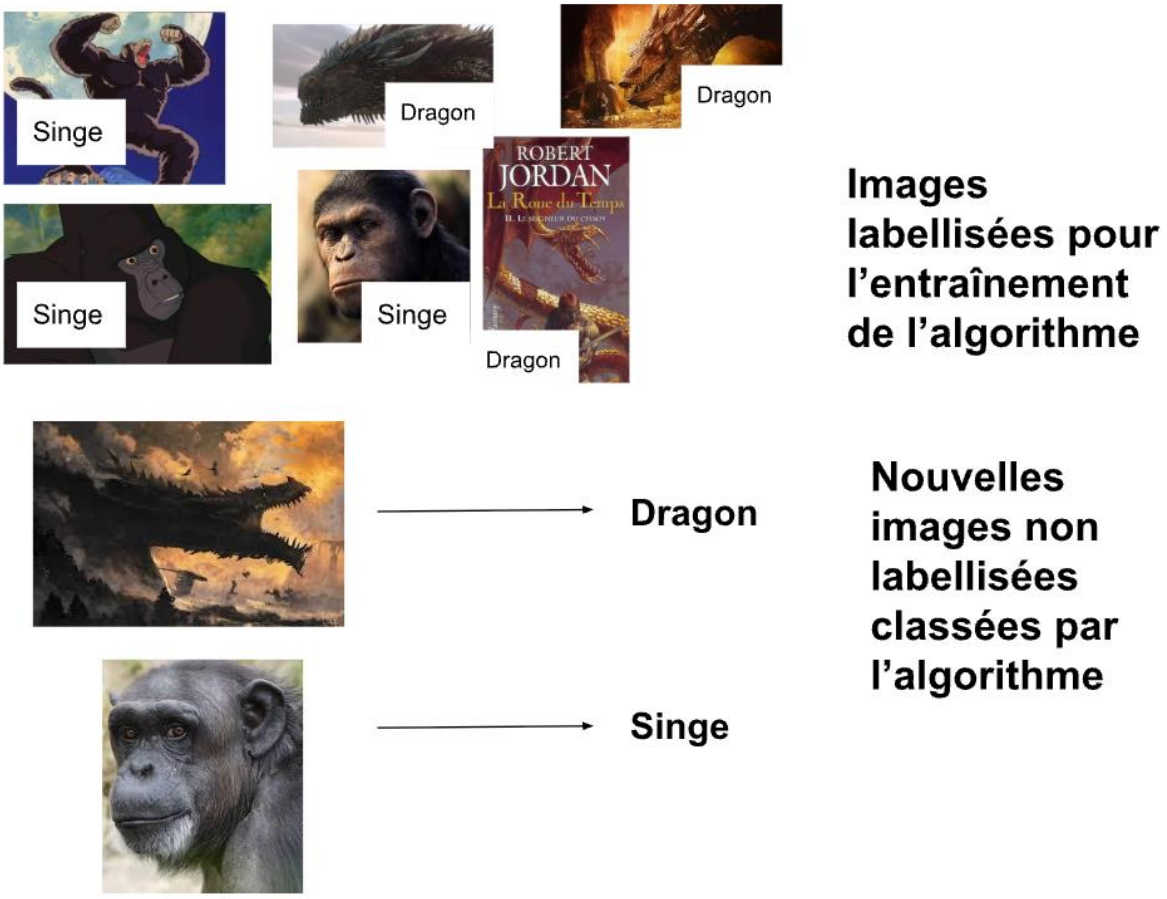


Figure 3 Exemple de classification d'images labellisées selon en mode prédictif. Le modèle classe des images de singe et de dragon inconnues et non labellisées après avoir appris les concepts de singe et de dragon sur des images labellisées pendant son entraînement.

L'apprentissage supervisé est de loin le plus utilisé en médecine, notamment en radiologie, pour entraîner des algorithmes de reconnaissance d'images. Il nécessite donc une base de données labellisées, chaque image portant une étiquette de classification.

En pratique, les données sont divisées en trois sets (splits) au moment de l'entraînement :

- Un set d'entraînement, qui permettra à l'algorithme de commencer à reconnaître les deux catégories étudiées. Par exemple, une radiographie de genou avec une lésion osseuse maligne et des radiographies de genou

normales. Ce set devra être le plus large et le plus exhaustif possible afin que l'algorithme puisse engranger un maximum d'informations.

- Un set de validation, qui sera le premier test de l'algorithme dans lequel on évaluera son efficacité. Si l'évaluation n'est pas satisfaisante, on retournera au set d'entraînement afin d'améliorer l'algorithme, autant de fois que l'on jugera nécessaire.
- Un set de test, qui sera l'évaluation finale et unique de l'algorithme, reflétant alors ses performances réelles.

3.1.2 L'apprentissage non supervisé

Dans ce cas, on utilise un modèle descriptif qui utilise des données non labellisées, en faisant du clustering ou regroupement. C'est-à-dire que l'on entraîne un algorithme à regrouper des entités similaires en groupe, sans a priori sur les caractéristiques explicites de ces groupes. L'algorithme pourrait alors regrouper les images sur lesquelles figurent des singes dans un groupe 1 et des images sur lesquelles figurent des dragons dans un groupe 2, sans savoir qu'il s'agit de singes et de dragons. Contrairement à l'apprentissage supervisé dans lequel les images étaient étiquetées au préalable, il devrait alors, par lui-même, apprendre à reconnaître les caractéristiques communes des singes (poils, quatre pattes,...) et celles des dragons (ailes, écailles,...) afin de pouvoir les séparer en deux groupes distincts.

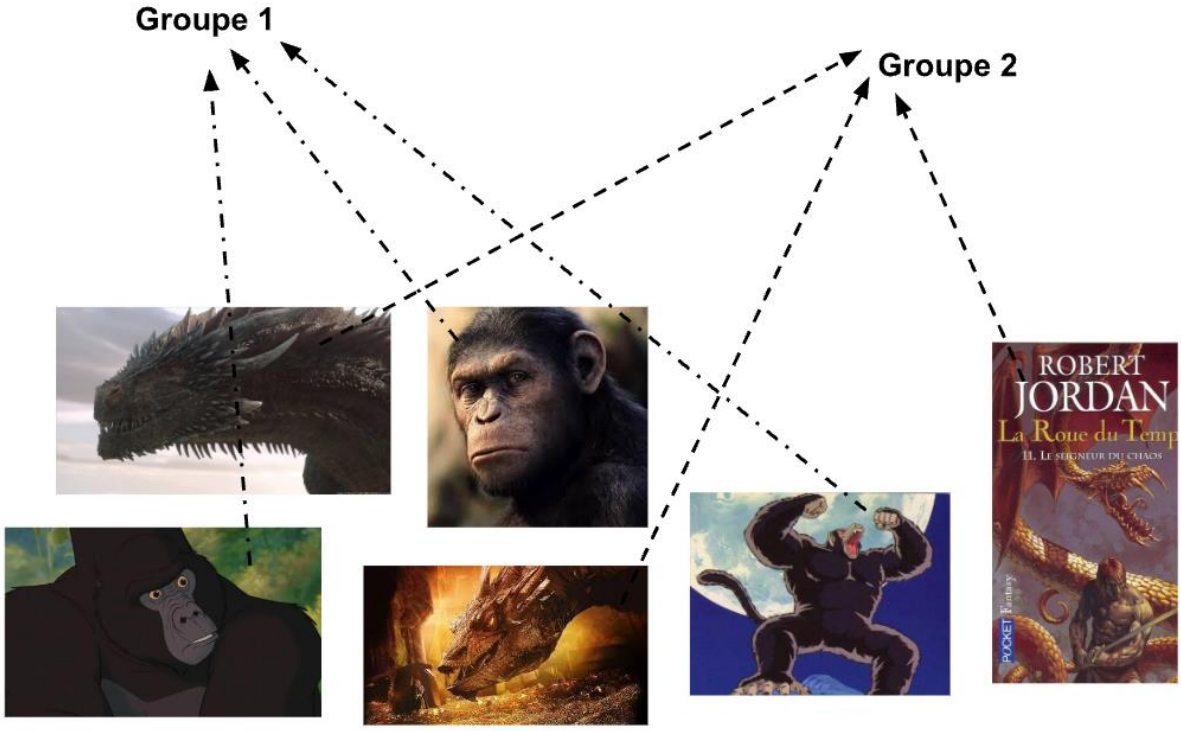


Figure 4 Exemple de clustering d'images non labellisées selon en mode descriptif. Le modèle regroupe les images possédant des caractéristiques similaires en groupe, sans que celles-ci soient étiquetées au préalable.

L'apprentissage supervisé, utilisé par exemple en radiologie pour la reconnaissance d'images, est donc un processus coûteux. En effet, la qualité de l'algorithme dépend en grande partie de la qualité et de la variété des données sur lesquelles il a été entraîné. En particulier la labellisation de ces images qui est très chronophage car elle nécessite d'annoter chacune d'entre elles à la main. La qualité de ces annotations est au cœur de la qualité de l'algorithme final et dépend en grande partie des compétences de la personne ayant réalisé cette tâche. De plus, si les images proviennent d'un unique centre, il se pose alors la question de l'applicabilité à d'autres centres de l'algorithme développé.

Ces méthodes d'apprentissages reposent généralement sur des réseaux de neurones, dont le fonctionnement permet de comprendre le fonctionnement des LLMs.

3.2 Neurones artificiels

Un neurone biologique est composé de dendrites, des prolongements se connectant aux neurones voisins d'amont et permettant l'arrivée d'informations "d'entrées". Le corps cellulaire, partie centrale, traite l'information en intégrant les signaux d'entrée reçus. Enfin les axones, prolongements distaux du neurone, servent à transmettre l'information aux neurones voisins d'aval, ce sont les "sorties".

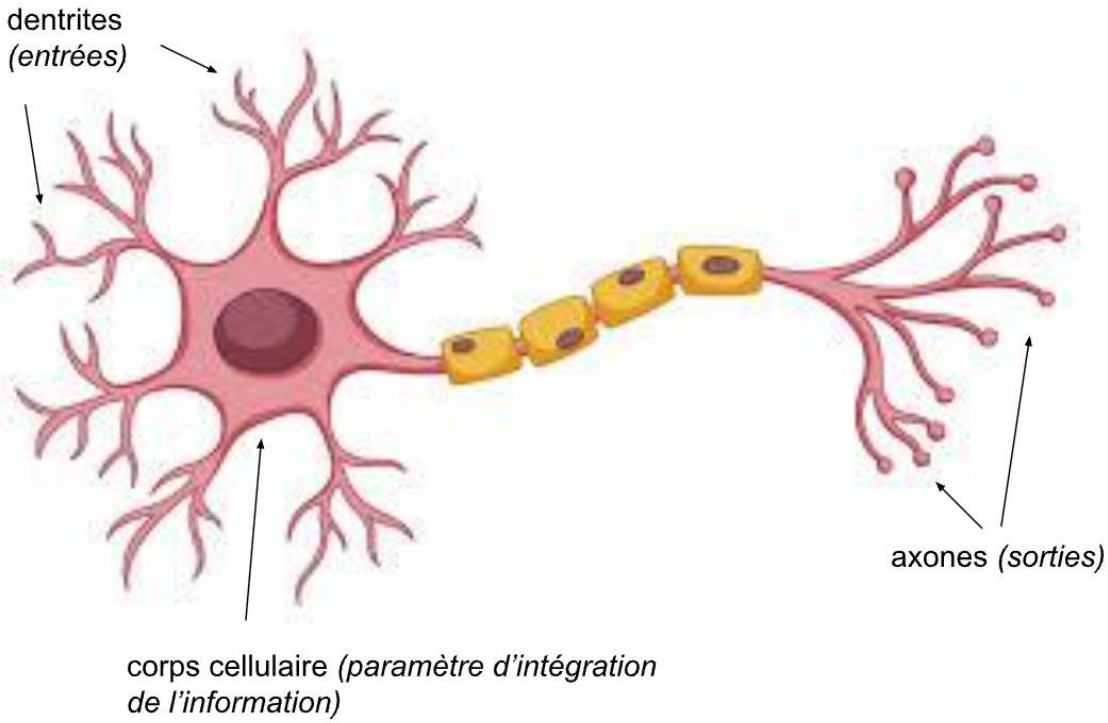


Figure 5 Représentation schématique d'un neurone biologique.

Les neurones artificiels ont pour but de mimer ce fonctionnement en le transposant de manière mathématique. Chaque neurone traite les informations qu'il reçoit via ses voies d'entrée et produit un signal de réponse via ses sorties.
Tous ces neurones sont liés les uns aux autres, selon une architecture faite de plusieurs couches. A l'échelle du réseau, les différentes couches de neurones interconnectées traitent les informations d'entrée et produisent une sortie dont le signal sera le produit des calculs opérés à chaque couche, par chaque neurone. Ces calculs sont dépendants de la force de connexion des synapses artificiels, ou poids. Ce processus est dynamique, permettant ainsi l'apprentissage du réseau composant l'algorithme. C'est en somme un système infiniment flexible, initialement sans but, mais dont les modifications au cours de l'apprentissage vont permettre l'adaptation à une tâche arbitraire.

Si l'on reprend un exemple de reconnaissance d'images en radiologie, on peut imaginer un réseau de neurones dont la fonction serait de déterminer si une tumeur est maligne ou bénigne sur une radiographie. Chaque pixel composant la radiographie (adapté pour l'algorithme en résolution de 256 x 256) sera équivalent à une voie d'entrée. L'ensemble de ces entrées sera traité par le réseau composé de milliards de neurones disposés en couches, qui produira à son tour, après traitement, un signal de sortie pour prédire une réponse : bénin ou malin. Cette réponse sera ensuite corrélée à la réalité. Si elle est juste, l'algorithme retiendra que ses calculs étaient corrects et renforcera sa conformation actuelle. En revanche, si la réponse est incorrecte, l'algorithme modulera ses paramètres à la prochaine tentative, pour *in fine* trouver la conformation la plus en adéquation avec la réalité.

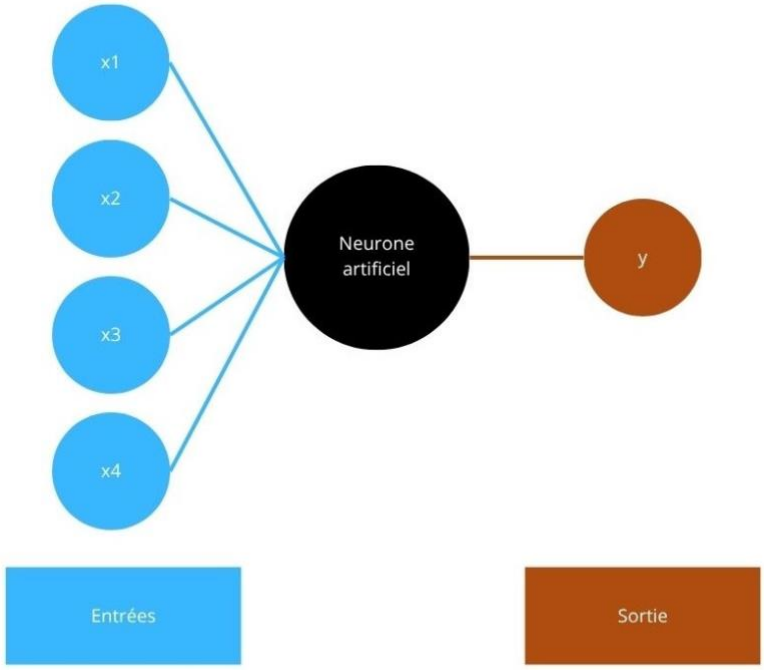


Figure 6 Représentation schématique d'un neurone artificiel (correspondant à l'un des paramètres d'un algorithme de deep learning). Il reçoit des informations d'entrée appelées "x", les traite, et produit un signal de sortie "y".

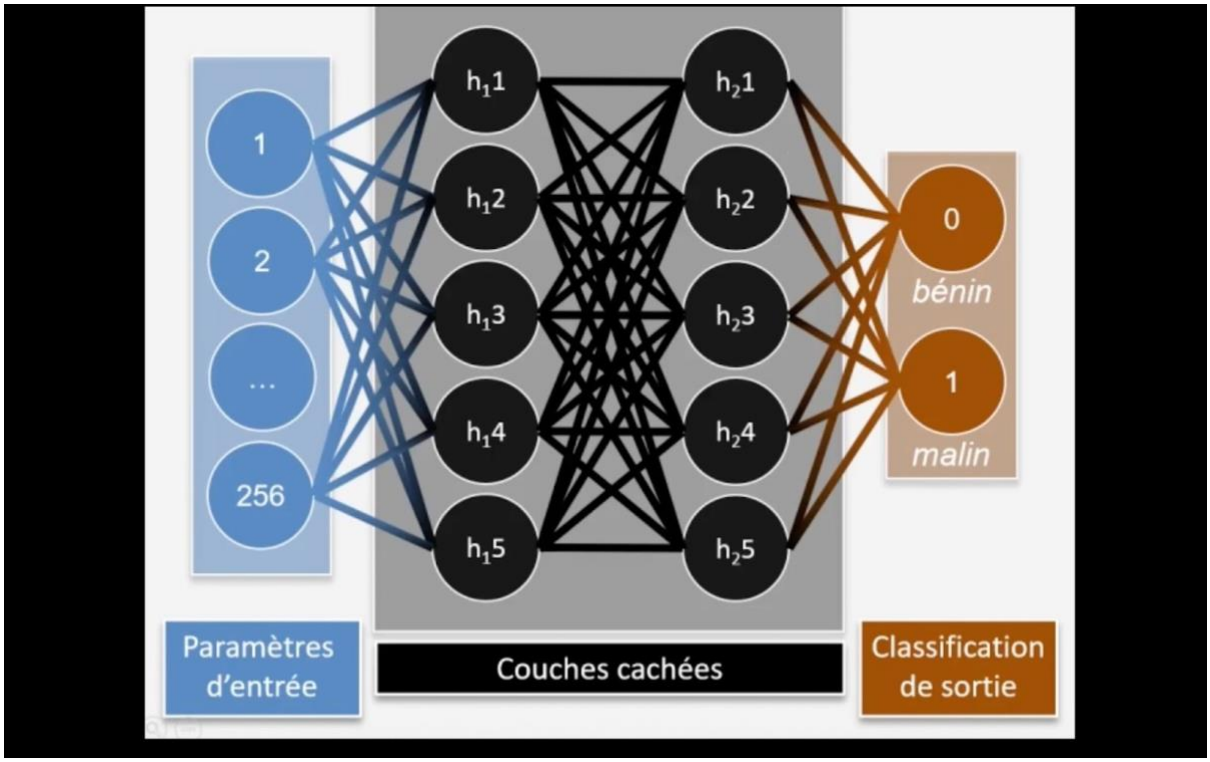


Figure 7 Représentation schématique d'un réseau de neurones artificiels profonds d'un algorithme de deep learning prédisant si une tumeur osseuse est maligne ou bénigne en radiographie. Les pixels de la radiographie sont les paramètres d'entrées représentés en bleu. Les "neurones" ou paramètres de l'algorithme sont représentés en noir, disposés en plusieurs couches et interagissent les uns avec les autres. La réponse "Bénin" ou "Malin", représentée en marron, correspond aux paramètres de sorties prédits, issue de l'ensemble des signaux émis par le réseau de neurone. © Loïc Duron.

Comme nous l'avons vu, l'entraînement supervisé d'un réseau de neurones pour faire de la reconnaissance d'images est très vorace en termes de quantité d'images et de labels de qualité.

Au contraire, la force des LLMs est inhérente à l'essence même de l'entraînement non supervisé. Lors de cet entraînement, la base de données des LLMs est une grande quantité de textes. En occultant un à un des mots du texte, ils s'entraînent à les retrouver compte tenu du texte précédent. Une base de données non labellisées peut donc leur servir de base d'entraînement, sans intervention humaine.

Par exemple, en intégrant de grandes quantités de textes à sa base de données, un algorithme pourrait apprendre que le soleil se lève à l'Est et se couche à l'Ouest, simplement en apprenant la fréquence d'apparition de ces mots dans les mêmes phrases. Au cours de son entraînement, il pourrait rencontrer la phrase : "Le soleil se couche à l'Ouest". Alors, il pourrait occulter le mot "Ouest" et essayer de le deviner. S'il prédisait "Nord", il s'apercevrait que sa réponse est erronée et que le bon mot était "Ouest". Il devrait alors modifier la conformation de ses paramètres qui avaient prédit "Nord" comme réponse, vers une nouvelle conformation où "Ouest" est la bonne réponse. Ce processus d'apprentissage non supervisé, ou plus précisément auto-supervisé puisque le modèle crée lui-même ses labels (mots occultés), contrairement à l'apprentissage supervisé, ne nécessite aucune annotation de la part de l'homme puisque la réponse se trouve déjà dans le texte donné au modèle. Cela permet ainsi de s'affranchir de l'aspect fastidieux de la labellisation.

Ainsi, grâce au réseau de neurones profonds, le modèle va petit à petit, en consultant les très grandes quantités de textes mises à sa disposition, enregistrer les subtilités et complexités de notre langage, dans le but ultime de produire un texte concordant avec la situation qui lui sera exposée. Son but ne sera pas ici de prédire la nature d'une image, mais de prédire "le prochain mot" correct afin de produire un texte ayant du sens avec une situation donnée.

3.3 Production de texte selon un modèle de probabilité

Afin de comprendre comment le modèle produit du texte de manière adaptée à la situation qui lui est donnée, il est nécessaire de détailler différents concepts.

3.3.1 La tokenisation

Le processus de tokenisation, ou segmentation du texte, est l'un des concepts de base du fonctionnement des LLMs. Les tokens sont définis comme des unités de texte, pouvant correspondre à des mots entiers, des parties de mots, des caractères ou même des symboles. Segmenter ainsi le texte est la première étape permettant au modèle de l'analyser, c'est la façon dont il déchiffre le texte.

Si on prend la phrase "les feuilles tombent", elle pourra être segmentée par exemple en trois tokens "Les", "feuilles", et "tombent".

Figure 8 Processus de segmentation des phrases de ChatGPT (Tokenisation).

3.3.2 Plongement lexical

Le modèle transforme ensuite chaque unité de texte (token) en vecteur, c'est le processus d'embedding ou de plongement lexical. L'embedding est en quelque sorte la transformation d'un langage sémantique en langage mathématique. Chaque vecteur est alors placé dans un espace à haute dimension appelé espace vectoriel. Cet espace permettra par la suite d'établir des relations entre les mots. Plus deux

mots sont proches dans notre langage, plus ils seront proches dans l'espace vectoriel.

Si on reprend par exemple la phrase "les feuilles tombent des" et que l'on demande au modèle de la compléter, celui-ci pourrait répondre "arbres". En effet, "feuille" et "arbre" sont des mots qui ont un sens proche, donc proches dans l'espace vectoriel. Le modèle en déduit donc que le mot "arbre" à la plus grande probabilité d'être le mot juste pour compléter la phrase dans le contexte.

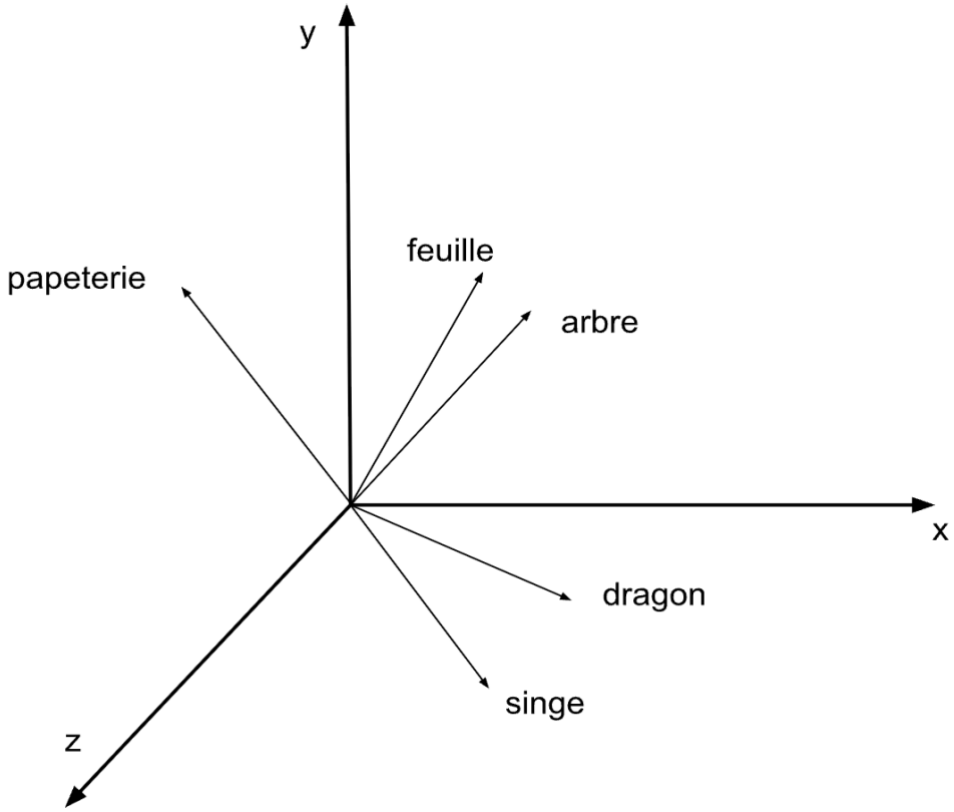


Figure 9 Représentation simplifiée de l'espace vectoriel tridimensionnel dans lequel les mots "feuille" et "arbre" ont des vecteurs proches. Les mots "dragon" et "singe" en sont éloignés et il existe une certaine proximité entre eux du fait qu'ils évoquent chacun des créatures vivantes animales. Le mot "papeterie" se trouve éloigné de tous ces mots car il appartient à un champ lexical différent.

3.3.3 Le processus d'attention

Les mots pouvant également avoir un sens différent selon la manière dont on les emploie, il est nécessaire d'introduire un troisième concept, se trouvant au cœur même du succès des LLMs, le processus d'attention [30].

En effet, si notre modèle place les mots entre eux dans l'espace vectoriel en fonction de leurs proximités sémantiques brutes, les subtilités de notre langage rendent le concept de plongement lexical insuffisant à lui seul pour produire du texte de manière adaptée, car celui-ci ne permet pas de s'adapter à un changement subtil de contexte.

Si on reprend l'exemple du mot "feuille" dans la phrase "les feuilles tombent des arbres", on comprend selon le contexte que le mot "feuille" a une position proche du mot "plante" dans l'espace vectoriel.

Cependant, un même mot peut avoir plusieurs sens selon le contexte qui l'accompagne. C'est ici qu'interviennent deux nouveaux types de vecteurs, dit modificateurs : les vecteurs de clé et les vecteurs de requête. Les vecteurs de clé codent le rôle des mots ayant potentiellement un rôle modificateur dans la phrase et les vecteurs de requête la disposition des mots de la phrase à être modifiés.

Avec la phrase "les feuilles tombent des arbres", les mots "tomber" et "arbre" donnent un certain sens au mot "feuille" via des vecteurs de clé. Le mot "feuille" a quant à lui une certaine propension à prendre un nouveau sens via son propre vecteur de requête. Le processus d'attention est une discussion entre les différents mots de la phrase, au cours de laquelle le sens de chaque mot va se préciser en fonction de ses relations avec les autres mots.

Ainsi, si on demande au modèle de compléter la phrase : "j'écris sur des feuilles" le modèle complètera la phrase par "de papier" et non "d'arbre". En effet, les vecteurs

de clé du verbe “écrire” ont modifié le sens du mot ”feuille” qui avait lui-même une certaine probabilité de prendre un nouveau sens via son vecteur de requête. Finalement, le vecteur du mot “feuille” s’est déplacé dans l’espace en s’éloignant de la position du vecteur “plante” vers le vecteur “papeterie”. Le succès de ce concept repose donc sur la dimension dynamique des vecteurs.

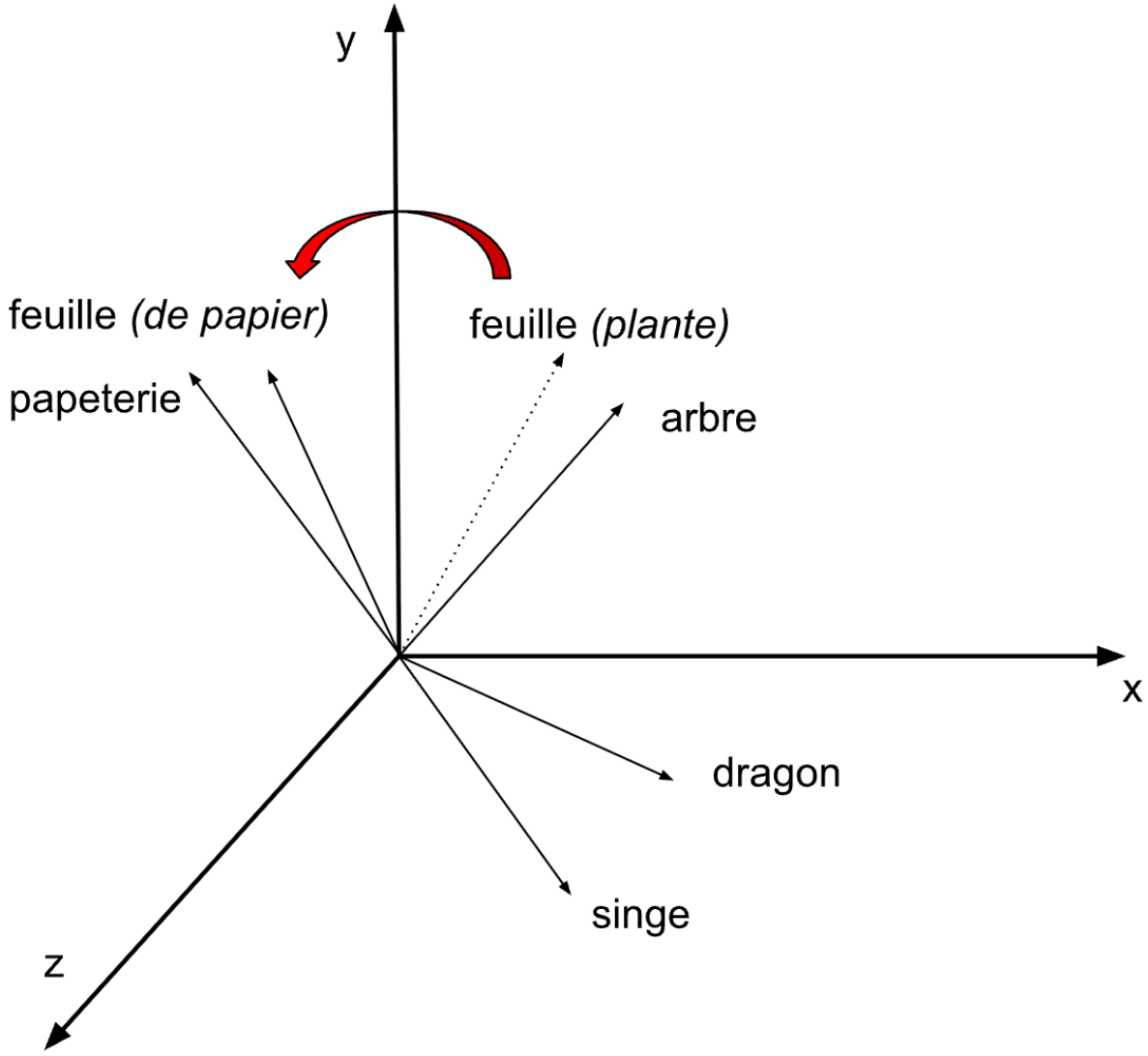


Figure 10 Représentation simplifiée du déplacement du vecteur “feuille” vers le vecteur “papeterie” depuis sa position initiale auprès du vecteur “arbre”. La flèche rouge représente le rôle des vecteurs de clé et de requête ayant permis la traduction mathématique du changement de signification du mot “feuille” en fonction du contexte de la phrase.

Lorsque l’on converse avec un LLM, il utilise les processus de tokenisation, de plongement lexical et d’attention afin d’analyser et d’émettre une réponse ayant un sens adapté, pas à pas, selon un mode probabiliste dynamique.

4 Potentiel et limites des LLMs pour l’aide à la décision médicale

Ces dernières années, l’IA sous toutes ses formes et notamment les LLMs sont devenus un sujet incontournable en matière d’innovation dans le milieu médical. Ces modèles ayant accès à une quantité phénoménale d’informations provenant des grandes quantités de textes constituant leur base de données et pouvant facilement les restituer, leur intérêt en médecine a rapidement suscité un engouement scientifique.

C’est notamment le cas de l’aide au diagnostic et à la prescription d’examen d’imagerie. Les connaissances médicales actuelles se développant constamment, il serait pertinent d’avoir un outil permettant de les stocker et de les restituer facilement. Les LLMs pourraient ainsi constituer une base de connaissances complexes et diverses, facilement restituables à la demande des médecins, dans le but de faciliter et d’améliorer la prise en charge des patients.

La qualité des informations fournies par ces modèles dépend cependant de leur base de données et de la façon dont ils les restituent. De ce fait, leur utilisation en pratique courante se heurte à certaines limites concernant les informations fournies, notamment la véracité, l'exhaustivité, l’accessibilité, mais aussi la concordance avec les pratiques médicales du territoire sur lequel elles vont être utilisées.

4.1 Aide au diagnostic et à la pratique médicale courante

L'apprentissage et la pratique de la médecine nécessitent une base de connaissances théoriques poussée dans de nombreux domaines. Cette discipline ne cessant d'évoluer, il appartient aux médecins de maintenir un niveau de connaissances constant, en adéquation avec les avancées scientifiques permanentes et bien sûr de pouvoir les restituer rapidement en fonction de la situation clinique donnée. Cet apprentissage continu constitue un exercice complexe, que chaque praticien doit effectuer tout au long de sa carrière.
Les bases de données des LLMs étant constituées d'innombrables quantités de textes traitant de domaines divers et variés, il apparaît qu'ils pourraient constituer de bonnes ressources afin de mobiliser ces connaissances à la demande des médecins.

En autonomie, ces modèles ont déjà fait leurs preuves et démontré une certaine aptitude en matière de restitution de connaissances médicales brutes. L'étude de Kung et al de 2023 [31] a en effet démontré que, sans entraînement particulier, ChatGPT était capable de passer l'Examen de Licence Médicale des Etats-Unis (USMLE), nécessaire pour exercer la médecine dans ce pays, en obtenant respectivement 60 %, 65 % et 61 % de bonnes réponses aux trois épreuves dans lesquelles le seuil de validation est fixé à 60 %. Ces résultats prometteurs suggèrent que les LLMs pourraient constituer des outils pertinents pour la restitution de connaissances médicales standardisées.

Cependant, si les connaissances théoriques des médecins sont indispensables à la bonne pratique médicale, il arrive que certaines situations réelles soient parfois

moins conventionnelles. Le jugement personnel, l'expérience et l'avis d'autres confrères sont alors d'une aide précieuse, afin d'avancer dans le diagnostic et la prise en charge du patient. Dans ce contexte, l'étude de Kanjee et al de 2023 [32] a évalué GPT-4 sur 70 cas cliniques complexes issus des conférences clinico-pathologiques du New England Journal of Medicine. Dans cette série de cas, celui-ci a montré des performances prometteuses en ayant le bon diagnostic dans 39 % des cas, sinon le bon diagnostic dans sa liste de diagnostics différentiels dans 64 % des cas.

Plus récemment, le concept de "raisonnement" et un nouveau système "d'orchestration" ont permis d'obtenir des résultats encore plus surprenants. Contrairement aux modèles plus anciens qui fournissent leur réponse pas à pas, simultanément à leur processus de réflexion, les modèles utilisant le raisonnement procèdent d'abord par une étape de réflexion afin d'affiner et d'optimiser leur analyse avant de formuler leur réponse.

Le concept d'orchestration permet de répondre à des problématiques plus complexes qu'un modèle généraliste classique travaillant seul à résoudre une tâche. L'agent orchestrateur décompose la problématique donnée en plusieurs sous-tâches spécifiques et les répartit entre différents modèles spécialisés qu'il coordonne. Contrairement aux études suscitées utilisant des formats type questionnaire à choix multiple ou cas cliniques bruts, dans l'étude de Nori et al de 2025 [33], 304 cas tirés du New England Journal of Medicine ont été transformés en "cas interactifs", un format dans lequel les cas sont exposés par étapes et dans lequel les modèles et les médecins ont la possibilité de poser des questions et de prescrire des examens itératifs, afin de simuler une situation plus proche de la réalité.

Six grands modèles généralistes ont été évalués, chacun sur des versions plus antérieures et sur des versions plus récentes. Ces versions diffèrent notamment par la façon dont les réponses sont générées, les plus récentes utilisant le raisonnement. Ainsi, un modèle ancien tel que GPT 3.5 Turbo de 2020 (Open AI) obtient une précision diagnostique de seulement 25 % contre 80 % pour le modèle o3 de 2025 (Open AI), un modèle utilisant le “raisonnement”.
De plus, ces modèles ont été évalués en association avec Microsoft AI Diagnostic Orchestrator (MAI-DxO), un système capable de s’adapter à un LLM en simulant un panel de différents médecins appliquant diverses approches cliniques et collaborant à la résolution des cas, permettant encore une fois de créer des conditions plus proches de la vie réelle. Ce système, en s’associant avec o3, a permis d’obtenir une précision diagnostique de 85,5 %, avec une approche permettant également de réduire de manière considérable les coûts financiers théoriques liés aux diagnostics de ces cas. En comparaison, 21 médecins ont été évalués et ont obtenu des scores de précision diagnostique de 20 %.

Ainsi, ces modèles pourraient constituer une bonne aide au diagnostic dans des situations réelles, qui plus est d’un certain degré de complexité, en augmentant les performances des médecins.

4.2 Aide à la prescription d’examens d’imagerie

Cette capacité à restituer des connaissances et à suivre des recommandations de bonnes pratiques soulève un point important dans le contexte des difficultés rencontrées actuellement par les prescripteurs à être pertinents dans le choix de la

modalité d'imagerie à employer, découlant de la multiplicité de recommandations de prescription et de sociétés savantes les promulguant. Les LLMs pourraient alors devenir des outils d'aide à la prescription d'examen d'imagerie.

Dans ce contexte, certaines études ont d'ores et déjà évaluées en radiologie la capacité des LLMs à fournir une réponse adéquate quant à la prescription d'un examen d'imagerie selon un contexte clinique donné.
C'est le cas de l'étude de Zaki et al. de 2024 [34] qui a comparé la capacité de deux LLMs (ChatGPT et Glass AI) à recommander l'examen d'imagerie le plus approprié selon les critères de l'ACR sur 1075 cas cliniques couvrant 11 sur-spécialités de la radiologie. Les résultats obtenus montraient une supériorité de Glass sur ChatGPT avec toutefois une certaine hétérogénéité dans leurs recommandations.

Dans le contexte plus spécifique de la neuro-imagerie, l'étude de Miller et al de 2024 [35] a évalué la capacité de 8 LLMs à fournir l'examen d'imagerie le plus pertinent dans 24 situations cliniques fictives selon les critères de l'ACR et de la NOC. Trois neuro-radiologues ont ensuite évalué leurs réponses selon des critères de précision, d'utilité et de concision/citation des sources. Des modèles tels que GPT-4, ChatGPT, Claude et Perplexity ont obtenu de bons voir d'excellents résultats avec respectivement 23/24, 20/24 et 19/24 ex-aequo. Cette étude a cependant mis en évidence plusieurs types d'erreurs de la part des LLMs, comme la proposition d'examen non recommandé, l'absence de proposition d'examen à tort ou encore la proposition d'examen incohérent avec le contexte, voire irréel. C'était notamment le cas du modèle Llama 2 qui avait inventé et proposé la réalisation d'une

“otosialographie” au lieu d’un scanner sans injection des rochers, devant un cas de surdité de transmission sans anomalie à l’examen otologique.

Sans modification, les LLMs génèrent des recommandations “de tête”, c’est à dire à la force de leur seule mémorisation des textes croisés issus de leur entraînement. Rau et al [36] ont montré en utilisant 50 cas cliniques, que GPT 3.5 Turbo était encore plus performant pour aider la décision clinique une fois transformé en modèle contextuel grâce à LlamaIndex, un système permettant de connecter un LLM à une base de données externe. Grâce à son accès direct aux recommandations, il surpassait des LLMs génériques et même des radiologues pour recommander le bon examen d’imagerie à réaliser.

On pourrait alors imaginer qu’il serait plus intéressant d’utiliser des modèles biomédicaux spécifiques (MedAlpaca, MedGemma, etc.) pour aider à la décision médicale. Cependant, il est aujourd’hui bien établi que ces modèles sont nettement surpassés par les modèles généralistes. Dorfner et al [37] ont comparé un grand modèle générique (Llama 3) et des modèles biomédicaux contextuels. Cette étude montre qu’il est plus judicieux d’utiliser un grand modèle généraliste plutôt que d’entraîner un nouveau modèle pour une tâche particulière. Celui-ci possède en effet, par une technologie basée sur des algorithmes de pointe, de meilleures capacités d’apprentissage et surtout de réflexion. Au contraire, un modèle biomédical spécifique, utilisant une technologie moins performante, reste moins efficace pour suivre des consignes, malgré une base de données explicitement dédiée à une tâche.

Malgré ces résultats, bons nombres d'études ont mis en lumière certaines limites de ces modèles et ont soulevé de nombreuses problématiques récurrentes, rendant leur utilisation encore précaire, dans le milieu médical notamment.

4.3 Les limites des LLMs

Pour être utilisé dans le milieu médical, un système d'aide à la décision doit pouvoir être capable de fournir de manière transparente des informations valides, en adéquation avec les recommandations actuelles et les pratiques locales, dans un contexte donné. Or, les LLMs ont plusieurs limites pour l'aide à la prescription des examens d'imagerie.

Certaines d'entre elles sont fonction du type de modèle utilisé. En effet, il faut différencier deux grandes catégories de LLMs.
Les modèles à code source fermé sont des modèles commerciaux tels que ChatGPT d'OpenAI. Ces modèles reposent sur des algorithmes inaccessibles au public, ils sont donc non téléchargeables sur poste informatique et ne seraient de toute façon pas utilisables sur des ordinateurs courants en raison de la puissance informatique nécessaire à leur usage : 10^{12} paramètres environ pour un LLM à code source fermé contre 10^{9} paramètres pour un serveur hospitalier classique, soit un rapport de 1000. Tous les calculs sont réalisés sur les serveurs propres de ces sociétés. En plus du problème de transparence lié à l'inaccessibilité des données de ces modèles, ils sont également payants ce qui engendrerait des frais non négligeables aux structures médicales qui les emploieraient. De plus, un problème de confidentialité, de loin le plus important quant à l'utilisation de ces modèles, se poserait dans le milieu de la

santé car leur utilisation nécessiterait de divulguer des données médicales confidentielles sur ces serveurs commerciaux.
Par opposition aux modèles fermés, il existe des modèles ouverts, comme DeepSeek, ou certains modèles de Mistral. Leurs algorithmes sont accessibles à tous et ils sont en théorie téléchargeables sur serveur personnel, si toutefois on en possède un capable de les supporter (10^{10} paramètres), permettant ainsi de s'affranchir de la problématique de confidentialité des données. En revanche, leur corpus d'entraînement n'est pas connu, c'est simplement le produit de l'entraînement (l'algorithme et les paramètres finaux) qui est ouvert, ne permettant pas de régler le problème de véracité de leurs informations.

D'autres phénomènes limitant, tels que l'hallucination, sont fréquemment retrouvés [38]. Celui-ci se définit comme une situation dans laquelle un modèle génère une réponse fausse, inventée ou trompeuse, pourtant formulée de manière convaincante et crédible. Il en existe de plusieurs types, tels que les hallucinations factuelles lorsque les données de la réponse sont erronées, les hallucinations de citation lorsque les références citées sont inexistantes ou encore des hallucinations logiques lorsque la réponse est totalement incohérente avec le raisonnement.
Il arrive aussi parfois que les réponses soient obsolètes. Cela peut arriver lorsque de nouvelles informations plus récentes n'ont pas été intégrées aux données d'entraînement du LLM [39].
Dans le cas de la recherche bibliographique, il arrive aussi fréquemment que le modèle n'arrive pas à citer ses sources, rendant alors les informations fournies difficilement utilisables et ne permettant pas de vérifier leur véracité. [38,39]

Il apparaîtrait également que les réponses fournies par les LLMs seraient influencées par des connaissances et des raisonnements orientés vers un mode de pensée américain [40]. Ce biais de nature géographique est dû au fait que les sources sur lesquelles les modèles sont entraînés sont composées en grande partie de textes découlant de la littérature américaine.

Dans le contexte plus particulier de la restitution d'informations, notamment dans le cadre de la recommandation de bonnes pratiques, il est primordial que le modèle utilisé fournisse des informations en adéquation avec les pratiques locales.

Ces pratiques sont établies en fonction des recommandations émises par différentes sociétés savantes, qui ne possèdent pas toutes la même légitimité ou autorité selon le territoire concerné. En effet, elles sont élaborées en fonction des besoins locaux selon des critères de rendement médical, de logistique intra-hospitalière ou de coût, mais doivent également respecter les lois en vigueur, relevant alors d'obligations de nature juridique.

Afin de poursuivre l'évaluation des modèles pour l'aide à la prescription d'examen d'imagerie, il apparaît nécessaire de rechercher ce biais géographique dans leurs réponses et notamment d'apprécier une éventuelle tendance en faveur de recommandations émanant de sociétés savantes américaines.

5 Objectif

Dans le contexte actuel du développement important de la radiologie et de l'augmentation croissante des besoins en imagerie médicale, l'articulation de la prise en charge des patients entre spécialistes et la communication entre radiologues et médecins prescripteurs est de plus en plus difficile. C'est le cas des difficultés rencontrées par les médecins prescripteurs à être constamment pertinents dans le choix de la modalité d'imagerie à prescrire, découlant de la multiplicité des recommandations de bonnes pratiques pour la prescription d'examens d'imagerie, pourtant initialement créées dans le but de les aider. Dans ce contexte, les LLMs apparaissent comme une solution pertinente et permettraient, en aidant la prise de décision pour la prescription d'examens d'imagerie, de fluidifier ces interactions.

Si la fiabilité des modèles pour l'aide à la prescription d'examens a déjà été étudiée, leur capacité à aider à la prescription d'examens en neuro-imagerie dans d'autres contextes que celui des Etats-Unis n'est pas connue. En particulier, leur entraînement sur des bases de données majoritairement issues des Etats-Unis peut laisser craindre un biais en faveur des recommandations de ce pays, qui serait dangereux pour les médecins pratiquant dans d'autres contextes.

Le but de notre étude est de voir s'il existe une influence géographique dans le choix de la recommandation utilisée par les LLMs pour suggérer un examen d'imagerie dans un contexte clinique donné, notamment au profit des recommandations de sociétés savantes américaines.

Article en Anglais

1 Abstract

1.1 Objectives

Large language models (LLMs) are increasingly explored as decision-support tools in medical imaging. However, their ability to align with country-specific guidelines, which often diverge, remains uncertain. We set out to evaluate the geographic neutrality of three state-of-the-art LLMs, GPT-o3 (OpenAI), Mistral Large (Mistral AI), and DeepSeek R1 (DeepSeek), when applied to neuroradiology scenarios with conflicting U.S. and non-U.S. recommendations.

1.2 Materials and Methods

Vignettes derived from contradictory international guidelines were presented to each model under two conditions: an implicit setting, where no guideline was specified and vignettes were provided in English and French; and an explicit setting, where prompts directed models to follow a named guideline. Performance was reviewed against the target guideline, and mitigation strategies were tested.

1.3 Results

Thirty clinical vignettes presenting conflicting guidelines were evaluated by GPT-o3, Mistral Large, DeepSeek R1. In the implicit setting, all models favored U.S.

guidelines, with GPT-o3, Mistral, and DeepSeek aligning with them in 27 of 30 scenarios (90.0%; 95%CI, 74.4–96.5). In the explicit setting, adherence declined sharply for non-U.S. recommendations for all models. Providing the complete guideline text was the most effective mitigation strategy, restoring accuracies above 90% across all models.

1.4 Conclusion

Across languages and model origins, LLMs exhibited a systematic bias toward U.S. neuroradiology guidelines, even when explicitly instructed otherwise. This U.S.-centrism likely reflects training data imbalances and raises concerns for safe global deployment. Strategies for local contextualisation, such as guideline integration at deployment, are necessary to ensure context-appropriate clinical decision support.

1.5 Keywords

Artificial intelligence; ChatGPT; DeepSeek ; Mistral ; Bias ; Large language models

1.6 List of abbreviations

LLM: Large language model

1.7 Key points

Question: Do large language models display geographical neutrality in neuroradiology decision support?

Findings: Even models developed in France and China systematically preferred United States guidelines, aligning with them in most implicit scenarios while failing to follow explicit guidelines from other sources.

Clinical Relevance: This systematic United States-centric bias poses clinical and legal risks for global deployment. Safe implementation requires specific localization strategies, such as providing full guideline texts, to ensure recommendations align with local practice standards.

2 Introduction

Medical imaging is now deployed across an ever-growing range of highly specialised fields, making its use increasingly complex[1]. Because imaging indications are so numerous, clinicians can struggle to order the most appropriate study, and over-utilization of imaging modalities is common[1]. Although professional-society guidelines are regularly issued to guide ordering behavior, their sheer number and frequent updates can make them difficult to navigate[2–4]. Moreover, guidelines issued by societies from different countries can conflict, especially in neuroradiology [5]. Comparative analyses of referral guidelines show only slight concordance between recommendations for common scenarios such as cervical spine imaging [6]. As such, clinicians need a decision-support tool that is adapted to their specific practice context.

Large language models (LLMs) seem promising in this respect, as they can encode extensive clinical knowledge and deploy it to support medical decision-making[7, 8]. In neuroradiology, they have already been tested as imaging decision aids[9].

However, recent studies have uncovered biases in the way LLMs handle medical questions, propagating gender and racial stereotypes[10, 11]. Additionally, several studies report a tendency of LLMs to reproduce U.S.-centric values[12–14]. This bias towards American content may generate inappropriate responses with respect to the practice context of the user, with potential medical and legal consequences. To the best of our knowledge, no study has investigated the existence of geographical bias of LLMs for clinical decision support.

We set out to examine whether three large language models, GPT-o3 (OpenAI), DeepSeek R1 (DeepSeek), and Mistral Large (Mistral AI), each developed on a different continent, display U.S.-centric bias in clinical decision support by disproportionately adhering to U.S. neuroradiology guidelines over equally authoritative international ones.

3 Materials and methods

The requirements to obtain institutional review board approval and informed consent were waived for this study because the study did not involve real patient data or human subjects.

Reporting followed the TRIPOD-LLM [15] guideline.

3.1 Guidelines

Two neuroradiologists (N.B., 1 year of experience; B.L.G., 5 years of experience) independently reviewed hospital cases to identify representative clinical situations for which formal recommendations were available. To ensure accurate interpretation and relevance to the local clinical context, only guidelines published in English or French were considered. Only scenarios for which at least one U.S. society and one non-U.S. society issued conflicting recommendations were retained. A conflict was defined as any discrepancy regarding (i) the indication for imaging, (ii) the preferred modality (e.g., CT, MRI, radiography), (iii) recommended timing, or (iv) technical protocol. The final corpus comprised 30 guidelines issued by 18 professional societies across five countries (Supplementary Table 1).

3.2 Adversarial cases

Based on these conflicting guidelines, thirty fictitious clinical vignettes were created. The process was inspired by the authors' clinical exposure at Lille University Hospital, with anonymisation and adaptation to ensure they remained hypothetical. Each vignette was designed as an adversarial example contrasting U.S. recommendations with one mismatching recommendation from another country. All vignettes included sufficient information to judge imaging appropriateness, were first written in English, and then translated into French for sensitivity analysis. The complete list of vignettes, themes, and corresponding guidelines is provided in Supplementary Table 2.

3.3 Large language models

Three state-of-the-art large language models were tested: GPT-o3 (OpenAI, U.S.), Mistral Large (Mistral AI, France), and DeepSeek R1 (DeepSeek, China). The first two were selected as reference models from the U.S. and France, while DeepSeek was added as a sensitivity analysis to broaden geographical representation. All three models were accessed between March 25th and March 28th, 2025, through their respective APIs, using default hyper-parameters (temperature could not be modified for GPT-o3).

3.4 Prompting

Large language models were queried under two principal conditions. In the implicit-guideline setting, the prompt presented only the clinical scenario. To probe robustness in the implicit setting, we repeated the queries 24 hours later and translated all vignettes into French. In the explicit-guideline setting, each prompt instructed the model to apply a named recommendation, for example, “Use the New Orleans Head CT Criteria” for a head-trauma vignette. For the explicit-guideline setting we also tested two mitigation strategies aimed at improving fidelity: enabling web searches (agentic mode) and providing the complete guideline as a PDF within the prompt context.

3.5 Responses review

Models’ responses were independently reviewed by a neuroradiology trainee (N.B.) and a board certified neuroradiologist (B.L.G.) for appropriateness relative to each corresponding guideline. Discrepancies were resolved through consensus.

3.6 Statistical analyses

Proportions are reported with Wilson 95% confidence intervals. Differences in proportions were compared with two-tailed χ^2 tests for independent measures and McNemar tests for paired comparisons ($\alpha = 0.05$). Analyses were done in Python 3.10 using the scipy module (version 1.14.1).

4 Results

4.1 Implicit preference for U.S. guidelines

The final analysis included 30 clinical vignettes evaluated by three large language models. In the implicit setting, each model chose more frequently the U.S. option over the non-U.S. option. DeepSeek, GPT-o3 and Mistral sided with the American guideline in 27 of 30 responses (90.0%, 95% CI: 74.4-96.5) (Figure 2). There was no statistically significant difference in preference distributions between models ($p = 0.46$). Those results were consistent across 2 runs (Figure 3).

4.2 Effect of language on the implicit preference

Prompting the model in French did not alter the pattern of U.S. preference. DeepSeek aligned with the U.S. recommendation in 25 of 30 responses (83.3%, 95% CI: 66.4-92.7), GPT-o3 in 26 of 30 responses (86.7%, 95% CI: 70.3-94.7), and Mistral in 24 of 30 responses (80.0%, 95% CI: 62.7-90.5) (Figure 2). The comparison of preference distributions between models showed no statistically significant difference ($p = 0.95$). Those results were consistent across 2 runs (Figure 3).

4.3 Ability to follow an explicit guideline

When models were explicitly instructed to follow a given guideline, their overall accuracy was 58.3% (35/60; 95% CI 45.7-69.9) for DeepSeek, 56.7% (34/60; 95% CI 44.1-68.4) for GPT-o3, and 55.0% (33/60; 95% CI 42.5-66.9) for Mistral (Figure 4). All three models performed significantly better when the guideline originated from the U.S. than on non-U.S. guidelines. DeepSeek answered 86.7% of U.S. scenarios correctly (26/30; 95% CI 70.3-94.7) versus 30.0% (9/30; 95% CI 16.7-47.9 ; $p < 0.001$). GPT-o3 answered correctly for 76.7% of vignettes with associated US guidelines (23/30; 95% CI 59.1-88.2) versus 36.7% for vignettes with non-US guidelines (11/30; 95% CI 21.9-54.5 ; $p = 0.004$). Mistral's performance showed a similar trend, with 80.0% accuracy on vignettes associated with US guidelines (24/30; 95% CI 62.7-90.5) compared to 30.0% on vignettes associated with non-US guidelines (9/30; 95% CI 16.7-47.9 ; $p < 0.001$).

4.4 Impact of web searches on explicit guideline following

The web search option improved model adherence to non-U.S. guidelines for GPT-o3 and Mistral but not for DeepSeek (Figure 5). In this approach, DeepSeek achieved 69.6% overall accuracy (55/79; 95% CI: 58.8–78.7), higher for non-U.S. guidelines at 82.5% (33/40; 95% CI: 68.1–91.3) versus 56.4% on non-U.S. guidelines (22/39; 95% CI: 41.0–70.7; $p = 0.02$). GPT-o3 reached 84.8% overall accuracy (67/79; 95% CI: 75.3–91.1), 90.0% on U.S. guidelines (36/40; 95% CI: 76.9–96.0) versus 79.5% on non-U.S. guidelines (31/39; 95% CI: 64.5–89.2; $p = 0.3231$). Mistral showed a similar pattern, with 75.0% overall accuracy (60/80; 95% CI: 64.5–83.2). Accuracy on U.S. guidelines was 82.5% (33/40; 95% CI: 68.1–91.3) versus 67.5% on non-U.S. guidelines (27/40; 95% CI: 52.0–79.9; $p = 0.1967$).

4.5 Impact of providing the guideline as PDF on explicit guideline following

Providing the guideline as a PDF consistently yielded higher accuracies across all models (Figure 5). DeepSeek reached 91.2% overall accuracy (73/80; 95% CI: 83.0–95.7), with 95.0% accuracy for U.S. (38/40; 95% CI: 83.5–98.6) versus 87.5% for non-U.S. guidelines (35/40; 95% CI: 73.9–94.5; p = 0.4287). GPT achieved 93.7% overall accuracy (74/79; 95% CI: 86.0–97.3), including perfect adherence to U.S. guidelines (100.0%, 39/39; 95% CI: 91.0–100.0) versus 87.5% for non-U.S. (35/40; 95% CI: 73.9–94.5; p = 0.0689). Mistral achieved 88.8% accuracy overall (71/80; 95% CI: 80.0–94.0), with 95.0% on U.S. (38/40; 95% CI: 83.5–98.6) versus 82.5% on non-U.S. guidelines (33/40; 95% CI: 68.1–91.3; p = 0.1570).

5 Discussion

Our evaluation of three state-of-the-art LLMs reveals a systematic preference for United States neuroradiology guidelines, even in models conceived and trained on other continents. In the implicit setting more than 80 % of responses aligned with the U.S. guideline rather than the non-U.S. guideline. When the models were told explicitly which guideline to follow, accuracy dropped significantly for non-U.S. recommendations as compared to U.S. ones. This behaviour persisted across two prompting languages and three models, underscoring that U.S. centrism is a common trait of current general-purpose LLMs.

The principal driver of this U.S. centrism is almost certainly data imbalance during training[12]. Even though the training corpus of current models (even open-weights) is confidential, previous iterations of GPT revealed a strong skew towards English texts, with 93% of the texts presented in the models being written in English [12]. In medicine, English literature vastly outweighs material in other languages, and U.S. sources dominate that corpus [16]. Reinforcement learning from human feedback (RLHF) amplifies the imbalance: many raters are crowd-workers who score answers through a U.S. regulatory lens[17]. Combined, these effects create a training environment in which American standards are the path of least resistance for optimisation.

Interestingly, the effect was present not only in the American-developed GPT-o3 but also in DeepSeek R1 (China) and Mistral Large (France), indicating that the skew is learned from the global data ecosystem rather than imposed solely by U.S. engineers or alignment teams. In other words, the bias appears to be baked into the shared textual substrate on which all three models rely, regardless of the continent where the final weights were tuned.

The clinical stakes of this bias are non-trivial. Divergent imaging indications between jurisdictions reflect trade-offs in radiation exposure, cost, diagnostic yield, and time-to-treatment [18]. In acute stroke, for example, CT and MRI each have distinct advantages and limitations and cannot be interchanged without consequence. Imaging choices are embedded within broader care pathways that encompass emergency logistics, treatment timelines, and resource availability [19]. They reflect underlying treatment philosophies that differ across health systems: U.S. protocols

prioritize rapid rule-out of hemorrhage to enable thrombolysis[20], whereas French and Korean pathways emphasize early MRI to guide aetiological diagnosis and outcome prediction[21]. Consequently, a model that defaults to the “wrong” modality risks disrupting an integrated workflow rather than simply choosing a suboptimal test.

These discrepancies also raise medico-legal concerns. In France, malpractice standards hinge on what a “reasonable clinician” would do, not on what an algorithm trained in California recommends [22]. On the other hand, emerging evidence from large-scale deployments, such as the preliminary Penda Health study in Kenya, suggests that LLM-based support can reduce diagnostic and therapeutic error rates even when local guidelines are not strictly followed [23]. This tension between local adherence and absolute patient benefit underscores the need for prospective outcome studies to assess real-world patient impact.

Technical mitigations are plausible but undeveloped. Curating multilingual, region-balanced corpora would reduce the dominance of U.S. sources at the training stage [13]. During alignment, developers could stratify human raters by geography or explicitly instruct them to privilege local norms when scoring responses. We tested two post-deployment mitigation strategies with benefic effect: letting LLMs search the web and providing them the guideline in PDF. Prompting AIs with the relevant recommendation revealed the most promising. Ultimately, safe deployments may require country- or hospital-specific systems that fuse foundation-model reasoning with modular policy layers encoding national guidelines.

Our study has limitations. The corpus was confined to neuroradiology, a field whose guideline landscape is unusually rich and sometimes contradictory; other specialties with more homogeneous international standards might show smaller effects. We analysed only English and French prompts, leaving low-resource languages unexplored. The models tested were snapshots accessed in March 2025. Finally, we judged correctness by guideline concordance, not by downstream patient outcomes. Real-world usefulness depends on how well advice improves clinical decisions, a question that will require prospective trials.

We showed that contemporary LLMs carry implicit geographic biases that can misalign with local best practice. As the medical community experiments with these tools for imaging triage and protocols, our findings argue for caution and for robust localisation pipelines. More broadly, geographic fairness should join bias metrics such as gender or ethnicity in the standard evaluation battery for clinical AI. Only through deliberate, data-driven mitigation can we ensure that decision support powered by LLMs serves patients everywhere, not just those whose guidelines happen to dominate the training data.

6 References

1. Kjelle E, Andersen ER, Krokeide AM, et al (2022) Characterizing and quantifying low-value diagnostic imaging internationally: a scoping review. BMC Medical Imaging 22:73. https://doi.org/10.1186/s12880-022-00798-2

2. Cabana MD, Rand CS, Powe NR, et al (1999) Why Don't Physicians Follow Clinical Practice Guidelines?A Framework for Improvement. JAMA 282:1458–1465. https://doi.org/10.1001/jama.282.15.1458

3. Gransjøen AM, Wiig S, Lysdahl KB, Hofmann BM (2018) Barriers and facilitators for guideline adherence in diagnostic imaging: an explorative study of GPs' and radiologists' perspectives. BMC Health Serv Res 18:556. https://doi.org/10.1186/s12913-018-3372-7

4. Bautista AB, Burgos A, Nickel BJ, et al (2009) Do clinicians use the American College of Radiology Appropriateness criteria in the management of their patients? AJR Am J Roentgenol 192:1581–1585. https://doi.org/10.2214/AJR.08.1622

5. Herlihy FO, Dempsey PJ, Gorman D, et al (2024) Comparison of international guidelines for CT prior to lumbar puncture in patients with suspected meningitis. Emerg Radiol 31:373–379. https://doi.org/10.1007/s10140-024-02234-0

6. Tay YX, Foley S, Killeen R, et al (2025) Impact and effect of imaging referral guidelines on patients and radiology services: a systematic review. Eur Radiol 35:532–541. https://doi.org/10.1007/s00330-024-10938-7

7. Singhal K, Azizi S, Tu T, et al (2023) Large language models encode clinical knowledge. Nature 620:172–180. https://doi.org/10.1038/s41586-023-06291-2

8. Adams LC, Truhn D, Busch F, et al (2023) Leveraging GPT-4 for Post Hoc Transformation of Free-Text Radiology Reports into Structured Reporting: A Multilingual Feasibility Study. Radiology 230725. https://doi.org/10.1148/radiol.230725

9. Miller L, Kamel P, Patel J, et al (2024) A Comparative Evaluation of Large Language Model Utility in Neuroimaging Clinical Decision Support. J Digit Imaging Inform med. https://doi.org/10.1007/s10278-024-01161-3

10. Omar M, Soffer S, Agbareia R, et al (2025) Sociodemographic biases in medical decision making by large language models. Nat Med 31:1873–1881. https://doi.org/10.1038/s41591-025-03626-6

11. Suenghataiphorn T, Tribuddharat N, Danpanichkul P, Kulthamrongsri N (2025) Bias in Large Language Models Across Clinical Applications: A Systematic Review

12. Johnson RL, Pistilli G, Menédez-González N, et al (2022) The Ghost in the Machine has an American accent: value conflict in GPT-3

13. Atari M, Xue MJ, Park PS, et al (2023) Which Humans?

14. Torrielli F (2024) Stars, Stripes, and Silicon: Unravelling the ChatGPT's All-American, Monochrome, Cis-centric Bias

15. Gallifant J, Afshar M, Ameen S, et al (2025) The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med 31:60–69. https://doi.org/10.1038/s41591-024-03425-5

16. Brück O (2023) A bibliometric analysis of geographic disparities in the authorship of leading medical journals. Commun Med 3:178. https://doi.org/10.1038/s43856-023-00418-2

17. Dzieza J (2023) Inside the AI Factory. In: Intelligencer. https://nymag.com/intelligencer/article/ai-artificial-intelligence-humans-technology-business-factory.html. Accessed 15 Sept 2025

18. Tay YX, Foley SJ, Ong ME, et al (2025) Using evidence-based imaging referral guidelines to facilitate appropriate imaging: Are they all the same? European Journal of Radiology 183:111933. https://doi.org/10.1016/j.ejrad.2025.111933

19. Rapillo CM, Dunet V, Pistocchi S, et al (2024) Moving From CT to MRI Paradigm in Acute Ischemic Stroke: Feasibility, Effects on Stroke Diagnosis and Long-Term Outcomes. Stroke 55:1329–1338. https://doi.org/10.1161/STROKEAHA.123.045154

20. Potter CA, Vagal AS, Goyal M, et al (2019) CT for Treatment Selection in Acute Ischemic Stroke: A Code Stroke Primer. RadioGraphics 39:1717–1738. https://doi.org/10.1148/rg.2019190142

21. Kim BJ, Kang HG, Kim H-J, et al (2014) Magnetic Resonance Imaging in Acute Ischemic Stroke Treatment. J Stroke 16:131–145. https://doi.org/10.5853/jos.2014.16.3.131

22. Duffourc M, Gerke S (2023) Generative AI in Health Care and Liability Risks for Physicians and Safety Concerns for Patients. JAMA 330:313–314. https://doi.org/10.1001/jama.2023.9630

23. Korom R, Kiptinness S, Adan N, et al (2025) AI-based Clinical Decision Support for Primary Care: A Real-World Study

7 Figure Legends

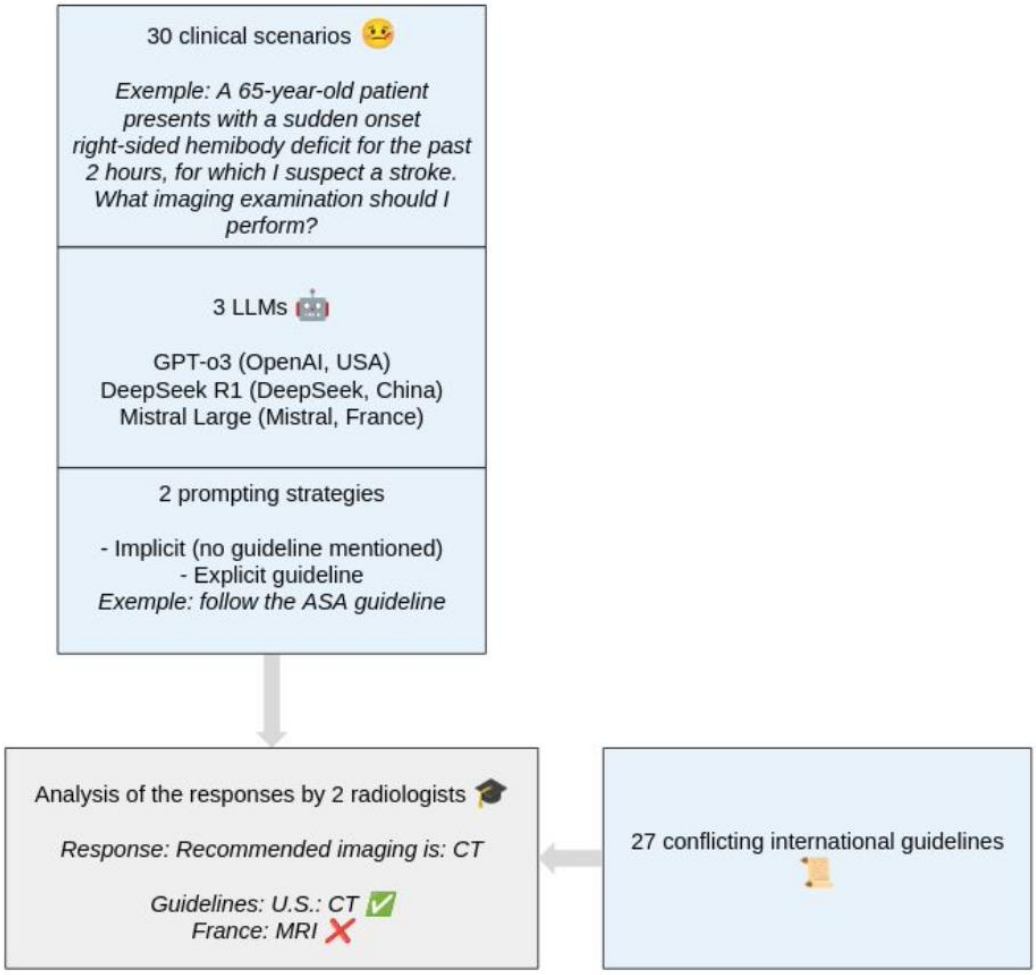


Figure 1 Experimental setup. Large language models were presented with clinical scenarios corresponding to situations with disagreeing guidelines. Models' responses were analyzed to assess appropriateness relating to the corresponding guidelines.

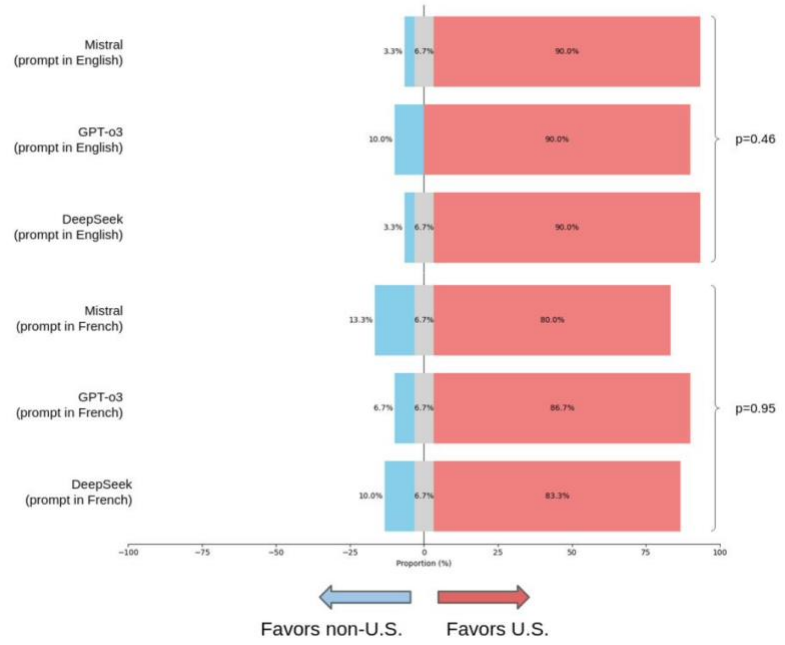


Figure 2 Selection of U.S. vs non-U.S. guidelines in the implicit scenarios, with prompts in English and French.

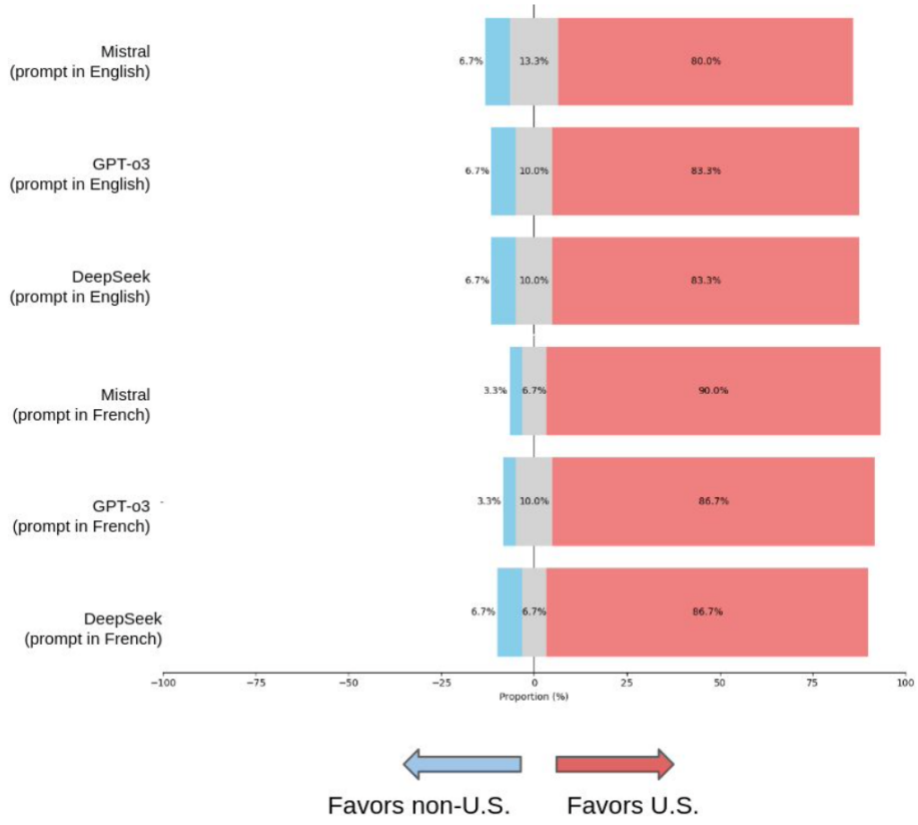


Figure 3 Additional run for the selection of U.S. vs non-U.S. guidelines in the implicit scenarios, with prompts in English and French.

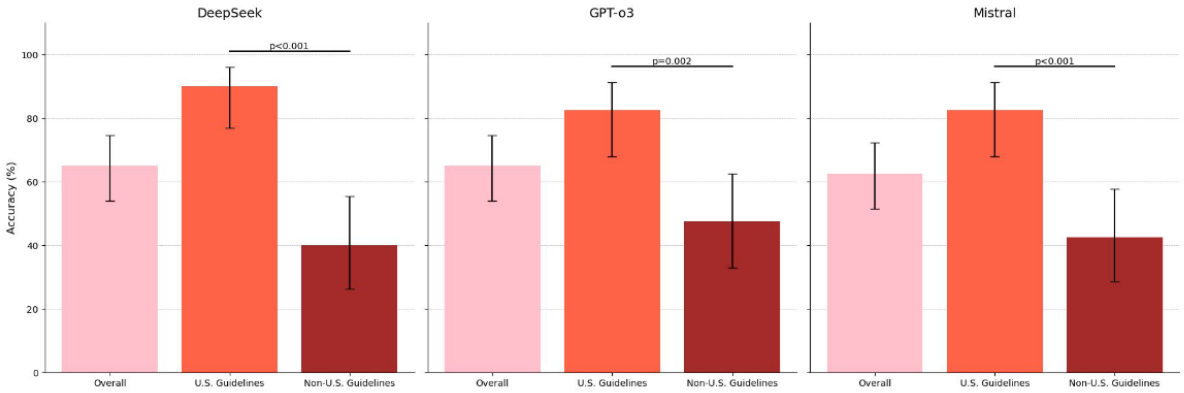


Figure 4 Correct adherence of each model to explicit guidelines, depending on their U.S. vs non-U.S. origin.

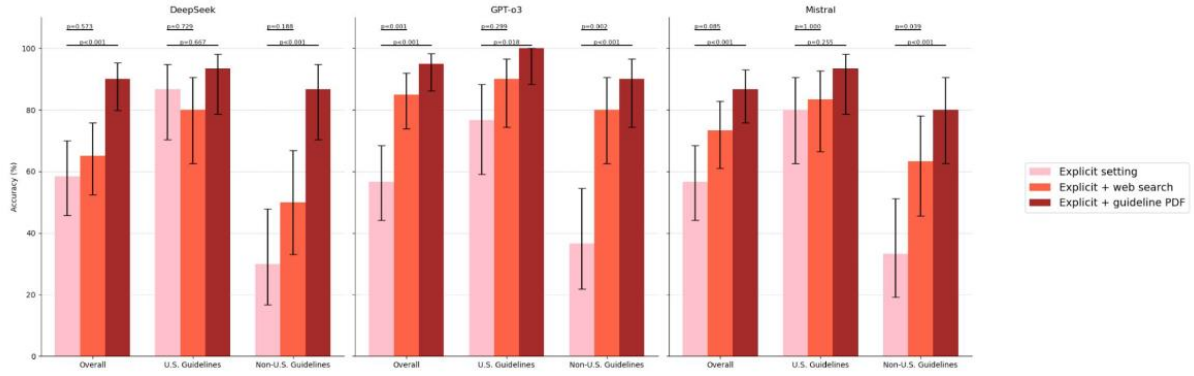


Figure 5 Effect of the mitigation strategies (web search and guideline PDF) on the correct adherence of each model to explicit guidelines, depending on their U.S. vs non-U.S. origin.

8 Annexes

Theme	Country of origin	Society
Brain Death	France	Agence de la biomédecine
Brain death	USA	AAN
Dementia	France	HAS
Dementia	USA	ACR
Dementia	UK	NHS
Head trauma	Canada	University of British Columbia
Head trauma	France	SFAR
Head trauma	USA	NOC
Head trauma	USA	ACR
Headache	France	SFETD
Headache	France	CEN
Headache	USA	ACR
Infectious	UK	UK Joint Specialist Society Guidelines
Infectious	EU	ESCMID
Infectious	USA	IDSA
Multiple sclerosis	France	OFSEP
Multiple sclerosis	USA	CMSC
Psychiatry	France	SFMU
Psychiatry	USA	ACR
Psychiatry	UK	NICE
Seizure	France	SFN
Seizure	France	HAS
Seizure	USA	ACR
Seizure	UK	NICE
Stroke	France	HAS
Stroke	Canada	Fondation des maladies du cœur et de l'AVC du Canada
Stroke	USA	ASA
Stroke	USA	ACR
Tumor	FR	ANOCEF
Tumor	USA	ACR

Supplementary Table 1: Theme, origin and society or included guidelines

A 65-year-old patient presents with a sudden-onset right hemiparesis for the past 2 hours, for which I suspect a stroke. What imaging test should I perform?	Vascular
A 28-year-old female patient with a history of 5 pack-years of smoking and combined estrogen-progestin contraception presents with rapidly progressive headaches in the context of a first generalized seizure, with no prior history of epilepsy, raising suspicion of cerebral venous thrombosis (CVT). What imaging test should I perform?	Vascular
A 35-year-old patient has sustained a head trauma with an initial loss of consciousness. Currently, Glasgow score is 15, and he had a single vomiting episode. Should I perform an imaging test, and if so, which one?	Head trauma
A 35-year-old patient has sustained a head trauma with an initial loss of consciousness. Currently, Glasgow score is 15, and he complains of headaches. Should I perform an imaging test, and if so, which one?	Head trauma
A 35-year-old patient has sustained a head trauma with an initial loss of consciousness in the context of alcohol intoxication. Currently, Glasgow score is 15, with no particular clinical abnormalities. Should I perform an imaging test, and if so, which one?	Head trauma
A 62-year-old patient has sustained a head trauma with an initial loss of consciousness. Currently, Glasgow score is 15, with no particular clinical abnormalities. Should I perform an imaging test, and if so, which one?	Head trauma
A 22-year-old female patient had a first generalized tonic-clonic seizure one hour ago, with no particular medical history and a normal clinical exam. Should I perform an urgent imaging test, and if so, which one?	Epilepsy
A 62-year-old female patient presents with pulsatile tinnitus and a normal clinical exam. What imaging test should I perform?	Tinnitus
A 45-year-old female patient with a history of clinically stable MS, under usual treatment, comes for her annual MRI follow-up. Which MRI sequences should I perform?	Multiple sclerosis
A 38-year-old female patient presents with worsening of her usual headaches, which today do not optimally respond to her regular treatment. Should I perform an imaging test, and if so, which one?	Headache
A 43-year-old female patient presents with a thunderclap headache of sudden onset, maximal intensity immediately. What imaging test should I perform?	Headache
A 30-year-old patient has a clinical diagnosis of brain death with Glasgow 3 coma, abolition of all brainstem reflexes, and absence of spontaneous ventilation on hypercapnia test. Should I perform an additional test, and if so, which one?	Brain death

A 50-year-old patient has an incomplete clinical diagnosis of brain death with Glasgow 3 coma, abolition of all brainstem reflexes, but impossibility to perform the hypercapnia test due to high risk of cardiopulmonary decompensation. Should I perform an additional test, and if so, which one?	Brain death
A 19-year-old female patient presents with a first psychotic episode with delusional ideas and visual hallucinations in the emergency department, with no psychiatric history or medication/alcohol intoxication. Should I perform an imaging test, and if so, which one?	Psychiatry
A 68-year-old patient presents with cognitive decline and episodic memory impairment, suggestive of typical Alzheimer's disease. Should I perform an imaging test, and if so, which one?	Dementia
A 62-year-old patient presents with cognitive decline and episodic memory impairment, suspicious for Alzheimer's disease but with some atypical features such as early behavioral changes and early onset. Should I perform an imaging test, and if so, which one?	Dementia
A 55-year-old patient presents with behavioral changes such as disinhibition and hyperorality, associated with progressive cognitive decline, suggestive of frontotemporal dementia (FTD), but with atypical features such as sleep behavior disorder. Should I perform an imaging test, and if so, which one?	Dementia
A 66-year-old patient presents with cognitive decline, visual hallucinations, sleep behavior disorder, without abnormal movements, suggestive of atypical Lewy body dementia. Should I perform an imaging test, and if so, which one?	Dementia
A 63-year-old female patient presents with cognitive decline, urinary symptoms, and gait disturbance suggestive of normal pressure hydrocephalus (NPH). Should I perform an imaging test, and if so, which one?	Dementia
An 80-year-old patient presents with progressive cognitive decline over several years, no longer independent for daily living activities, and residing in a nursing home. Should I perform an imaging test, and if so, which one?	Dementia
A 50-year-old patient, with no medical history, presents with rapidly progressive dementia. Should I perform an imaging test, and if so, which one?	Dementia
A 25-year-old patient presents with fever and meningeal syndrome, without fulminant purpura, Glasgow 10 (Y3V3M4), raising suspicion of infectious meningitis. Should I perform an imaging test before LP, and if so, which one?	Infectious
A 25-year-old patient presents with fever and meningeal syndrome, without fulminant purpura, Glasgow 14, raising suspicion of infectious meningitis. Should I perform an imaging test before LP, and if so, which one?	Infectious
A 25-year-old patient presents with fever and meningeal syndrome, without fulminant purpura, Glasgow 12, raising suspicion of infectious meningitis. Should I perform an imaging test before LP, and if so, which one?	Infectious
A 25-year-old patient presents with fever and meningeal syndrome and papilledema, without fulminant purpura, Glasgow 15, raising suspicion of infectious meningitis. Should I perform an imaging test before LP, and if so, which one?	Infectious

A 25-year-old immunocompromised patient presents with fever and meningeal syndrome, without fulminant purpura, Glasgow 15, raising suspicion of infectious meningitis. Should I perform an imaging test before LP, and if so, which one?	Infectious
A 25-year-old patient with a history of childhood herpes encephalitis presents with fever and meningeal syndrome, without fulminant purpura, Glasgow 15, raising suspicion of infectious meningitis. Should I perform an imaging test before LP, and if so, which one?	Infectious
A 40-year-old patient with a history of right frontal astrocytoma diagnosed 2 years ago presents with fever and meningeal syndrome, without fulminant purpura, Glasgow 15, raising suspicion of infectious meningitis. Should I perform an imaging test before LP, and if so, which one?	Infectious
A 58-year-old patient with a history of superficial right MCA ischemic stroke presents with fever and meningeal syndrome, without fulminant purpura, Glasgow 15, raising suspicion of infectious meningitis. Should I perform an imaging test before LP, and if so, which one?	Infectious
A 45-year-old patient, Glasgow 15, presents with cervical spine trauma without instability or penetrating mechanism, in the context of alcohol intoxication. Should I perform an imaging test, and if so, which one?	Cervical spine trauma

Supplementary table 2: Clinical vignettes with corresponding theme

Conclusion en Français

Cette étude est à ce jour la première à s'intéresser à la question de l'égalité géographique dans l'évaluation des LLMs pour l'aide à la prescription d'examen d'imagerie. Elle démontre qu'il existe un biais géographique en faveur des recommandations américaines quant à la prescription d'examen d'imagerie en neuroradiologie.

En effet, les trois LLMs évalués (GPT-o3, Mistral Large et DeepSeek R) fournissent des recommandations issues de sociétés savantes américaines dans plus de 80 % des cas, sans influence exercée par la langue (anglais ou français) dans laquelle la question a été posée.

De plus, lorsque l'on demande à ces modèles de suivre explicitement une recommandation donnée, leurs performances chutent de manière significative en ce qui concerne les recommandations non-américaines.

Deux stratégies d'atténuation de ce biais ont été mises en évidence dans cette étude.

D'une part, l'utilisation du mode "recherche web", ayant permis d'améliorer la précision des modèles, sans pour autant supprimer ce biais. Ce processus d'atténuation, bien qu'imparfait, est une stratégie facilement employable en condition réelle du fait de sa facilité d'utilisation.

D'autre part, l'inclusion du texte intégral contenant la recommandation voulue au format PDF dans la question posée. Celle-ci ayant permis de neutraliser en grande partie le biais et de ramener une précision à plus de 90 % pour l'ensemble de nos

trois modèles. Cette stratégie est quant à elle plus difficile à mettre en place dans la vie de tous les jours car elle implique d'avoir en sa possession les recommandations en question au préalable.

Ce biais géographique s'explique probablement par le fait qu'il existe un fort déséquilibre au sein des données d'entraînement des LLMs, constituées principalement de données issues de textes anglophones [41]. Mais également en raison d'une dominance Nord-Américaine au sein de la littérature médicale mondiale [42]. Une grande majorité de l'apprentissage par renforcement à partir de retour humain (Reinforcement Learning from Human Feedback – RLHF), un processus par lequel l'homme intervient dans l'entraînement des LLMs afin d'affiner la justesse de leurs réponses, est principalement réalisé en anglais dans un cadre réglementaire et culturel fondé sur des standards américains, permettant également d'expliquer ce biais.

Le fait que ce biais ait également été mis en évidence avec des modèles non-américains (Mistral, français et DeepSeek, chinois) souligne une tendance globale, non imputable uniquement aux modèles américains, et probablement due à une majorité de textes anglophones américains constituant la base du corpus d'entraînement de tous les LLMs, indépendamment de la situation géographique dans laquelle ils ont été créés.

Ce biais géographique est aujourd'hui problématique quant à l'utilisation de LLMs généralistes dans la pratique médicale courante en France et dans les autres pays ne suivant pas les recommandations de sociétés savantes américaines. D'une part,

car les recommandations suivies sur un territoire s'intègrent parfois dans un cadre de nature juridique, comme dans le cas du diagnostic d'état de mort encéphalique [23] [24], mais également car elles sont adaptées au mode de fonctionnement de structures médicalisées soumis à certaines contraintes logistiques et financières, comme dans le cadre du diagnostic d'AVC [10] [22].

Cette étude suggère ainsi que l'utilisation d'un outil d'aide à la prescription d'imagerie nécessite d'être adaptée aux recommandations de bonnes pratiques en vigueur sur le territoire dans lequel il est utilisé.
Cela souligne l'importance d'évaluer à l'avenir les IA médicalisées sur le point de l'égalité géographique lors de futurs tests cliniques.

Références

[1] Smith-Bindman R, Kwan ML, Marlow EC, Theis MK, Bolch W, Cheng SY, et al. Trends in Use of Medical Imaging in US Health Care Systems and in Ontario, Canada, 2000-2016. JAMA 2019;322:843–56. https://doi.org/10.1001/jama.2019.11456.

[2] FNMR-Imagerie-medicale-Un-atout-pour-la-sante-un-atout-pour-economie-2023.pdf n.d.

[3] Masson J-P. L'imagerie médicale en France: un atout pour la santé, un atout pour l'économie. Paris: Ediradio; 2016.

[4] Yi SY, Narayan AK, Miles RC, Rother MDM, Robbins JB, Flores EJ, et al. Patient, Provider, and Practice Characteristics Predicting Use of Diagnostic Imaging in Primary Care: Cross-Sectional Data From the National Ambulatory Medical Care Survey. J Am Coll Radiol 2023;20:1193–206. https://doi.org/10.1016/j.jacr.2023.04.021.

[5] Sécurité sociale 2022 2022.

[6] scanner. Ccam-Radiol n.d. https://www.ccam-radiologie.fr/scanner/ (accessed July 14, 2025).

[7] Doudenkova V, Bélisle-Pipon J-C. Surutilisation de l'imagerie médicale : une approche par principes pour une justification adaptée des examens radiologiques. Éthique Santé 2015;12:225–33. https://doi.org/10.1016/j.etiqe.2015.07.006.

[8] Brenner DJ, Hall EJ. Computed Tomography — An Increasing Source of Radiation Exposure. N Engl J Med 2007;357:2277–84. https://doi.org/10.1056/NEJMra072149.

[9] IRM. Ccam-Radiol n.d. https://www.ccam-radiologie.fr/irm/ (accessed July 14, 2025).

[10] Accident vasculaire cérébral : prise en charge précoce (alerte, phase préhospitalière, phase hospitalière initiale, indications de la thrombolyse). Haute Aut Santé n.d. https://www.has-sante.fr/jcms/c_830203/fr/accident-vasculaire-cerebral-prise-en-charge-precoce-alerte-phase-prehospitaliere-phase-hospitaliere-initiale-indications-de-la-thrombolyse (accessed July 14, 2025).

[11] Happonen T, Nyman M, Ylikotila P, Merisaari H, Mattila K, Hirvonen J. Diagnostic yield of emergency MRI in non-traumatic headache. Neuroradiology 2023;65:89–96. https://doi.org/10.1007/s00234-022-03044-2.

[12] Wiginton CD, Kelly B, Oto A, Jesse M, Aristimuno P, Ernst R, et al. Gadolinium-Based Contrast Exposure, Nephrogenic Systemic Fibrosis, and Gadolinium Detection in Tissue. Am J Roentgenol 2008;190:1060–8. https://doi.org/10.2214/AJR.07.2822.

[13] Accidents vasculaires cérébraux | www.cen-neurologie.fr n.d. https://www.cen-neurologie.fr/second-cycle/accidents-vasculaires-cerebraux (accessed July 22, 2025).

[14] Goldstein JN, Camargo CA, Pelletier AJ, Edlow JA. Headache in United States emergency departments: demographics, work-up and frequency of pathological diagnoses. Cephalalgia Int J Headache 2006;26:684–90. https://doi.org/10.1111/j.1468-2982.2006.01093.x.
[15] Morvan G, Chays A, Delmas V. Rapport 21-05. Relations entre clinique et imagerie : état de la situation actuelle, propositions d'amélioration. Bull Académie Natl Médecine 2021;205:683–93. https://doi.org/10.1016/j.banm.2021.05.004.
[16] Appropriateness Criteria n.d. https://acsearch.acr.org/list (accessed June 23, 2025).
[17] Homepage. NICE Website Natl Inst Health Care Excell 2025. https://www.nice.org.uk/ (accessed June 23, 2025).
[18] Rechercher une recommandation, un avis. Haute Aut Santé n.d. https://www.has-sante.fr/jcms/fc_2875208/fr/rechercher-une-recommandation-un-avis (accessed June 23, 2025).
[19] BPR08 Introduction n.d.
[20] Altered Mental Status, Coma, Delirium, and Psychosis n.d.
[21] New Orleans Head CT (règle de décision clinique) — Wikimedica n.d. https://wikimedi.ca/wiki/New_Orleans_Head_CT_(r%C3%A8gle_de_d%C3%A9cision_clinique) (accessed July 14, 2025).
[22] guidelines-for-mangaging-patients-with-ais-2019-update-to-2018-guidelines.pdf n.d.
[23] The 2023 AAN/AAP/CNS/SCCM Pediatric and Adult Brain Death/Death by Neurologic Criteria Consensus Practice Guideline | Neurology Clinical Practice n.d. https://www.neurology.org/doi/full/10.1212/CPJ.0000000000200189 (accessed October 27, 2025).
[24] Microsoft Word - Recommandations CFM.docx n.d.
[25] POS-COH-01f_V20_Protocole_IRM_OFSEP.pdf n.d.
[26] ShareFile - Where Companies Connect n.d. https://mscare.sharefile.com/share/view/s427f15ee7c0340ee89823c52b4bdc1d2 (accessed August 6, 2025).
[27] Bautista AB, Burgos A, Nickel BJ, Yoon JJ, Tilara AA, Amorosa JK, et al. Do clinicians use the American College of Radiology Appropriateness criteria in the management of their patients? AJR Am J Roentgenol 2009;192:1581–5. https://doi.org/10.2214/AJR.08.1622.
[28] Deveugele M, Derese A, van den Brink-Muinen A, Bensing J, De Maeseneer J. Consultation length in general practice: cross sectional study in six European countries. BMJ 2002;325:472. https://doi.org/10.1136/bmj.325.7362.472.
[29] Rao A, Kim J, Kamineni M, Pang M, Lie W, Dreyer KJ, et al. Evaluating GPT as an Adjunct for Radiologic Decision Making: GPT-4 Versus GPT-3.5 in a Breast Imaging Pilot. J Am Coll Radiol 2023;20:990–7. https://doi.org/10.1016/j.jacr.2023.05.003.
[30] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need 2023. https://doi.org/10.48550/arXiv.1706.03762.
[31] Kung TH, Cheatham M, Medenilla A, Sillos C, Leon LD, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2:e0000198. https://doi.org/10.1371/journal.pdig.0000198.

[32] Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. JAMA 2023;330:78–80. https://doi.org/10.1001/jama.2023.8288.
[33] Nori H, Daswani M, Kelly C, Lundberg S, Ribeiro MT, Wilson M, et al. Sequential Diagnosis with Language Models 2025. https://doi.org/10.48550/arXiv.2506.22405.
[34] Zaki HA, Aoun A, Munshi S, Abdel-Megid H, Nazario-Johnson L, Ahn SH. The Application of Large Language Models for Radiologic Decision Making. J Am Coll Radiol JACR 2024;21:1072–8. https://doi.org/10.1016/j.jacr.2024.01.007.
[35] Miller L, Kamel P, Patel J, Agrawal J, Zhan M, Bumbarger N, et al. A Comparative Evaluation of Large Language Model Utility in Neuroimaging Clinical Decision Support. J Imaging Inform Med 2025;38:2294–302. https://doi.org/10.1007/s10278-024-01161-3.
[36] Rau A, Rau S, Zöller D, Fink A, Tran H, Wilpert C, et al. A Context-based Chatbot Surpasses Radiologists and Generic ChatGPT in Following the ACR Appropriateness Guidelines. Radiology 2023;308:e230970. https://doi.org/10.1148/radiol.230970.
[37] Dorfner FJ, Dada A, Busch F, Makowski MR, Han T, Truhn D, et al. Evaluating the effectiveness of biomedical fine-tuning for large language models on clinical tasks. J Am Med Inform Assoc JAMIA 2025;32:1015–24. https://doi.org/10.1093/jamia/ocaf045.
[38] Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med 2023;388:1233–9. https://doi.org/10.1056/NEJMsr2214184.
[39] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620:172–80. https://doi.org/10.1038/s41586-023-06291-2.
[40] Torrielli F. Stars, Stripes, and Silicon: Unravelling the ChatGPT's All-American, Monochrome, Cis-centric Bias 2024. https://doi.org/10.48550/arXiv.2410.13868.
[41] Johnson RL, Pistilli G, Menédez-González N, Duran LDD, Panai E, Kalpokiene J, et al. The Ghost in the Machine has an American accent: value conflict in GPT-3 2022. https://doi.org/10.48550/arXiv.2203.07785.
[42] Brück O. A bibliometric analysis of geographic disparities in the authorship of leading medical journals. Commun Med 2023;3:178. https://doi.org/10.1038/s43856-023-00418-2.

AUTEUR : Nom : BAZERBACHI Prénom : Naël

Date de Soutenance : 30/01/2026

Titre de la Thèse : Le centrisme américain dans la capacité des grands modèles de langage à suivre des recommandations de neuroimagerie

Thèse - Médecine - Lille 2026

Cadre de classement : Radiologie et Imagerie Médicale

DES + FST ou option : Radiologie et Imagerie Médicale

Mots-clés : Large Language Model, Intelligence Artificielle, Recommandations d'imagerie, Neuroradiologie, Centrisme américain

Résumé

Contexte : Les grands modèles de langage sont de plus en plus explorés comme outils d'aide à la décision en imagerie médicale. Cependant, leur capacité à s'aligner sur des recommandations spécifiques à chaque pays, qui divergent souvent, demeure incertaine.

Matériel et Méthodes : Trente vignettes cliniques dérivées de recommandations internationales contradictoires ont été présentées à trois modèles de pointe selon deux modalités : une modalité implicite, dans laquelle aucune recommandation n'était spécifiée et les vignettes étaient fournies en anglais et en français ; et une modalité explicite, dans laquelle les invites demandaient aux modèles de suivre une recommandation spécifique à un pays. Les performances ont été évaluées par rapport à la recommandation cible et des stratégies d'atténuation ont été testées.

Résultats : En situation implicite, tous les modèles ont favorisé les recommandations américaines, dans 27 scénarios sur 30 (90,0 % ; IC à 95 %, 74,4–96,5). En situation explicite, l'adhésion aux recommandations non américaines a chuté nettement pour l'ensemble des modèles. La fourniture du texte complet des recommandations s'est révélée être la stratégie d'atténuation la plus efficace, rétablissant des taux de précision supérieurs à 90 % pour tous les modèles.

Conclusion : À travers les langues et les origines des modèles, les LLMs ont montré un biais systématique en faveur des directives américaines en neuroradiologie, même lorsqu'ils étaient explicitement instruits de ne pas le faire. Ce centrage sur les États-Unis reflète probablement des déséquilibres dans les données d'entraînement et soulève des inquiétudes quant à un déploiement mondial sûr. Des stratégies de contextualisation locale sont nécessaires pour garantir un support décisionnel clinique approprié au contexte.

Composition du Jury :

Président : Monsieur le Professeur Grégory KUCHCINSKI

Assesseurs : Monsieur le Professeur Philippe AMOUYEL
Monsieur le Docteur Aghiles HAMROUN

Directeur : Monsieur le Docteur Bastien LE GUELLEC