



IA générative en médecine générale : attentes des professionnels et évaluation de modèles de langue sur le raisonnement diagnostique

Sofiane Djillali Sid-Ahmed, Josselin Corvellec

► To cite this version:

Sofiane Djillali Sid-Ahmed, Josselin Corvellec. IA générative en médecine générale : attentes des professionnels et évaluation de modèles de langue sur le raisonnement diagnostique. Médecine humaine et pathologie. 2024. <dumas-04918701>

HAL Id: dumas-04918701

<https://dumas.ccsd.cnrs.fr/dumas-04918701v1>

Submitted on 29 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

IMPORTANT : OBLIGATIONS DE LA PERSONNE CONSULTANT CE DOCUMENT

Conformément au *Code de la propriété intellectuelle*, nous rappelons que le document est destiné à un **usage strictement personnel**. Les "analyses et les courtes citations justifiées par le caractère critique, polémique, pédagogique, scientifique ou d'information" sont autorisées sous réserve de mentionner les noms de l'auteur et de la source (article L. 122-4 du *Code de la propriété intellectuelle*). Toute autre représentation ou reproduction intégrale ou partielle, faite sans le consentement de l'auteur ou de ses ayants droit, est illicite.

De ce fait, nous vous rappelons notamment que, **sauf accord explicite** de l'auteur de la thèse ou du mémoire, **vous n'êtes pas autorisé** à rediffuser ce document sous quelque forme que ce soit (impression papier, transfert par voie électronique, ou autre). Tout contrevenant s'expose aux peines prévues par la loi.

NANTES UNIVERSITE

FACULTE DE MEDECINE

Année 2024

N°

THESE

pour le

DIPLOME D'ETAT DE DOCTEUR EN MEDECINE

(DES de MEDECINE GENERALE)

par

Sofiane-Djillali SID-AHMED

et

Josselin CORVELLEC

Présentée et soutenue publiquement le 29 octobre 2024

**IA générative en médecine générale : attentes des professionnels et évaluation
de modèles de langue sur le raisonnement diagnostique**

Président : Monsieur le Professeur Jean-Pascal FOURNIER

Directeur de thèse : Monsieur le Professeur Pierre-Antoine GOURRAUD

Co-directeur de thèse : Monsieur le Docteur Adrien BAZOGE

Membre du jury : Monsieur le Docteur Grégoire HINZELIN

REMERCIEMENTS COMMUNS

Merci à monsieur le Professeur Gourraud Pierre-Antoine de nous avoir offert l'opportunité de travailler sur ce sujet passionnant et nous avoir fait profiter de son expérience.

Merci à monsieur le Docteur Bazoge Adrien pour sa disponibilité, son aide indispensable et l'intérêt porté à notre travail.

Merci à monsieur le Professeur Fournier Jean-Pascal de nous faire l'honneur de présider notre jury et l'intérêt porté à notre travail.

Merci à monsieur le Docteur Hinzelin Grégoire pour sa participation à notre jury de thèse et l'intérêt porté à notre travail.

Nous remercions l'URML des Pays de la Loire pour leur aimable participation à la diffusion de notre questionnaire de thèse.

REMERCIEMENTS DE JOSSELIN

Pour cette thèse comme pour tout le reste, je remercie avant tout Marine Le Louet, dont la présence, l'humour, l'intelligence, le sourire et l'amour auront facilité bien des choses... Dix ans ensemble que je n'ai pas vu passer, ce n'est que le début. Merci. Je t'aime.

A mon père. Tu es et resteras l'image que je me fais « du bon médecin ». Merci de m'avoir donné le goût d'apprendre et de comprendre les choses, de m'avoir appris que les soucis du quotidien ne sont pas si graves. Merci à maman et toi d'être toujours là pour moi. Merci pour tout ce que vous m'avez donné. Je vous aime tous les deux.

Merci à mes bien chères sœurs. Je suis heureux et fier de vous avoir. Dédicace spéciale à Steven et Amédée. Je vous aime tous.

Merci à ma belle-famille de m'avoir accueilli, je sais que je peux compter sur vous.

Mention spéciale à Christopher. Mon grand frère. Merci. Je vous embrasse, Flora, Apolline et toi.

A tous les vieux de la vieille, mes amis de « Vendée ». Merci pour les moments passés et ceux à venir. Merci Jérémy, James, Marion, Julie, Maxime, Nolwenn, Clément, Martin S., Martin P., Marie, Xavier, Tanguy, Gabrielle, Alex, Amaury.

Merci à ceux avec qui j'ai partagé ces études de médecine, dans les bons et les mauvais moments. Merci à Alexandre, Pierre, Lucille, Charlotte, Isaline, Félix, Juliette, Sindy, Anna et les autres. Merci « On s'la colle » pour tous ces moments, je n'en pense pas moins. Merci aux étudiants du master d'anat., cru 2014-2015, j'aurais aimé aller au bout. Merci aux « Wapiti » pour ces moments ensembles.

Merci Edouard, à qui je me dois de dédier un texte un peu plus long. Merci pour tous ces moments, merci de lutter contre ma réticence naturelle à à peu près tout, tu m'as montré qu'on gagne souvent à tenter le coup.

Un grand merci à ceux que j'ai rencontré pendant cette période dite de « l'internat ». La fine équipe de Challans, merci Clarisse, Pierre-Alain, Margot, Nico et Marc. Merci à Pauline et Alicia, la JAP a été le grand plus de mon séjour sablais ! Merci Valentine et Héléna.

Petite mention spéciale à tous ceux qui m'ont redonné foi dans la médecine, ce qui n'était pas une sinécure... Merci à Nicolas Baril et Charlotte Grimault pour leur écoute. Merci à Joshua et au Dr Marot. Merci à Frédérique Carton pour son humanité. Merci à Coline et Cécile. Merci à Soizic, Fanny et Jean-Yves pour vos enseignements et votre gentillesse. Merci Marion, David, Marion, Pauline, Stéphanie et Patrick.

Merci au Professeur Tirard Stéphane, pour m'avoir sorti de mon errance quant à la recherche d'un directeur de thèse.

Pour finir, merci Kiwi. Tu resteras mon premier chat, mon chat préféré. Grandir avec toi aura été un des grands bonheurs de ma vie. J'espère que tu es heureux là où tu es, au paradis des chats.

REMERCIEMENTS DE SOFIANE

Je dédie cette thèse avant tout à mes parents. A mon père, pour m'avoir inspiré à faire ce beau métier mais aussi pour le modèle de droiture que tu incarnes tous les jours et qui me tire toujours vers le haut. A ma mère, pour ton amour sans limites, ta joie constante, ton attention et les petits soins dont as toujours fait preuve avec moi. Je ne serais jamais arrivé jusqu'ici sans vous et je n'oublierai jamais les immenses sacrifices que vous avez fait tous les deux pour offrir un avenir meilleur à vos enfants. Je vous aime.

Je la dédie à mes tendres sœurs, Louiza, Lila et Meriem, pour nos fous rires, votre bonne humeur contaminante, et votre soutien indéfectible qui m'ont permis de survivre à des études interminables. A Mounir et Olivier, que je suis fier de compter dans ma famille. A Neïla, Aydin et Marwan, peut-être que vous lirez un jour ces pages et qu'elles vous inspireront. Je vous aime tous.

Je la dédie aussi à Amel et Nini, qui malgré la distance sont toujours là pour moi.

Je tiens à remercier mes amis. Stuart, pour ton soutien à travers les âges, quelle que soit la ville on parvient toujours à se retrouver. Merci à Clément et Estelle, on ne se voit pas souvent mais à chaque fois j'ai l'impression que c'était comme hier.

Merci à Nadjib et Yazane, pour toutes ces heures passées à refaire le monde et à ouvrir sans cesse de nouvelles perspectives en quête du mieux.

Merci à Ghiles et Sid Ali, pour tous nos étés passés ensemble et votre soutien.

Merci bien sûr aux Sablais : Thomas le tigre rouge, Mélissa, Gwendal, Grégoire, Alizée, Nico, Inès, Vivien, Marie Nicolas, Sarah, William, Marie Cottenceau, Arthur, Claire. Vous avez réussi à rendre l'internat facile, et malgré la distance aujourd'hui je serai là.

Je tiens aussi à remercier mes anciens co-externes, particulièrement Alexandre et Hugo. Je n'ai que de bons souvenirs de nos années à Brest et je les garderai toujours.

Un merci particulier au Professeur Alain Sarraux au CHU de Brest, pour m'avoir aiguillé et sans qui je n'aurais pas pu participer à cette thèse.

PREFACE DE JOSSELIN

« Nous nous sommes émerveillés de notre magnificence dès la venue au monde de l'I.A »

Morpheus - Matrix, 1999.

Il ne me semble pas que l'IA actuelle soit en voie d'asservir l'humanité à court terme. Mais je ne peux pas m'empêcher d'avoir cette phrase en tête lorsque j'entends certains discours sur l'IA, expliquant à quel point tout va changer grâce à elle. Il me semble que l'IA va effectivement modifier beaucoup de choses dans nos sociétés, y compris la médecine. Mais je reste mal à l'aise vis à vis des discours trop enthousiastes ou au contraire trop anxiogènes, qui me semblent tous deux relever du fantasme, plutôt que d'une appréhension correcte de la réalité. Quoi qu'il en soit, la crainte d'être un jour remplacé par un robot-médecin est devenue pour ma part un peu plus palpable au tournant des années 2020. Les héros de nos séries médicales américaines, ces médecins aux connaissances encyclopédiques, perspicaces à souhait, capables à eux-seuls de faire à peu près tout ce que la médecine peut faire, semblent désormais dépassés. Si le mythe du clinicien incollable, résolvant l'énigme la plus ardue relevait peut-être lui-aussi du fantasme, tendre vers une certaine finesse clinique, vivre le bonheur de savoir et comprendre, d'appliquer ses connaissances au réel, étaient des objectifs jusqu'à présents légitimes et réalisables. Mais tout cela paraît bien peu de choses face aux technologies de l'IA et à ses capacités infiniment plus grandes que celles de l'être humain. La question se pose. A-t-on encore besoin de médecins ? Certains répondent non, sans hésiter. Ma petite expérience de l'exercice médical me rend plus nuancé. Je crains que l'arrivée de l'IA dans nos métiers ne s'accompagne peut-être d'un certain désenchantement de la profession. Ma crainte est aussi que ces évolutions ne nous fassent « jeter le bébé avec l'eau du bain ». Les promoteurs de l'IA semblent parfois ne pas se poser la question : a-t-on vraiment besoin d'IA partout ? Les humains qui pratiquent la médecine font-ils les choses si mal qu'ils ne soient plus capables de faire sans l'IA ? Je ne suis pas persuadé qu'un monde où nous serions dépendant à 100% de l'IA pour nous soigner soit souhaitable. Mais si nous voulons garder des humains dans la boucle, sans doute faudra-t-il que ces derniers soient compétents en médecine. Mais comment l'être si on ne la pratique plus, après l'avoir déléguée à l'IA ?

Ce sont les raisons qui m'ont poussé à travailler sur ce sujet. Regarder en face ces progrès et mieux les comprendre, rester un acteur du soin, au service de la santé humaine.

PREFACE DE SOFIANE

« L'intelligence, c'est la capacité de s'adapter au changement. »

Stephen Hawking.

Depuis quelques années, nous assistons à une révolution majeure dans nos vies. Pourtant, l'intelligence artificielle est un domaine dont on parle depuis plus de 70 ans, mais ChatGPT semble avoir précipité la conscience collective au point de susciter les réactions les plus extrêmes. La mise à portée d'outils aussi puissants soient-ils a réveillé chez beaucoup l'angoisse de voir émerger des machines que nous ne pouvions imaginer qu'au cinéma jusqu'à présent. A l'inverse, elle a aussi provoqué de l'émerveillement chez certains, convaincus que nous pourrions bientôt résoudre des équations complexes comme la crise climatique ou le cancer.

En ce qui me concerne, bien que je sois médecin, j'ai toujours été fasciné par le monde de l'informatique. Mon côté « geek » est très présent et c'est probablement pour ça que je n'ai jamais réellement « craint » cet essor de l'intelligence artificielle. En effet, je pense qu'il reste important de tempérer nos réactions : peut-on réellement la comparer à la révolution industrielle par exemple, qui aurait fait tripler la population du Royaume-Uni en 100 ans ? J'ai du mal à y croire, mais là encore seul l'avenir nous le dira.

Nous disposons tous aujourd'hui de « vastes encyclopédies dotées de supercalculateurs » capables de répondre à de nombreuses interrogations en un temps record. Il est clair que la médecine est un domaine dans lequel l'IA va exceller et devenir prédominante, pour le meilleur et pour le pire. Mais en réalité, il n'appartient qu'à nous de définir le cadre de son utilisation, d'éclaircir nos propres interrogations, de choisir les données que nous lui mettrons à sa portée, et surtout d'imaginer ce que nous ferons des résultats qu'elle nous offrira. En médecine comme en informatique, il peut parfois être bien plus difficile de savoir quelle est la question à poser et comment le faire, plutôt que de chercher la réponse.

Mais aujourd'hui, je crois que le plus gros travail à faire est sur nous-même. Serons-nous capables d'accepter les transformations portées par ces nouvelles technologies sur notre propre pratique ? Serons-nous suffisamment humbles pour sortir de notre cocon, sans pour autant renoncer à notre compétence, notre humanité et notre rapport aux patients ? Il me semble que c'est dans cette capacité à embrasser le changement que chacun prendra la pleine mesure de sa propre intelligence.

TABLE DES MATIERES

INTRODUCTION	13
I. INTRODUCTION A L'INTELLIGENCE ARTIFICIELLE	13
A. Définitions	13
B. Essor de l'intelligence artificielle	16
II. INTELLIGENCE ARTIFICIELLE GENERATIVE	19
A. Remise en contexte	19
B. Promesses et dangers	23
III. INTELLIGENCE ARTIFICIELLE ET SANTE	24
A. Applications de l'IA en santé, quelques exemples	24
B. Médecine générale et intelligence artificielle	26
IV. RATIONNEL DU TRAVAIL DE THESE	35
A. Enquête auprès des médecins généralistes	36
B. Évaluation de grands modèles de langues sur le raisonnement diagnostique	36
MATERIEL ET METHODE	38
I. ENQUETE AUPRES DES MEDECINS GENERALISTES	38
A. Objectifs de l'étude	38
B. Type d'étude	38
C. Population visée	38
D. Élaboration du questionnaire	38
II. ÉVALUATION DE MODELES DE LANGUE SUR LE DIAGNOSTIC CLINIQUE	40
A. Conception d'une banque de cas clinique	40
B. Choix de la métrique de test	48
C. Choix des modèles testés et méthodologie des tests	55
RESULTATS	57
I. ENQUETE AUPRES DES MEDECINS GENERALISTES	57
A. Données sociodémographiques	57
B. Réponses au questionnaire	59
C. Question ouverte	68
D. Résumé des résultats	69
II. ÉVALUATION DES GRANDS MODELES DE LANGUE	71
A. Caractéristiques des réponses	71
B. Moyennes, écarts types, médianes, percentiles	73
C. Distribution des notes par modèle	73
D. Profil des notes par item et par modèle	77
E. Analyse comparative des performances	78
F. Performance des modèles selon la spécialité	84
G. Comparabilité inter-auteurs	91

H. <i>Éléments d'analyse qualitative</i>	91
I. Résumé des résultats	94
DISCUSSION	95
I. ENQUETE AUPRES DES MEDECINS GENERALISTES	95
A. <i>Discussion des résultats</i>	95
B. <i>Limites de l'étude</i>	97
II. ÉVALUATION DES MODELES DE LANGUE	98
A. <i>Discussion des résultats</i>	98
B. <i>Limites de l'étude</i>	102
C. <i>Forces de notre étude</i>	107
D. <i>Perspectives</i>	108
CONCLUSION	110
BIBLIOGRAPHIE	111
ANNEXES	115

LISTE DES ABREVIATIONS

GAN	Generative Adversarial Network
GPT	Generative Pre-trained Transformer
HAS	Haute Autorité de Santé
IA	Intelligence Artificielle
LLM	Large Language Model
MSP	Maison de Santé Pluridisciplinaire
NLP	Natural Language Processing
PENSO	Programme d'Enseignement du Numérique en Santé de l'Ouest
PDL	Pays de la Loire
SMOG	Simple Measure of Gobbledygook
URML	Union Régionale des Médecins Libéraux
WONCA	World Organization of National Colleges, Academies and Academic Associations of General Practitioners/Family Physicians ou World Organisation of Family Doctors
ANN	Artificial Neuron Network
API	Application Programming Interface
KDE	Kernel Density Estimation
PDL	Pays de la Loire

TABLE DES ILLUSTRATIONS

Figure 1 - Un réseau neuronal profond, en plusieurs couches (6).....	15
Figure 2 - Application du deep learning en santé (6).....	16
Figure 3 - Nombre de publications scientifiques sur l'IA par année	17
Figure 4 - Répartition des publications scientifiques par pays.....	17
Figure 5 - Investissements dans l'IA dans le monde, entre 2013 et 2023.....	18
Figure 6 - Chronologie non exhaustive des évolutions depuis le début de la conception de cette thèse.....	19
Figure 7 - Logo d'OpenAI	20
Figure 8 - Illustration de la réponse générée en quelques secondes par GPT-4o au prompt « Quels sont les symptômes de la méningite ? ».....	21
Figure 9 - Un médecin et son patient. Illustration photoréaliste générée via DALL-E, modèle créé par OpenAI (prompt : « Génère une image représentant un docteur assis à son bureau, en face de son patient, au cours d'une consultation. »).	22
Figure 10 - Illustration du raisonnement d'un LLM dans la génération d'une suite de mots (9).	23
Figure 11 - Actualisation du carré de White par Green (reproduit d'après Kandel, Bousquet et Chouilly) (36).....	27
Figure 12 - Evolution du nombre de références sur la base PubMed année par année (consulté en juin 2024). A : nombre de références pour l'équation de recherche « general practice » AND « artificial intelligence » ; B : nombre de références pour l'équation de recherche « general practice » AND « large language model »	30
Figure 13 - Répartition du nombre de cas clinique par tranche d'âge.	45
Figure 14 - Proportion de cas par type de source	45
Figure 15 - Répartition des cas par type de raisonnement associé.....	46
Figure 16 - Répartition des cas selon le type de diagnostic attendu	47
Figure 17 - Mise en place des tests et d'évaluation des modèles	56
Figure 18 - Répartition des participants par sexe.....	57
Figure 19 - Répartition des participants par tranche d'âge.....	57
Figure 20 - Répartition des secteurs d'activité des participants.....	58
Figure 21 - Répartition des milieux d'exercice des participants.....	58
Figure 22 - Répartition des participants par compétence en informatique	59
Figure 23 - Répartition des notes à la question "Sur une échelle de 1 à 10, quel serait selon vous le potentiel des IA génératives à transformer votre activité professionnelle au quotidien ?"	60

Figure 24 - Répartition des participants sur la fréquence d'utilisation des outils d'IA générative au quotidien	61
Figure 25 - Répartition des participants sur la fréquence d'utilisation professionnelle des outils d'IA générative	61
Figure 26 - Choix des participants sur les tâches dans lesquelles l'IA générative serait la plus utile	62
Figure 27 - Degré de favorabilité des médecins à l'arrivée des intelligences artificielles génératives dans leur pratique	63
Figure 28 - Répartition des réponses à la question "Dans quelle proportion de vos consultations pensez-vous que l'utilisation de l'IA générative pourrait vous aider de façon significative ?"	64
Figure 29 - Répartition des réponses sur le besoin de suivre une formation.....	65
Figure 30 - Répartition des réponses sur le format souhaité pour la formation	65
Figure 31 - Comparaison des écarts-types des longueurs des réponses, par modèle et par langue.	72
Figure 32 - Distribution des notes pour GPT-4, en français (à gauche) et anglais (à droite). 74	
Figure 33 - Distribution des notes pour Llama, en français (à gauche) et anglais (à droite). .75	
Figure 34 - Distribution des notes pour BioMistral, en français (à gauche) et anglais (à droite).	76
Figure 35 - Notes moyennes obtenues par item selon le modèle et la langue	78
Figure 36 - Comparaison des notes entre GPT en français (à gauche) et en anglais (à droite) sous forme de boîtes à moustaches.	79
Figure 37 - Comparaison des notes entre Llama en français (à gauche) et en anglais (à droite) sous forme de boîtes à moustaches.	80
Figure 38 - Comparaison des notes entre GPT en français (à gauche) et Llama en français (à droite) sous forme de boîtes à moustaches.	82
Figure 39 - Comparaison des notes entre GPT en anglais (à gauche) et Llama en anglais (à droite) sous forme de boîtes à moustaches.	83
Figure 40 - Moyennes obtenues par GPT en français selon la spécialité.....	85
Figure 41 - Comparaison en sous-groupes des notes de GPT en français sous forme de carte de chaleur.....	86
Figure 42 - Moyennes obtenues par GPT en anglais selon la spécialité.....	87
Figure 43 - Moyennes obtenues par Llama en français selon la spécialité	88
Figure 44 - Comparaison en sous-groupes des notes de Llama en français sous forme de carte de chaleur.....	89
Figure 45 - Moyennes obtenues par Llama en anglais selon la spécialité.....	90

Tableau 1 - Répartition finale du nombre de cas par libellé.....	42
Tableau 2 - Nombre de cas concernés par chaque "co-discipline".....	48
Tableau 3 - Grille d'évaluation des raisonnements diagnostiques des LLM, version finale....	54
Tableau 4 - Tableau récapitulatif des réponses.....	67
Tableau 5 - Minimums, maximums, et moyennes des longueurs des réponses fournies par les LLM	71
Tableau 6 - Moyennes et écarts-type des indices de SMOG selon le modèle, en anglais.....	73
Tableau 7 - Tableau des moyennes, écarts types, médianes et percentiles par modèle selon la langue.....	73
Tableau 8 - Moyennes par spécialité selon le modèle et la langue.....	84

Introduction

I. Introduction à l'intelligence artificielle

A. Définitions

1. *Qu'est-ce que l'intelligence artificielle ?*

Aborder la définition de l'intelligence artificielle (IA) est une tâche complexe en raison de la large gamme de sous-disciplines qu'elle recouvre. Nous n'aurons pas la prétention de couvrir intégralement ce sujet, mais nous devons d'apporter avant ce texte quelques définitions essentielles qui permettront à tout lecteur de comprendre nos travaux.

Le terme « d'intelligence artificielle » est utilisé pour la première fois, lors d'une conférence d'été à l'université de Dartmouth en 1956, organisée par John McCarthy et Marvin Minsky, deux des fondateurs de ce domaine. Il y est défini comme **le domaine des sciences et l'ingénierie dédié à la fabrication de machines intelligentes** (1). Cette définition est encore aujourd'hui reprise par l'université de Stanford, pionnière dans le domaine de l'intelligence artificielle (2).

2. *Une IA forte ou faible ?*

Au cœur de la révolution de l'IA, nous retrouvons deux concepts fondamentaux : celui d'**IA forte** et d'**IA faible**. Ces notions incarnent à la fois les espoirs et les dangers qui entourent l'intelligence artificielle. C'est dans un article publié en 1980 (3) que le philosophe John Searle introduit ces concepts :

- L'IA faible désigne des systèmes d'intelligence artificielle conçus pour effectuer des tâches spécifiques et limitées. Ces systèmes ne possèdent pas de conscience, de compréhension ou de véritable intelligence. Ils fonctionnent selon des algorithmes permettant l'exécution de certaines tâches prédéfinies. Bien qu'ils puissent simuler de manière convaincante des capacités humaines dans des contextes restreints, ils restent fondamentalement des outils sans cognition propre.
- L'IA forte, en revanche, fait référence à des systèmes d'intelligence artificielle dotés d'une conscience et d'une intelligence comparable à celles des humains. Cette idée implique que les machines pourraient un jour penser et comprendre de façon autonome, et posséder une forme de conscience semblable à celle de l'humain.

Il est crucial de comprendre que le concept d'IA forte reste purement théorique. A ce jour, aucun système d'IA ne s'approche de la conscience et de l'intelligence humaine.

Ces programmes ne relèvent que du domaine de la science-fiction. Si de telles intelligences artificielles venaient à voir le jour, de nombreux scientifiques mettent déjà en garde contre les dangers qu'elles pourraient représenter pour notre espèce (4).

3. Connexionnisme ou symbolisme ?

Nous avons pu observer aussi différentes façons d'approcher le développement de l'intelligence artificielle au fil du temps. Nous avons ainsi pu parler du connexionnisme ou du symbolisme.

L'IA symbolique, principalement développée dans les années 1950 et 1960, repose sur l'utilisation de symboles et de règles explicites pour représenter et manipuler des informations. Ces règles sont définies manuellement pour codifier le raisonnement. Les systèmes vont se baser sur ces règles pour inférer de nouvelles informations à partir de données préexistantes. Pour illustrer avec un exemple appliqué à la médecine, une règle d'un tel système pourrait être : si un patient a de la fièvre (F) **et** une toux (T), alors le patient pourrait avoir la grippe (G). Cette règle serait alors notée $F \wedge T \rightarrow G$.

L'IA connexionniste s'inspire du fonctionnement du cerveau humain et se concentre sur l'apprentissage à partir de données qui lui sont soumises, plutôt que sur des règles explicites. Ces systèmes nécessitent souvent de grandes quantités de données pour s'entraîner. Par exemple, dans le domaine de la santé, un système d'IA connexionniste pourrait être formé à partir d'une importante collection de photos de maladies dermatologiques, pour identifier des lésions suspectes chez des patients. Ainsi, en 2017, un groupe de chercheurs a mis en place un tel système, basé sur 130 000 images cliniques de cancers cutanés, permettant d'identifier des carcinomes avec une précision proche de celle d'un dermatologue confirmé (5). Comme nous le verrons plus tard, cette approche tend à supplanter le symbolisme, en grande partie grâce à l'essor des « big data » qui permettent d'entraîner ces modèles.

4. Une technologie en constante évolution

Nous aurons encore l'occasion de le voir, mais l'intelligence artificielle est un domaine en ébullition depuis plusieurs années. Cette effervescence est alimentée par **l'émergence rapide de nouvelles technologies innovantes**, et peut être difficile à suivre. Ces avancées technologiques redéfinissent constamment les capacités et les applications de l'intelligence artificielle, rendant le domaine dynamique et en constante évolution. Nous proposons dans cette section d'expliquer certaines de ces technologies afin d'aider à la compréhension de notre travail (6).

- **Réseaux de neurones artificiels :**

Les réseaux de neurones artificiels (en anglais *Artificial Neuron Network*, *ANN*) sont des systèmes informatiques conçus pour imiter le fonctionnement du cerveau humain. Ils se composent de couches de neurones interconnectés : une couche d'entrée qui reçoit les données, des couches cachées qui traitent l'information, et une couche de sortie qui donne le résultat final. Chaque neurone dans ces couches effectue des calculs simples et transmet les résultats aux neurones suivants.

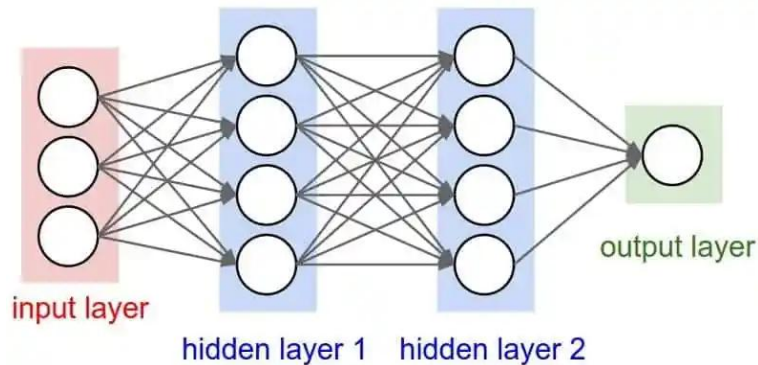


Figure 1 - Un réseau neuronal profond, en plusieurs couches (6).

- **Apprentissage automatique :**

L'apprentissage automatique (*machine learning*) est un domaine de l'intelligence artificielle qui cherche à permettre aux ordinateurs d'apprendre et de s'améliorer à partir de données, sans nécessiter une programmation explicite pour chaque règle. Ce processus repose sur l'utilisation d'algorithmes capables de détecter des motifs et des relations au sein des données. Il existe plusieurs types, l'apprentissage supervisé, non supervisé, et par renforcement. Il s'agit de l'application du concept d'intelligence artificielle connexionniste, où la machine « apprend » elle-même les règles.

- **Apprentissage profond :**

L'apprentissage profond (*deep learning*) est quant à lui un sous-domaine de l'apprentissage automatique qui utilise des réseaux de neurones artificiels avec de nombreuses couches, appelés réseaux neuronaux profonds. Ces réseaux sont capables de modéliser des abstractions complexes dans les données, ce qui les rend particulièrement efficaces pour traiter des images, des vidéos et des textes. Par exemple, dans le domaine de la santé, l'apprentissage profond est utilisé pour analyser des images médicales, comme des radiographies ou des IRM.

Grâce à leurs multiples couches, les réseaux neuronaux profonds peuvent automatiquement apprendre à identifier des caractéristiques simples dans les premières couches, comme les contours des organes, et des caractéristiques plus complexes dans les couches supérieures, comme des anomalies ou des tumeurs.

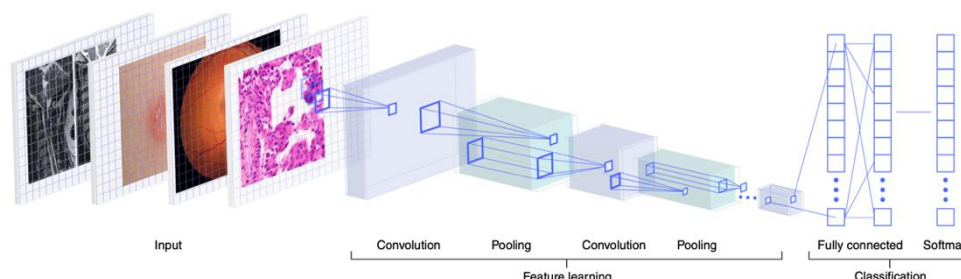


Figure 2 - Application du deep learning en santé (6).

- Traitement du langage naturel :

Le traitement du langage naturel (*Natural Language Processing, NLP*) est une discipline de l'intelligence artificielle qui se concentre sur l'interaction entre les ordinateurs et les humains par le biais du langage naturel. Le NLP permet aux machines de comprendre, interpréter et de répondre dans différentes langues de façon convaincante.

Ces technologies, parmi d'autres, font partie de l'écosystème de l'intelligence artificielle et sont essentielles pour comprendre les progrès et les applications de l'IA dans divers domaines comme la santé. En effet, aujourd'hui, la grande majorité des systèmes d'IA modernes utilisent le machine learning et ses sous-disciplines comme le deep learning. Ce sont grâce à ces avancées que les récentes applications ont permis d'arriver à des degrés de performance proches de l'être humain.

B. Essor de l'intelligence artificielle

Nous avons pu le voir précédemment, l'intelligence artificielle n'est pas une discipline récente. Néanmoins, au fil des dernières décennies, elle a connu une expansion remarquable, soutenue par des avancées technologiques significatives et un intérêt mondial croissant. Les progrès rapides du matériel informatique et les capacités de calcul des superordinateurs ont joué un rôle déterminant dans cette croissance. Selon le Centre pour la sécurité et les technologies émergentes, le nombre de publications scientifiques dédiées à l'IA a plus que doublé entre 2010 et 2021, passant d'environ 200 000 à plus de 450 000 articles (7). Cette augmentation témoigne de l'engagement accru de la communauté scientifique mondiale dans ce domaine en pleine expansion.

Annual scholarly publications on artificial intelligence

English- and Chinese-language scholarly publications related to the development and application of AI. This includes journal articles, conference papers, repository publications (such as arXiv), books, and theses.

Our World
in Data

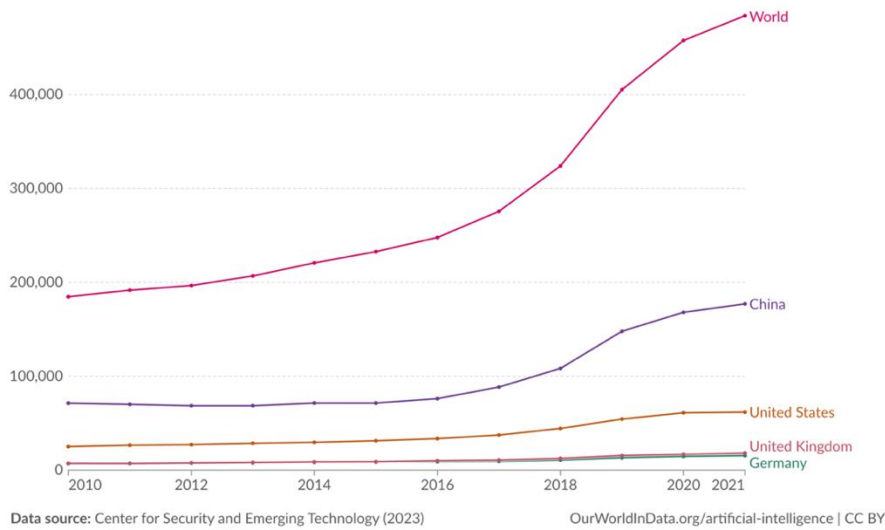


Figure 3 - Nombre de publications scientifiques sur l'IA par année

Dans cette dynamique globale, la France accuse un retard notable avec environ 9 500 publications en 2021, la plaçant au 10e rang mondial. En comparaison, ce chiffre est six fois inférieur à celui des États-Unis et près de dix-huit fois inférieur à celui de la Chine, soulignant un écart significatif à combler (7). Cette situation est particulièrement préoccupante pour le secteur de la santé, où les applications potentielles de l'IA pourraient révolutionner les pratiques médicales, des diagnostics aux traitements, en passant par la gestion des services de soins publics (8).

Annual scholarly publications on artificial intelligence, 2021

English- and Chinese-language scholarly publications related to the development and application of AI. This includes journal articles, conference papers, repository publications (such as arXiv), books, and theses.

Our World
in Data

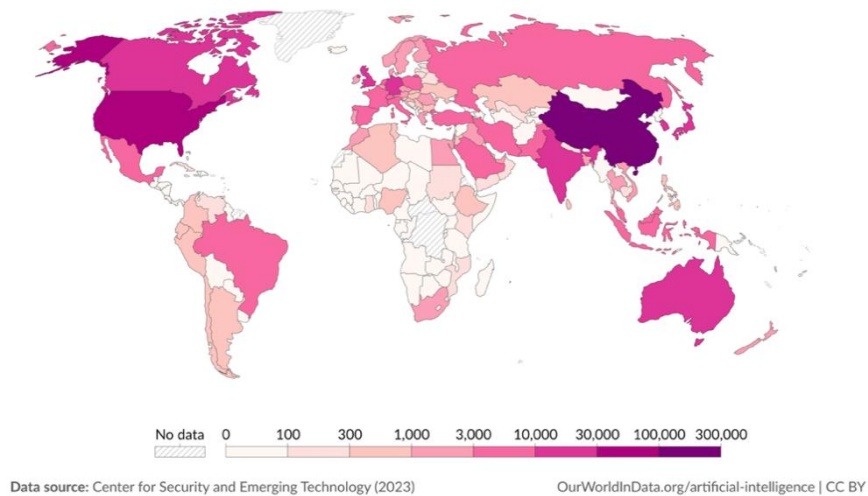


Figure 4 - Répartition des publications scientifiques par pays

Pour pallier ce retard, un rapport récent de la Commission de l'Intelligence Artificielle fait en mars 2024 propose des mesures ambitieuses, notamment un investissement annuel de 5 milliards d'euros sur cinq ans. Ce plan vise à renforcer la recherche et l'innovation en IA en France, à travers des initiatives telles que la formation des professionnels, la facilitation de l'accès aux données, et le soutien aux entreprises technologiques. Les projections économiques indiquent que l'adoption de l'IA pourrait potentiellement accroître le PIB jusqu'à 420 milliards d'euros, avec un impact positif sur l'emploi (9). Cependant, ces avancées ne sont pas sans risques, comme nous l'examinerons ultérieurement.

Parallèlement à ces initiatives nationales et devant de telles perspectives, les investissements mondiaux dans l'IA ont atteint des sommets, avec un pic à 337 milliards de dollars en 2021 selon le rapport annuel de l'université de Stanford (10), démontrant l'ampleur de l'engagement économique dans ce secteur en rapide expansion.

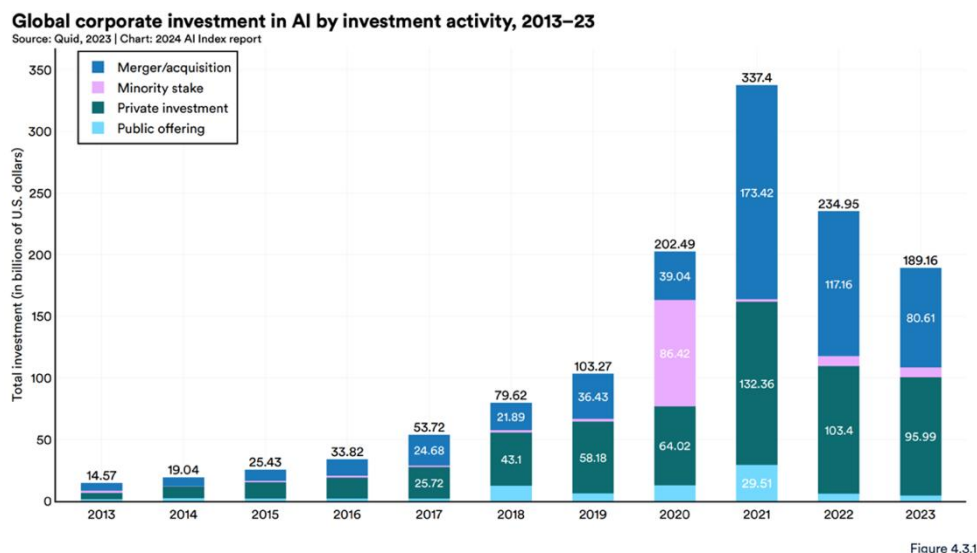


Figure 5 - Investissements dans l'IA dans le monde, entre 2013 et 2023.

Sur le plan réglementaire, l'Europe a pris une avance significative en adoptant en mars 2024 le « EU Artificial Intelligence Act », une législation régulant les systèmes d'IA au sein de l'Union Européenne. Cette initiative vise à équilibrer l'innovation et la protection des droits fondamentaux, en interdisant des pratiques comme par exemple la notation citoyenne, la surveillance de masse et l'identification biométrique à distance dans les lieux publics (11).

Il est d'une manière générale difficile d'être tout à fait à jour des données de la science en médecine, et cela est encore plus vrai avec l'intelligence artificielle. La seule période de réalisation de cette thèse d'exercice a vu évoluer le sujet sur bien des aspects qui n'ont pas tous pu être pris en compte. Ce dynamisme de la recherche rend le travail d'analyse et d'évaluation lent et parfois déjà dépassé avant même d'avoir été mené à bien.

Il nous paraissait utile de rappeler cet écart entre la vitesse d'évolution de l'IA et le temps nécessaire à la réalisation de travaux scientifiques rigoureux de l'autre. Les résultats et conclusions de nos travaux devront être regardés à l'aune de cette notion. La figure 6 illustre cette idée.

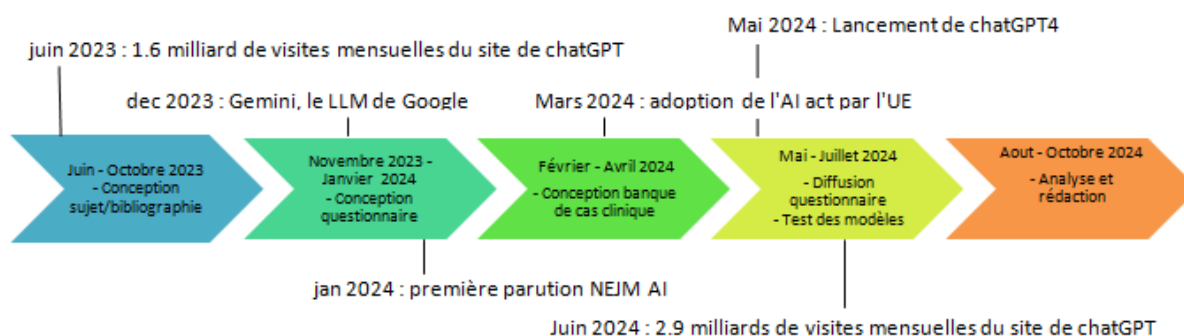


Figure 6 - Chronologie non exhaustive des évolutions depuis le début de la conception de cette thèse.

En conclusion, l'essor de l'IA représente une transformation majeure de notre société et ce dans de nombreux secteurs, et il est primordial que les médecins eux-mêmes participent à cette révolution sous peine de se voir y être soumis. Les actions entreprises à l'échelle mondiale et nationale dans ces domaines seront déterminantes pour façonner l'avenir et la position de la France et de l'Europe vis-à-vis de cette technologie prometteuse.

II. Intelligence artificielle générative

A. Remise en contexte

1. Qu'est-ce qu'une IA générative ?

Les travaux présentés dans cette thèse s'intéressent plus particulièrement aux intelligences artificielles génératives. Une intelligence artificielle générative est un système d'IA capable de produire du contenu. Ce contenu peut être au format texte, audio ou image. Il doit être suffisamment convaincant pour qu'on puisse croire qu'il a été créé par un être humain, passant ainsi le fameux « test de Turing ».

En 2014, Ian Goodfellow introduit le concept de réseaux antagonistes génératifs (*generative adversarial networks*, *GAN*). Il s'agit d'une classe d'algorithmes d'apprentissage non supervisé, où deux réseaux de neurones sont mis en compétition. Le premier réseau, appelé générateur, produit par exemple une image, tandis que son adversaire, le discriminateur, essaie de détecter si l'échantillon est réel ou bien généré artificiellement. L'arrivée de cette technologie permet la création de quantités de grandes données très crédibles (12).



Figure 7 - Logo d'OpenAI

En novembre 2022, l'entreprise américaine OpenAI met à disposition du grand public GPT-3, un chatbot utilisant l'intelligence artificielle pour répondre aux requêtes des utilisateurs. GPT, ou *Generative Pre-trained Transformer*, est un large modèle de langue (*large language model, LLM*), c'est-à-dire un système d'intelligence artificielle capable de produire du texte. Il entre donc dans la catégorie des intelligences artificielles génératives. Ces modèles sont créés grâce à l'apprentissage profond (*deep learning*) sur de très grandes quantités de données. GPT-3, par exemple, a été entraîné sur un vaste corpus de textes. Ces données restent le plus souvent confidentielles pour des raisons de sécurité, mais on sait que GPT-3 est constitué d'environ 175 milliards de paramètres. Les versions ultérieures disposent d'un plus grand nombre de paramètres, d'un *tuning* (c'est-à-dire un ajustement fin du modèle à certaines tâches), et d'une vitesse améliorée (13).

A l'heure de l'écriture de cette thèse, OpenAI a sorti, le 13 mai 2024, la version GPT-4o avec des capacités de synthèse vocal et de reconnaissance visuelle.

Pour utiliser un modèle de langue, il suffit de lui présenter une requête, qu'on appelle aussi un ***prompt***, et le bot génère une réponse en se basant sur ses connaissances. Les applications possibles des intelligences artificielles génératives sont vastes et variées, allant de la création de contenu artistique à l'automatisation de tâches complexes.



La méningite est une inflammation des méninges, les membranes entourant le cerveau et la moelle épinière. Les symptômes peuvent varier en fonction de la cause (virale, bactérienne, fongique, etc.), mais les signes classiques incluent :

1. **Fièvre** élevée soudaine
2. **Maux de tête** sévères
3. **Raideur de la nuque**
4. **Nausées et vomissements**
5. **Photophobie** (sensibilité à la lumière)
6. **Confusion** ou altération de la conscience
7. **Somnolence** ou difficultés à se réveiller
8. **Convulsions** dans certains cas graves
9. **Éruption cutanée** (surtout dans les formes bactériennes comme la méningite à méningocoques)
10. **Troubles sensoriels** comme l'intolérance au bruit

Chez les **nourrissons**, les symptômes peuvent être plus difficiles à détecter, avec des signes tels que :

- Refus de s'alimenter
- Irritabilité excessive
- Fontanelle bombée (chez les bébés)
- Pleurs constants

Si une méningite est suspectée, il est important de consulter immédiatement un médecin, car certaines formes, notamment la méningite bactérienne, peuvent être rapidement mortelles.

Figure 8 - Illustration de la réponse générée en quelques secondes par GPT-4o au prompt « Quels sont les symptômes de la méningite ? »

Nous prenons l'exemple de ChatGPT, car depuis sa sortie en novembre 2022 il est devenu l'outil d'intelligence artificielle générative le plus populaire en France et dans le monde (14). Néanmoins, on peut retrouver de nombreux autres exemples, comme Llama développé par Meta ou MistralAI, un modèle français appartenant à l'entreprise Mistral.

Ces outils sont aussi capables de produire des sons ou des images.



Figure 9 - Un médecin et son patient. Illustration photoréaliste générée via DALL-E, modèle créé par OpenAI (prompt : « Génère une image représentant un docteur assis à son bureau, en face de son patient, au cours d'une consultation. »).

2. Comment fonctionne une IA générative ?

Nous en avons déjà brièvement parlé, les systèmes d'IA générative se basent sur les techniques d'apprentissage automatique. Ils sont entraînés sur des données préexistantes. Celles-ci peuvent être dans différents formats : texte, audio, image, vidéo...

Les systèmes peuvent être entraînés sur de très grandes quantités de données, formant ce qu'on appelle aussi des modèles de fondation. Certains systèmes peuvent être affinés, en sélectionnant les données d'entraînement, à la réalisation de certaines tâches. On parle à ce moment-là de modèles affinés ou ajustés.

Lorsque le système reçoit une instruction, il décompose le texte en ce qu'on appelle des *tokens*. Un token correspond à une série de lettres, en général plus petit qu'un mot complet. Il établit alors des probabilités pour déterminer le token suivant dans la phrase. Il est important de comprendre que le système ne comprend pas les informations. Il n'a pas de réelle conscience ou d'intelligence, ne cherche pas vraiment de réponse ou de « vérité », il établit uniquement des probabilités et détermine la suite logique de la phrase. Cela est essentiel car on ne peut pas établir que les IA génératives soient entièrement fiables (étant donné qu'elles n'appréhendent pas le sens des mots ou des images).

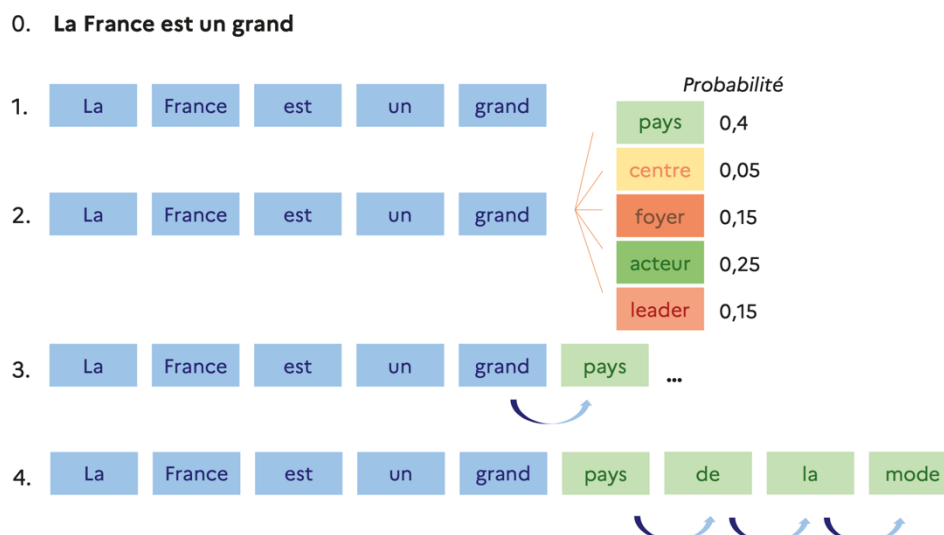


Figure 10 - Illustration du raisonnement d'un LLM dans la génération d'une suite de mots (9).

B. Promesses et dangers

Les modèles d'IA générative possèdent des aptitudes impressionnantes. De plus en plus d'études sont faites, montrant leurs capacités dans la réalisation de tâches complexes. Une étude récente a montré que ces algorithmes présentaient des meilleurs résultats que 90% des candidats à certains examens du barreau aux États-Unis (15).

Leurs applications peuvent être diverses et variées. Cela va de la découverte de nouvelles molécules par AlphaFold (16) à la suggestion de nouveaux matériaux pour équiper nos futures voitures (17). Une étude américaine a montré que l'introduction de l'IA générative auprès de 5000 agents d'un support client a augmenté leur productivité de 14%, avec un gain plus notable chez les travailleurs novices (18).

De la même façon que toutes les nouvelles technologies, elles s'accompagnent aussi de risques non négligeables.

Les modèles d'IA générative étant entraînés sur des corpus de texte confidentiels, il est difficile d'en juger la qualité. Il en découle ce qu'on appelle un effet « boîte noire » où on ne peut juger du procédé de l'algorithme pour générer le résultat. Ces systèmes raisonnant par probabilités, ils peuvent parfois générer des informations inventées, c'est ce qu'on appelle une « hallucination ». Cette limitation est importante surtout dans le domaine de la santé où les résultats peuvent avoir des conséquences graves.

Il n'est pas toujours possible de juger de la qualité des données sur lesquelles sont entraînées les IA génératives privées. Elles peuvent donc être sujettes à débats, que ce soit d'un point de vue scientifique mais aussi éthique. En effet, si une IA est capable de produire du texte comme un être humain, il est aussi capable de dériver de la même façon.

Ainsi, en 2021, une IA sud-coréenne supposée incarner une adolescente de 20 ans s'est mise à proférer des insultes envers les minorités sexuelles (19).

La génération de contenus réalistes pose de nombreux problèmes. Un rapport publié par l'organisme non gouvernemental Freedom House explique que l'intelligence artificielle générative menace de surcharger Internet de campagnes de désinformation. Selon ce rapport, cette technologie a été utilisée par au moins 16 pays pour créer du doute et influencer le débat public au cours de l'année 2023. L'intelligence artificielle a par ailleurs poussé de nombreux gouvernements à revoir et intensifier leur politique de censure en ligne (20).

La confidentialité et la collecte des données est toujours au centre du débat, surtout dans le milieu médical. Les règles autour de celles-ci peuvent parfois être très floues selon les modèles utilisés. En avril 2023, OpenAI a été la cible de deux plaintes en France pour non-respect des règles européennes sur la récolte de données personnelles (21). Leur plateforme avait même été bloquée par l'Italie en mars 2023 pour ces mêmes raisons.

La souveraineté des données, la désinformation, la confidentialité, l'écologie (que nous n'avons pas développé ici) sont autant d'enjeux et de défis à relever pour installer de façon sereine l'intelligence artificielle générative dans notre milieu.

III. Intelligence artificielle et santé

A. Applications de l'IA en santé, quelques exemples

Les modèles d'IA générative semblent capables de répondre correctement à nombre de questions et d'usages, aussi complexes soient-ils. Le domaine de la santé est invariablement cité comme l'un des domaines d'application privilégié pour cette technologie.

Tous les acteurs de la santé et tous les aspects de leur pratique professionnelle sont concernés. Cela va de l'organisation des soins à l'orientation des patients, de l'implémentation des dossiers informatisés à l'interprétation d'examens paracliniques, en passant par le suivi, les choix thérapeutiques, la communication et l'information aux patients (22). La quasi-totalité des champs de l'intelligence artificielle – reconnaissance d'image, de vidéos, traitement du langage naturel, apprentissage automatique, robotique, etc. – peuvent trouver des applications en matière de santé.

Il nous paraît vain de proposer une liste exhaustive des applications de l'IA en santé, ne serait-ce que du fait de l'augmentation constante du nombre de ces applications. Nous ne donnerons donc que quelques exemples marquants pour illustrer les possibilités extrêmement larges d'utilisation de ces outils en santé.

Il est logique que ce soient surtout les domaines pour lesquels il existe des grands ensembles de données, typiquement les images, qui aient été les terrains les plus fertiles au développement de l'IA en santé. Les progrès de l'analyse automatisée ont été fulgurants en imagerie médicale. Parmi les algorithmes d'IA pour la santé évalués par la FDA, plus de 76% concernaient la radiologie (23).

L'intelligence artificielle est utilisée de multiples façons en médecine. En premier lieu, elle intervient dans la lecture automatisée d'images, permettant le diagnostic. Par exemple, *Teare et al.* (24) décrivent un système de détection de processus malins dans les mammographies aussi performant que les radiologues experts. L'IA peut aussi assister des radiologues pour des tâches chronophages ou fastidieuses. Des solutions sont déjà en place comme Arterys qui mesure les nodules pulmonaires et hépatiques ou Viz.ai qui détecte automatiquement les thrombus sur des scanners (25). Dans ces cas, l'IA assure une partie du travail du radiologue. En oncologie, elle pourrait optimiser les doses de rayons reçues limitant ainsi les effets secondaires pour le patient (8). Les performances quant à l'automatisation de la lecture d'image sont telles que ces outils permettraient de voir ce que le radiologue ne voit pas, donnant naissance à un nouveau domaine : la *radiomique*.

Même lorsque l'IA est moins performante que l'humain seule, elle semble améliorer de manière synergique les performances de ce dernier (26). Ces systèmes ne se cantonnent pas aux analyses complexes, mais permettent aussi la lecture automatisée d'examens courants et largement utilisés, comme la détection de fracture sur des clichés standards (27) ou l'analyse de radiographies de thorax de dépistage (28).

L'anatomo-pathologie est aussi largement investie par l'IA. Des travaux montrent la précision de l'intelligence artificielle dans le diagnostic pathologique (29). Certains montrent même une meilleure performance de l'IA par rapport aux humains pour la classification de certaines lésions (30). En dermatologie, la concurrence des IA pour la caractérisation clinique de lésions cutanées suspectes de cancer semble déjà être une réalité (5). En ophtalmologie, la lecture automatisée des images de fond d'œil pour le diagnostic d'affection rétinienne est là aussi probante (31).

La liste pourrait se poursuivre. Notons que les exemples les plus médiatisés, comme ceux que nous avons cités, concernent surtout des tâches très spécifiques qui ne recouvrent certainement pas toute la complexité de la pratique médicale concernée. Il s'agit souvent d'interprétation de données paracliniques, et d'utilisation dans des contextes très spécifiques.

Les systèmes d'IA semblent destinés à se développer pour couvrir l'ensemble des tâches habituellement attribuées aux médecins. En effet, au-delà de leur savoir-faire technique, certaines compétences sont souvent jugées comme purement humaines.

On peut citer en particulier l'empathie, l'intuition, la réalisation de gestes complexes, l'expérience, ou même le raisonnement médical en général. Ces compétences sont souvent citées comme étant inaccessibles pour l'intelligence artificielle. Néanmoins, comme nous le rappelle le Dr Molcard dans sa revue de la littérature (25), il serait possible de présenter pour chacune de ces compétences au moins un article montrant qu'elles ne sont plus l'apanage de l'humain. Pour ne citer que l'empathie¹, un article paru en 2023 a montré une meilleure empathie dans les réponses d'un LLM à des questions de santé par rapport à celles de professionnels humains (32).

Malgré tout, n'oublions pas qu'il existe une différence de nature fondamentale entre l'échange avec un LLM et un humain. Les modèles de langue génèrent du texte par des moyens stochastiques, sans intégrer de représentation logique ou de critères de vérité et ils ne connaissent pas les émotions au sens humain du terme (22). Aussi impressionnante qu'elle puisse paraître, l'empathie exprimée par ces modèles reste simulée. Certains auteurs commencent à évoquer cette idée, rappelant l'importance de l'empathie réelle dans le contexte du soin (33).

Ces considérations nous invitent à rester critiques vis-à-vis des performances annoncées de l'IA sur ces aspects de la santé et le travail du soignant. Au-delà de la performance brute, il sera peut-être nécessaire de s'interroger sur ce que ces avancées apportent concrètement au patient et au système de santé en général.

B. Médecine générale et intelligence artificielle

La médecine générale se distingue des autres spécialités médicales par bien des aspects. Son histoire, sa pratique, les situations qu'elle rencontre, et la place de l'omnipraticien dans le système de santé et dans la société en font une discipline singulière. Bien loin des questionnements autour de la lecture automatique d'image de radiologie ou de lames d'anatomopathologie, le généraliste et sa pratique peu technique peuvent parfois sembler exclus de l'imaginaire autour du futur de l'IA en santé (34).

Nous rappelons ici quelques concepts clés qui font la spécificité de la discipline. Il s'agit de quelques repères pour mettre en perspective la question des usages de l'intelligence artificielle générative en médecine générale.

La médecine générale, est une médecine de premier recours. C'est celle qui ne s'intéresse ni spécifiquement à un organe, ni à un type de patient. Celle que l'on sollicite directement, sans adressage préalable. Le médecin généraliste est le plus souvent le premier contact des patients avec le système de santé.

¹ L'empathie est la faculté intuitive de se mettre à la place d'autrui, de percevoir ce qu'il ressent.

C'est le médiateur entre le patient et les soins plus spécialisés (35). Il s'agit d'entretenir une relation prolongée avec les patients, allant de la prévention et du dépistage au suivi de pathologies chroniques, de la prise en charge de certaines urgences aux pathologies fréquentes, sans oublier un rôle de conseil et d'accompagnement. Pour la WONCA, l'omnipraticien apporte une réponse à une grande majorité de besoins individuels (36).

En soins primaires, l'incidence et la prévalence des maladies sont faibles. Les symptômes n'y sont pas toujours synonymes de maladie (37), de plus, le généraliste est celui qui est contraint d'agir à un stade précoce et non différencié des maladies (35). Le généraliste prend en charge des problématiques propres à sa pratique, qui ne seront le plus souvent prises en charge que par lui. Le Carré de White donne un aperçu de cette place particulière qu'occupe la médecine générale dans le système de santé et des problématiques qu'elle rencontre. Cette figure, proposée en 1961, est un carré représentant 1000 habitants exposés à un « problème de santé » pendant 1 mois.

Chaque carré plus petit, représente l'effectif de patients parmi ces 1000 personnes qui iront vers telle ou telle orientation dans le système de soin. Il illustre la différence majeure entre le nombre de personne « exposées à un problème de santé », et le nombre de personne entrant dans le système de soin. Parmi ces dernières, seulement quelques unes consulteront un spécialiste, et encore moins seront hospitalisées (figure 11) (38).

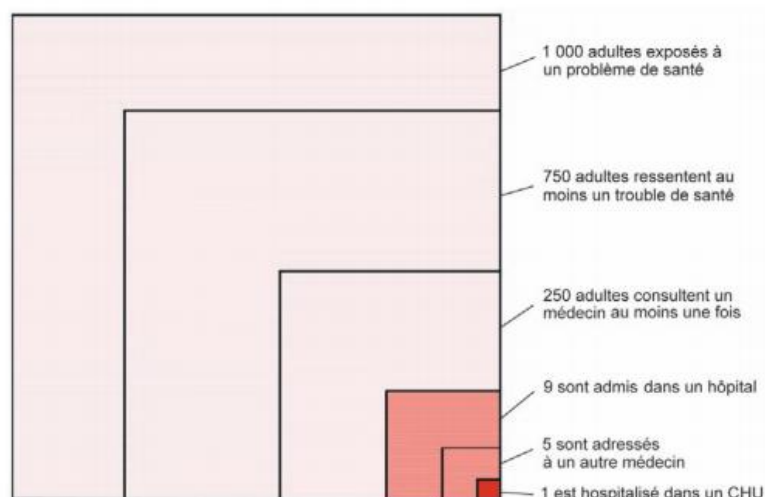


Figure 11 - Actualisation du carré de White par Green (reproduit d'après Kandel, Bousquet et Chouilly) (35).

Les patients et les problématiques de santé vus à l'hôpital ne sont pas les mêmes que ceux vus par le médecin généraliste. Entre autres conséquences, cette réalité impacte l'application des données hospitalières à la médecine générale.

1. Le diagnostic en médecine générale :

Du fait des problématiques de santé qu'il rencontre et des moyens à sa portée, le généraliste peut rarement poser un diagnostic de certitude au moment de la consultation. Il devra souvent se contenter d'évaluer et de nommer la situation à travers un « tableau clinique ». La plupart de ses décisions sont prises en contexte d'incertitude diagnostique. Les mêmes symptômes peuvent être le signe d'une pathologie tout à fait bénigne comme le mode de révélation d'une pathologie plus grave. Tout tableau clinique n'est pas le stade précoce d'une maladie et la plupart des états morbides vus par le médecin généraliste n'évolueront pas et seront spontanément résolutifs (jusqu'à 70% selon la société française de médecine générale (39)). Cela sous-entend que dans un certain nombre de cas, il n'y a pas de diagnostic.

Le médecin doit donc savoir ne pas se lancer dans des investigations ou traitements pouvant être inutiles, coûteux et anxiogènes. Nommer le diagnostic implique donc certaines subtilités. Un exemple classique : on préférera le terme « état fébrile » à « syndrome grippal ». On évite alors de confondre effets et causes, hypothèses et diagnostics. Il existe tout un continuum pour ce que l'on appelle « diagnostic », de la caractérisation des symptômes en termes médicaux au diagnostic final. Selon Pouchain, seulement 9% des consultations aboutissent à un tel résultat (35). Aboutir à un diagnostic précis final n'est donc pas une fin en soit, cela d'autant que dans la réalité de la pratique médicale, un diagnostic formel n'est souvent pas indispensable à la prise de décision.

Rappelons également que la plainte du patient se manifeste dans un contexte humain, culturel et matériel complexe avec une interaction forte entre les composantes multiples de sa situation. On parle parfois de diagnostic de situation, plus large que le diagnostic de maladie. Il englobe le contexte, les dimensions relationnelles affectives et les liens entre toutes ces dimensions.

Le concept de diagnostic est mouvant au même titre que celui de pathologie. Ce que nous appelons maladie ou considérons comme pathologique est en évolution. L'évolution des sociétés humaines dans leurs usages, leurs comportements et leurs rapports à l'environnement fait évoluer en permanence les problématiques de santé qui les concernent. C'est le concept de pathocénose. Il existe de nombreux exemples illustrant cette dynamique : la mondialisation favorise de nouvelles épidémies comme la Covid-19. Les troubles neurocognitifs majeurs n'existaient pas lorsque l'espérance de vie ne dépassait pas 50 ans. La notion de maladie et son identification évoluent aussi au fur et à mesure de l'avancée des connaissances scientifiques, mais aussi en fonction des représentations et des préoccupations de la société à leur égard.

De nouvelles maladies apparaissent désignant ce qui étaient peut-être des pathologies anciennes (l'endométriose, la maladie coéliquue). Certaines pathologies ne sont plus considérées comme telles (l'OMS a cessé de considéré l'homosexualité comme un trouble psychiatrique en 1992 (40)). L'évolution des pratiques médicales apporte aussi son lot de nouveautés, on parle désormais d'incidentalome par exemple, pour désigner les lésions asymptomatiques, découvertes sur des examens qui n'avaient pas pour but de trouver cette lésion. De ces évolutions émergent alors le concept de prévention quaternaire, et la nécessité de ne pas induire des troubles là où il n'y en a pas : *primum non nocere*.

Le concept de diagnostic, particulièrement en médecine générale, est donc beaucoup plus difficile à appréhender qu'il n'y paraît. Le médecin généraliste (et le médecin humain d'une manière générale) aura toujours pour lui cette particularité d'évoluer dans la même société que son patient, et qu'il est de ce fait bien placé pour y percevoir les évolutions au jour le jour. Il est un témoin privilégié de la pathocénose.

Enfin, rappelons que le rôle du médecin généraliste, ne se cantonne pas toujours au couple diagnostic/prescription. Un grand nombre de consultations ne posent pas de questions diagnostiques, qu'il s'agisse des consultations de prévention, de suivi de l'enfant, de discussion autour de la santé et de conseils, etc.

2. Le raisonnement en médecine :

La faible prévalence des pathologies contraint le plus souvent le médecin à une démarche probabiliste. Il est classique de segmenter la démarche diagnostique, en temps d'interrogatoire, examen physique et synthèse. Néanmoins ces différentes étapes se chevauchent, n'existent pas, ou n'interviennent pas dans cet ordre. Les raisonnements mis en place par le clinicien ont en réalité lieu tout au long de la consultation. Ces raisonnements sont nombreux et divers. Ils permettent au médecin de cheminer du motif vers le « diagnostic » retenu. Il s'appuie pour cela sur la sémiologie apprise et son expérience.

Il existe plusieurs types de raisonnements diagnostiques. Il s'agit d'un phénomène dynamique et continu, plus complexe que ses descriptions théoriques (voir annexe 2).

Plusieurs difficultés inhérentes à la rencontre de soins constituent des niveaux supplémentaires de réflexion dans la démarche clinique. Le motif de consultation et la manière dont il est présenté peuvent nécessiter une attention particulière. Au-delà du motif annoncé, on peut retrouver des motifs cachés. Il n'est alors pas toujours simple de saisir l'attente réelle du patient. Le motif sera souvent multiple, complexifiant encore la démarche. Tous ces aspects du langage du patient doivent être décryptés.

Pour toutes ces raisons, les besoins du médecin généraliste sont différents de ceux de ses confrères spécialistes. Il est important de garder à l'esprit que cela sera probablement également vrai pour l'intelligence artificielle appliquée à ce domaine.

3. Recherche en médecine générale

a) Généralités

A l'instar de la recherche médicale globale, il semble que la recherche sur l'IA en soins primaires et en médecine générale soit moins prolifique que dans d'autres domaines comme l'imagerie.

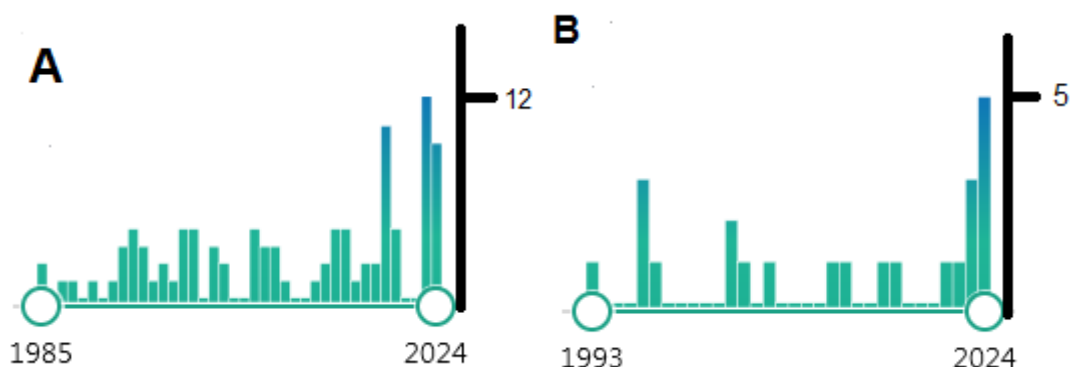


Figure 12 - Evolution du nombre de références sur la base PubMed année par année (consulté en juin 2024). A : nombre de références pour l'équation de recherche « general practice » AND « artificial intelligence » ; B : nombre de références pour l'équation de recherche « general practice » AND « large language model »

Une équipe canadienne a réalisé une revue de la littérature sur 405 articles traitant de l'IA en soins primaires (41). Dans ce travail, seulement 14% des articles sur l'IA appliquée à la médecine de premier recours avaient parmi leurs auteurs un médecin généraliste, la majorité des auteurs étant informaticiens. Leur travail montre aussi que moins de 7% des articles concernent un essai de technologie en conditions réelles d'utilisation.

Un article de 2021 sur l'utilisation de l'IA en soins primaires retrouve aussi une implication quasi-inexistante des soignants dans la conception et l'évaluation des systèmes d'IA utilisés (42). Cette étude rapportait une tendance à la hausse du nombre de publication autour de l'IA en soins primaires. Dans ce même article, la France n'arrivait qu'en dixième position des pays en nombre de publications sur l'utilisation de l'IA en soins primaires. Le Dr Molcard a réalisé un travail de revue de la littérature sur l'intelligence artificielle et la médecine générale pour sa thèse d'exercice (25). Il explique que l'IA en médecine générale est à un stade très précoce et que s'il existe de nombreux articles qualitatifs pour comprendre les usages, peu d'articles démontrent l'utilisation en conditions réelles d'outils fonctionnels.

Cet état de fait doit faire tempérer l'enthousiasme autour des avancées de ces technologies, du moins en médecine générale. Par ailleurs, la place de la médecine générale dans la recherche médicale laisse supposer que les données sur lesquelles les intelligences artificielles sont entraînées comportent peu de données produites par des généralistes dans leur pratique. C'est une limite non négligeable à l'utilisation de ces technologies dans ce contexte. Par ailleurs, et c'est un argument supplémentaire qui plaide pour l'implication des médecins généralistes dans le développement de ces outils, les solutions proposées par les start-up et autres industriels ne prennent pas nécessairement en compte les aspects particuliers de la pratique de l'omnipraticien détaillés plus haut.

b) Utilisation actuelle de l'IA par les médecins généralistes

Notre recherche documentaire n'a pas permis de répondre de manière satisfaisante à la question de la proportion d'utilisateurs de l'IA en pratique courante. Les quelques données disponibles semblent discordantes. L'entreprise PulseLife par exemple dans son « Baromètre IA en santé » en 2024 s'appuyait sur un questionnaire adressé aux professionnels de santé utilisant leurs propres applications pour annoncer une proportion de 53% de soignants intégrant l'IA à leur pratique quotidienne (43). Cela semble surprenant au regard de ce que dit l'Académie Nationale de médecine dans son rapport sur l'IA, expliquant que les technologies d'IA génératives « *trouvent peu d'écho auprès des cliniciens* » (44). C'est aussi l'avis de la Haute Autorité de Santé (HAS) concernant les dispositifs médicaux numériques embarquant de l'IA, rappelant à ce sujet que « leur utilisation et leur adoption par les professionnels restent encore limitées en pratique courante de soins » (45). Le Dr Molcard dans sa revue de la littérature, n'apporte pas de réponse chiffrée à cette question. Il cite une étude allemande de 2022 (46), portant sur des entretiens avec 18 médecins généralistes, retrouvant seulement 22% d'entre eux comme ayant « une expérience » de ces technologies « sans pour autant les avoir utilisés dans leur pratique ».

Pour le Dr Moukrim dans sa thèse de 2019 (47), 30% des médecins interrogés (187 réponses de médecins généralistes français) déclaraient avoir déjà utilisé une technologie d'IA. Notons que dans ce travail, « intelligence artificielle » était accepté dans un sens très large, dépassant de loin les seules IA génératives. En tenant compte d'un probable biais de sélection vis à vis des réponses, émanant probablement de médecins déjà sensibilisés au sujet, il est possible que cette proportion soit surestimée.

La difficulté à trouver des données précises sur l'utilisation des IA en médecine générale nous renvoie à la difficulté de définir la notion d'intelligence artificielle.

Les arguments marketing des entreprises proposant ces solutions ne doivent pas nous faire perdre de vue que l'utilisation d'IA générative en pratique courante ne semble pas, dans l'immédiat, largement répandue chez les généralistes français.

Par ailleurs, cette question nous pose aussi celle des applications disponibles. Quelles sont les solutions d'IA validées et accessibles au médecin généraliste pour sa pratique courante ? Un inventaire exhaustif des solutions existantes est difficile à réaliser. Les données disponibles, largement proposées par des entreprises privées vendant ces technologies, ne nous ont permis d'obtenir qu'une idée assez floue de ce qu'il existe réellement.

Varghese *et al* en 2020 (48) met les chiffres de l'augmentation des travaux de recherche sur l'IA en perspective. En incluant dans sa recherche de publications sur l'IA en santé le critère « essai clinique », reflétant théoriquement le nombre de dispositifs à l'essai sur le terrain, il montre que ce nombre a peu changé depuis 2005. On retrouve l'idée que malgré un grand enthousiasme, peu de ces dispositifs sont véritablement prometteurs même si l'offre semble s'étoffer.

Lorsque nous interrogeons ChatGPT sur les applications d'IA générative disponibles en France pour les médecins généralistes, la réponse est décevante. Aucune des applications citées n'est orientée vers la médecine générale. La consultation des sites commerciaux de ces applications n'a pas retrouvé de données concernant le nombre précis d'utilisateurs.

Parmi les applications les plus médiatisées, on peut citer Nabla (49), qui a pour but d'enregistrer les données de consultation et de générer des comptes rendus ou des ordonnance pré remplies. Nabla serait utilisé par « 30 000 praticiens ». La plateforme Doctolib propose une solution équivalente, avec un assistant de consultation en cours de développement (50). Il s'agit là surtout d'automatiser des tâches pouvant être fastidieuses, en particulier l'implémentation des dossiers informatisés.

c) Concernant les grands modèles de langue ?

L'irruption de ChatGPT auprès du grand public a rendu bien plus concrète la possibilité d'utilisation de l'IA générative au quotidien. Nombre de travaux se sont alors intéressés à l'utilisation des LLM en pratique médicale. Ces articles ne semblent pas vraiment apporter d'informations sur l'utilisation en pratique courante de telles technologies en médecine générale en France.

L'article de Rahimi (42) sur l'utilisation de l'IA en soins primaire retrouve dans sa revue de la littérature une majorité de technologies utilisées basées sur le machine learning. Nombre d'articles étudiant ChatGPT s'attachent à le tester sur des examens de médecine comme l'ECN.

Il n'est pas étonnant de voir que ces programmes, entraînés sur des grands corpus de textes de référence, performant très bien sur ce genre de tâche. Là encore, des tests « en pratique réelle » ne semblent pas avoir été largement effectués.

Les articles déjà mentionnés portaient sur de nombreux aspects de la pratique médicale. Très souvent sur le raisonnement clinique et diagnostique, mais aussi sur la documentation clinique (*implémentation des dossier, recherche de données utiles*), la recherche (51), jusqu'à l'aide à la rédaction d'article scientifiques ou la discussion de problèmes de santé publique (52).

Malgré ces éléments nous ne pouvons pas répondre précisément à la question de l'utilisation de l'intelligence artificielle générative en pratique courante par les médecins généralistes français. Il semble qu'à l'heure actuelle, en France, il existe peu de systèmes d'intelligence artificielle générative à destination des généralistes validées et largement utilisés en pratique courante.

d) L'IA pour quoi faire ?

Au-delà du nombre d'utilisateur et des technologies utilisées, pourquoi les médecins généralistes auraient-ils besoin d'IA générative ? Nous l'avons vu, les promoteurs de ces systèmes proposent d'automatiser de très nombreuses tâches, certaines jusqu'à présent dévolues au seul médecin. Mais qu'en attendent vraiment ces derniers ? Si les données disponibles sur l'utilisation réelle de systèmes d'IA par les généralistes sont limitées, elles semblent plus nombreuses concernant leurs attentes. Le Dr Molcard, déjà cité, mentionne l'article d'une équipe canadienne (41) dans laquelle les intervenants, issus de diverses professions (professionnels de santé, informaticiens, décideurs ou politiques, managers) étaient interrogés afin de dégager des sujets d'intérêt pour une utilisation efficace de l'IA en médecine générale. Les auteurs mentionnaient les 9 points suivants, par ordre de priorité :

1. Prévention et détermination du risque : aide décisionnelle en cas d'incertitude diagnostique, et libérer du temps dans la consultation
2. Gestion de sa maladie par le patient lui-même
3. Gestion et synthèse de plusieurs sources d'information médicale
4. Meilleure communication entre les médecins et les spécialistes en IA
5. Partage des données entre médecins
6. Aide à la décision clinique, par exemple recommandations individuelles.
7. Appui administratif : planification des visites, rendez-vous, partage de l'information nécessaire avant la consultation, etc.

8. Rôle de clerc et d'assistant documentaliste : diminuer la charge de tâches de routine (documentation des cas, transcription automatique du dialogue entre patient et médecin, génération automatique de l'observation)
9. Améliorer la prise en charge psychiatrique, par exemple en diminuant l'effet d'ancrage (le fait de s'attacher à une première impression) en suggérant la poursuite de l'entretien sur d'autres aspects

Le Dr Moukrim, déjà cité, a étudié l'effet de l'IA envisagé par des médecins généralistes sur tel ou tel aspect de leur pratique. Les items liés aux diverses tâches du médecin ont été classés par les participants en fonction de l'impact attendu de l'IA sur ces tâches, que cet impact soit positif, négatif ou neutre. Pour la relation médecin malade, une majorité d'items étaient étiquetés « effet négatif ». Une majorité de réponses « effet positif » avaient été données pour la prise en charge au long cours et les conditions d'exercice des généralistes. Il y avait aussi une majorité d'« effets positifs » pour la démarche décisionnelle, mais moins marqué, avec beaucoup de réponses « effet neutre » (47). Notons que dans ce cas la méthodologie utilisée ne met pas en avant les attentes concrètes des médecins.

Chez le Dr Tabla (53), on retrouve des attentes similaires (évaluation du risque et médecine personnalisée, assistant documentaliste, implication du patient dans sa prise en charge). Il faut noter que les éléments qui seraient dissuadant quant à l'utilisation de ces modules hypothétiques étaient ceux renvoyant à une perte de contrôle sur la décision finale.

Les Dr Lafarge et Derradji (54), sur la base d'entretiens qualitatifs avec 14 généralistes ont mis en avant six utilisations préférentielles : collecter les données, gérer l'incertitude, améliorer la prise en charge et résoudre les situations complexes, aider au diagnostic des maladies peu fréquentes, *empowerment*² du patient, interprétation automatique de certains examens paracliniques, organisation du cabinet.

Ainsi, la plupart des travaux s'accordent sur la question des attentes des médecins. Notons qu'il semble en être de même quant à ce que les médecins n'attendent pas, ou redoutent de l'intelligence artificielle.

Nous soulignons deux éléments mettant en perspective ces attentes.

Premièrement l'évolution très rapide du secteur les rend changeantes. Un médecin qui découvre l'IA générative peut avoir bien d'autres attentes pour sa pratique que celui qui n'en a jamais entendu parler. Dans la mesure où il y a peu de solutions disponibles, il paraît difficile de répondre à cette question.

² Le fait de donner de l'autonomie au patient.

Les attentes d'aujourd'hui ne seront peut-être pas celles de demain, et la plupart des travaux que nous avons mentionnés demandent aux médecins interrogés « d'imaginer » ce qu'il pourrait faire, avec des technologies qu'ils doivent aussi imaginer.

Deuxièmement, les travaux de Rahimi (42) montrent que les IA en soins primaires étaient finalement plus utilisées pour le diagnostic ou le pronostic que pour d'autres tâches souvent mentionnées parmi les attentes des médecins. Il existe donc probablement un écart entre l'utilisation attendue ou souhaitée, et l'utilisation réelle. L'utilisation réelle dépend de nombreux facteurs, et les technologies disponibles influencent probablement fortement l'utilisation qui en est faite.

Nous pouvons dans tous les cas constater un certain écart entre des solutions d'IA offrant la possibilité d'automatiser presque toutes les tâches du médecin, et les attentes de ce dernier. Par ailleurs, l'implication attendue de l'humain dans l'utilisation, la conception et la validation de l'IA, appelée de leurs vœux par nombre d'institutions et d'auteurs, invite à une approche *bottom-up* (issues des besoins du terrain), ce qui ne semble pas toujours être le cas. Ce genre d'approche demandera une plus grande implication des médecins eux-mêmes, et une plus grande collaboration avec les informaticiens.

IV. Rationnel du travail de thèse

Nous l'avons vu, les médecins y compris les généralistes, doivent s'impliquer dans la recherche sur l'IA s'ils ne veulent pas être dépassés, ou que s'imposent à eux des technologies dont ils n'auraient pas besoin. Les progrès de l'IA ne doivent pas faire oublier le rôle fondamental de l'humain dans une médecine au service du patient.

Dans la grande diversité des offres existantes, les IA génératives fondées sur le deep learning, et en particulier les LLM font l'actualité et semblent les plus à même d'induire des bouleversements profonds dans nos sociétés. Comme expliqué précédemment, c'est sur ce type de technologie que nous nous concentrons.

Nous l'avons vu, les données manquent sur la question de l'utilisation de ces IA par les médecins généralistes.

Pour tenter d'apporter des éléments de réflexion à ces problématiques, nous avons mené un double travail. D'une part la réalisation d'une enquête auprès de médecins généralistes sur leurs utilisations des IA génératives, d'autres part un travail concret d'évaluation des LLMs. Nous espérons que ces deux aspects de notre travail apporteront des premiers éléments de réponse.

A. Enquête auprès des médecins généralistes

Le projet ANR MALADES (55) est un projet de recherche, porté (*entre autres*) par le CHU de Nantes, et explorant l'intégration et l'adaptation des Grands Modèles de Langue pour des applications cliniques spécifiques.

Nous avons eu l'occasion de nous greffer à ce projet, avec l'objectif d'évaluer les utilisations actuelles de l'IA par les soignants ainsi que leurs attentes. Nous avons initialement travaillé à l'élaboration d'un questionnaire auprès des professionnels du CHU de Nantes. Dans le prolongement de cette enquête, nous nous sommes proposés d'interroger les médecins généralistes libéraux de la région, avec le même objectif de faire émerger des données « du terrain » qui à terme pourront guider le développement de certains outils.

Cela nous paraissant d'autant plus important pour une génération de soignants qui aura, comme les générations précédentes, à intégrer dans sa pratique de nouveaux outils qui seront numériques et intégreront des algorithmes d'intelligence artificielle de plus en plus pointus.

B. Évaluation de grands modèles de langues sur le raisonnement diagnostique

La seconde partie de notre travail s'est attachée à « faire de l'IA en vrai ». Nous avons décidé d'évaluer les performances de différents modèles de langues sur des tâches concrètes. Notre questionnement s'est porté sur l'une des tâches emblématiques à laquelle le médecin est confronté, celle du diagnostic clinique.

La question du diagnostic a déjà largement été traitée. Lorsque l'on saisit, au moins dans les grandes lignes, le fonctionnement des LLM, il ne paraît pas très étonnant qu'ils soient à même de produire un diagnostic exact sur la base de données d'entrée suffisamment précises et complètes.

Si la capacité des LLM à produire un diagnostic exact n'a rien de très étonnant, le médecin a sans doute plus besoin d'éléments pour raisonner que d'une réponse finale apparaissant *ex nihilo*. Il semble donc pertinent de s'intéresser à la capacité des modèles à expliciter leurs réponses, c'est à dire, pour ce qui est du diagnostic, **la capacité à produire un raisonnement pertinent**. Chercher à obtenir de l'IA ce genre de raisonnement présente au moins deux avantages. Premièrement, on évite l'effet « boîte noire », souvent cité comme une limite à l'utilisation de l'IA (25,45). Deuxièmement, cela permet d'utiliser l'intelligence artificielle non pas comme un substitut au soignant, mais bien comme un outil d'aide au raisonnement (et donc au diagnostic).

Notre travail s'est porté sur la capacité de l'IA à raisonner sur la base d'un prompt unique en langage naturel, auquel un médecin pourrait être confronté. Nous n'avons pas étudié spécifiquement la technique de prompting et son influence sur les réponses de l'IA (56,57).

Enfin, plusieurs constats détaillés plus haut nous poussent à nous interroger sur certains points spécifiques. Y a-t-il des différences significatives entre les modèles ? Le choix du langage d'entrée influence-t-il les performances ? Les modèles sont-ils aussi bons en médecine générale que pour les autres disciplines ? Ce sont autant de questions auxquelles nous avons tenté d'apporter des éléments de réponse avec ce travail.

Matériel et méthode

I. Enquête auprès des médecins généralistes

A. Objectifs de l'étude

Le premier objectif de l'étude était d'évaluer l'utilisation de l'intelligence artificielle générative par les médecins généralistes.

L'objectif secondaire était d'évaluer le degré de favorabilité des médecins envers ces technologies ainsi que les freins à leur adoption.

Le questionnaire proposé devait s'aligner sur celui qui était initialement prévu pour l'hôpital. Néanmoins pour des raisons techniques et administratives ce dernier n'a pas pu être lancé dans les temps et nous ne ferons pas d'analyse comparative.

B. Type d'étude

Il s'agissait d'une étude quantitative descriptive réalisée à l'aide d'un questionnaire contenant 19 questions dont une ouverte. Elle a été réalisée du 14 juin au 30 août 2024.

C. Population visée

La population d'étude comprenait l'ensemble des médecins généralistes pratiquant une activité libérale en Pays de la Loire.

D. Élaboration du questionnaire

1. Composition du questionnaire

Le questionnaire a été élaboré sur FramoForms.

Nous n'avons pas donné de définition précise de l'intelligence artificielle, tant le sujet est vaste comme nous l'avons précédemment expliqué. Nous avons néanmoins rappelé à quoi correspondaient les intelligences artificielles génératives, en citant l'exemple de ChatGPT, modèle le plus populaire à ce jour depuis sa sortie fin novembre 2022.

Le questionnaire était divisé en deux grandes sections. La première section est composée de plusieurs parties :

- La première partie évalue le contact des professionnels via l'IA et leur perception vis-à-vis de l'impact que celle-ci peut avoir dans leur milieu professionnel. Ces premières questions présentent aussi l'avantage de bien dissocier intelligence artificielle « classique » et générative pour la suite du questionnaire.

Il est à noter que cette partie (les quatre premières questions) était censée être partagée avec le questionnaire adressé aux médecins hospitaliers. Afin de maintenir la comparabilité, nous ne l'avons pas modifié.

- La seconde partie s'intéresse spécifiquement aux intelligences artificielles génératives. Elle interroge sur leur utilisation à la fois dans leur activité professionnelle et en dehors, sur les secteurs d'activités où les médecins pensent qu'elles auront le plus d'impact, sur les principaux enjeux dans leur utilisation, et enfin sur leur degré de favorabilité vis-à-vis de leur installation pérenne.
- Nous avons ajouté deux questions sur l'intérêt et le souhait des médecins d'en apprendre plus sur ces technologies dans le cadre de l'étude PENSO (Projet d'Enseignement du Numérique en Santé de l'Ouest, l'un des programmes du cluster DELPHI).

La seconde partie du questionnaire apporte des informations sur le profil de médecin interrogé, à savoir la tranche d'âge, le sexe, le mode d'exercice ainsi que le degré de compétence informatique. Concernant cette question, nous n'avons pas défini de degrés de compétence informatique et avons laissé les participants faire leur propre auto-évaluation.

Enfin, le questionnaire se terminait par une question libre, où les participants étaient invités à préciser de quelle façon ils utilisaient les intelligences artificielles génératives, ou dans quelles activités ils auraient voulu les voir se développer.

Le questionnaire est disponible en annexe.

2. Diffusion du questionnaire

Le questionnaire était disponible en ligne sur le site FramaForms du **14 juin au 30 août 2024**.

Le questionnaire a été diffusé principalement par courriel. Il était destiné à l'ensemble des médecins généralistes de la région Pays de la Loire. Les liens étaient accessibles sur la rubrique dédiée aux thèses par le CDOM 44. L'URML Pays de la Loire a été sollicitée afin de diffuser le questionnaire au sein d'une liste de 3000 médecins.

Les auteurs ont fait directement appel par téléphone aux cabinets et maisons médicales des Pays de la Loire pour répondre aux questionnaires, en recourant aux pages jaunes ou à Google. Ils ont aussi sollicité leurs propres contacts personnels et une page Facebook.

Les réponses ont été recueillies sur un tableau Excel et les analyses ont été réalisées en utilisant Python. Les graphiques ont été réalisés en utilisant la bibliothèque matplotlib.

II. Évaluation de modèles de langue sur le diagnostic clinique

A. Conception d'une banque de cas clinique

Dans le cadre de notre deuxième travail, et pour tester la capacité des LLM à raisonner sur une question diagnostique, nous avons fait le choix d'utiliser le format classique du cas clinique. Les cas sont des vignettes de texte décrivant la situation clinique d'un patient à la manière d'une observation médicale, reprenant de façon plus ou moins détaillés l'histoire de la maladie, les éléments d'interrogatoire et les données de l'examen physique, parfois des éléments paracliniques. Le cas clinique est souvent utilisé comme un exercice, où l'objectif est d'arriver à trouver le diagnostic et/ou une conduite à tenir. Les vignettes cliniques sont un mode de formation et de communication largement utilisé en médecine. Plusieurs bases de données de cas existent et ont déjà été utilisées par nombre de travaux sur l'intelligence artificielle (58). Cependant, leur accès n'est pas toujours facile et la plupart d'entre elles sont en anglais. Nous avons fait le choix de constituer notre propre base de cas clinique. Celle-ci présente l'intérêt d'être originale, d'avoir été conçue en français et de contenir des cas spécifiques de médecine générale.

1. Nombre de cas

Le souhait des auteurs était qu'un maximum de spécialités cliniques soient représentées. Pour rappel, ci-après la liste des spécialités médicales accessibles à l'ECN en 2024 (total : 44) (59):

Allergologie, Anatomie et cytologie pathologiques, Anesthésie Réanimation, Biologie médicale, Chirurgie maxillo-faciale, Chirurgie Orale, Chirurgie orthopédique et traumatologique, Chirurgie pédiatrique, Chirurgie plastique, reconstructrice et esthétique, Chirurgie thoracique et cardiovasculaire, Chirurgie vasculaire, Chirurgie viscérale et digestive, Dermatologie et vénéréologie, Endocrinologie-diabétologie-nutrition, Génétique médicale, Gériatrie, gynécologie médicale, Gynécologie obstétrique, Hématologie, Hépatogastro-entérologie, Maladies infectieuses et tropicales, Médecine cardio-vasculaire, Médecine d'urgence, Médecine et santé au travail, Médecine générale, Médecine intensive et réanimation, Médecine interne et immunologie clinique, Médecine légale et expertises médicales, Médecine nucléaire, Médecine physique et réadaptation, Médecine vasculaire, Néphrologie, Neurochirurgie, Neurologie, Oncologie, Ophtalmologie, Oto-rhino-laryngologie chirurgie cervico-faciale, Pédiatrie, Pneumologie, Psychiatrie, Radiologie et imagerie médicale, Rhumatologie, Santé publique et médecine sociale, Urologie

Nos cas devaient avant tout porter sur le diagnostic clinique, c'est pourquoi nous avons sélectionné uniquement les spécialités médicales et médico-chirurgicales. Nous avons exclu les disciplines techniques (radiologie, biologie médicale, anatomopathologie, médecine nucléaire) ainsi que les disciplines purement chirurgicales.

Il ne nous a pas semblé pertinent d'inclure des cas autour des disciplines suivantes : santé publique, médecine du travail, médecine légale, génétique médicale, médecine physique et réadaptation, allergologie, gériatrie. La question diagnostique y est souvent moins centrale, et certains libellés utilisés dans d'autres disciplines pouvaient déjà les couvrir. La dermatologie a également été exclue en raison de l'importance de l'image dans cette spécialité. Enfin, nous avons regroupé certaines disciplines sous un seul libellé :

- Médecine d'urgence, médecine intensive et réanimation, et anesthésie réanimation ont été regroupés dans « Urgences-Réanimation ».
- Gynécologie obstétricale et médicale, regroupées dans « Gynécologie »
- Oncologie et hématologie, regroupées dans « Onco-Hématologie »
- Médecine vasculaire, immunologie clinique et médecine interne dans « Médecine interne »
- Ophtalmologie et ORL dans « Tête et cou »
- Urologie et néphrologie dans « Urologie-Néphrologie »

Les libellés finalement retenus ont été les suivants (*total : 16*) :

Urgence-Réanimation, Endocrinologie-diabétologie-nutrition, Gynécologie, Onco-Hématologie, Hépatogastro-entérologie, Maladies infectieuses et tropicales, Médecine cardio-vasculaire, Médecine générale, Médecine interne, Neurologie, Tête et cou, Pédiatrie, Pneumologie, Psychiatrie, Rhumatologie, Urologie-Néphrologie.

Ces choix permettent d'affirmer que notre base de cas cliniques couvre l'ensemble du champ de la pratique médicale clinique.

Une fois les libellés définis, nous avons décidé de réaliser environ dix cas cliniques par libellé. La médecine générale étant notre discipline, nous avons proposé d'obtenir environ 30 cas pour cette dernière. La réalisation des cas a été répartie équitablement entre les deux auteurs, avec **180 cas au total**. Le tableau 1 résume la répartition finale du nombre de cas par libellé.

Spécialité	Nombre de cas
Urgence-Réanimation	10
Endocrinologie-diabétologie-nutrition	9
Gynécologie	9
Onco-Hématologie	9
Hépatogastro-entérologie	15
Maladies infectieuses et tropicales	17
Médecine cardio-vasculaire	12
Médecine générale	32
Médecine interne	6
Neurologie	11
Tête et cou	7
Pédiatrie	7
Pneumologie	10
Psychiatrie	8
Rhumatologie	9
Urologie-Néphrologie	9
Total	180

Tableau 1 - Répartition finale du nombre de cas par libellé

2. Règles de conception a priori et sources

Dans l'optique de tester les LLM, chaque cas constituait une vignette unique et avait un diagnostic prédéfini. Les vignettes devaient inclure l'ensemble des informations nécessaires à leur résolution, c'est-à-dire permettre d'aboutir au diagnostic. En raison de la nature de certains diagnostics, certains cas comportaient des éléments paracliniques, le plus souvent simple ou « de première intention ».

Dans notre démarche, les diagnostics finals doivent être entendus comme diagnostics « attendus » indépendamment du niveau de certitude. Il s'agit, selon les situations, des diagnostics les plus pertinents, les plus probables ou bien les plus utiles en pratique. Par exemple, pour un patient dans un état critique, c'est le diagnostic de choc septique qu'il est important de porter en priorité, pas forcément celui du germe en question. On peut parler d'un certain degré de « profondeur » du diagnostic, qui sera ici variable d'un cas à l'autre. Les diagnostics proposés seront étiologiques, syndromiques, ou encore paracliniques, en fonction des situations et des données disponibles. Pour les cas inspirés de référentiels, le diagnostic final était celui indiqué dans le référentiel.

Les sources utilisées pour élaborer les cas ont été variées. On peut distinguer trois grands groupes de cas selon leurs sources : les cas inventés, les cas inspirés de textes de référence et les cas inspirés de situations réelles rencontrées en pratique.

Les cas inventés ont été conçus sur la base des recommandations en vigueur, lesquelles sont citées lorsque cela est pertinent. Les cas issus de référentiels sont ceux inspirés de la littérature scientifique, notamment de référentiels de sociétés savantes, d'ouvrages de préparation à l'ECNi³, ainsi que de cours et conférences d'internat consultés par les auteurs au cours de leur formation (les sources utilisées sont également citées dans la base). Chaque cas a été adapté pour répondre aux critères mentionnés précédemment, notamment en étant synthétisé en un texte unique (vignette). Concernant les cas réels, les données initiales ont été adaptées afin de garantir l'anonymisation parfaite du cas.

Nous appuyer sur des sources de références a permis de s'assurer d'une « validité » experte du cas et du diagnostic final. La nomenclature proposée pour le diagnostic était celle des référentiels utilisés comme source. Pour les cas réels, la nomenclature utilisée relevait du choix de l'auteur, dans la mesure du possible actualisée ou a minima, largement utilisée en pratique courante.

Pour chaque cas, les auteurs proposaient un raisonnement diagnostique permettant de le résoudre. Ce raisonnement servait de support pour la relecture des réponses de l'IA. Il permettait également certaines analyses annexes. Deux points clés concernant ces raisonnements :

1. Il s'agit de propositions des auteurs. Un même cas clinique, soumis à différents médecins, pourra produire différents raisonnements, tous valides et aboutissant au même diagnostic. Le raisonnement proposé par les auteurs à l'avantage d'être « humain » mais ne saurait constituer l'unique réponse possible.
2. Il s'agit d'un cadre théorique. La réalité du raisonnement est souvent plus complexe et le clinicien pourra utiliser plusieurs types de raisonnements de manière concomitante.

La validité des raisonnements proposés pouvait être considérée comme satisfaisante. Pour les cas issus de référentiels ou créés à partir de ceux-ci, les raisonnements proposés s'appuyaient sur ces mêmes référentiels. Pour les cas réels, c'est le raisonnement de l'auteur lui-même qui était proposé.

³ De 2016 à 2021, les Épreuves Classantes Nationales informatisées (ECNi) constituaient l'examen final du second cycle des études de médecine. Parmi les épreuves, figurait une série de cas cliniques. Les étudiants se préparaient à ce concours en travaillant sur des séries de cas cliniques corrigées en groupe, sous la supervision d'un enseignant, lors de sessions appelées « conférences d'internat ».

3. *Forme*

Pour identifier les cas, nous nous référons à un code simple, composé d'une lettre et d'un nombre. La lettre, S ou J, fait référence au prénom de l'auteur du cas (*Sofiane ou Josselin*), le nombre, à la position du cas dans l'ordre de création. Chaque cas a donc un identifiant unique, allant de S1 à S90 et de J1 à J90.

Chaque cas présente quatre éléments : la vignette clinique, c'est à dire le cas à proprement parlé, le diagnostic attendu, le raisonnement diagnostique et une traduction du cas en anglais proposée par l'auteur.

Notons que si la forme finale des cas est homogène du fait des règles de conception sus mentionnées, le ton et le style sont variables d'un cas à l'autre. Cela est dû au fait qu'ils sont proches des cas dont ils sont inspirés, eux-mêmes issus de diverses sources. Pour les cas inventés et inspirés de cas réels, on peut retrouver une certaine diversité de style entre les deux auteurs.

4. *Description de la base finale de cas cliniques*

Les données qui suivent ont été obtenues après la réalisation des cas. Les différences de proportions vis-à-vis de certaines caractéristiques ont donc été « découvertes » après coup, un équilibre parfait entre ces caractéristiques n'ayant pas été une contrainte posée a priori.

Chaque cas se référait à un seul patient.

a) *Taille des cas :*

Les cas de la base de données comportent en moyenne **144 mots**. La médiane est à **128 mots**. Le plus petit cas était le J15 avec **40 mots**, le plus long est le J25 avec **389 mots**.

b) *Profil des patients :*

L'âge moyen de ces patients était de **38 ans** pour une médiane de **35 ans**. Le cas avec le patient le plus jeune était le S12, évoquant un enfant de **3 mois**. Celui avec le patient le plus âgé était le J68, avec une patiente de **92 ans**.

Un seul cas ne mentionnait pas l'âge du patient (J16). Dans ce dernier, bien qu'il y ait peu d'indication, les données permettaient d'envisager qu'il s'agisse d'un homme d'âge mur. Ce manque n'a pas porté préjudice à notre travail. On observe un plus grand nombre de cas aux âges les plus bas (0 à 10 ans), autour des 30 ans, et un dernier pic autour de 50 ans. On note un nombre de cas moins important pour les âges les plus élevés. La figure 13 illustre le nombre de cas par tranches d'âge de 5 ans. Notre base de cas clinique englobe tous les âges.

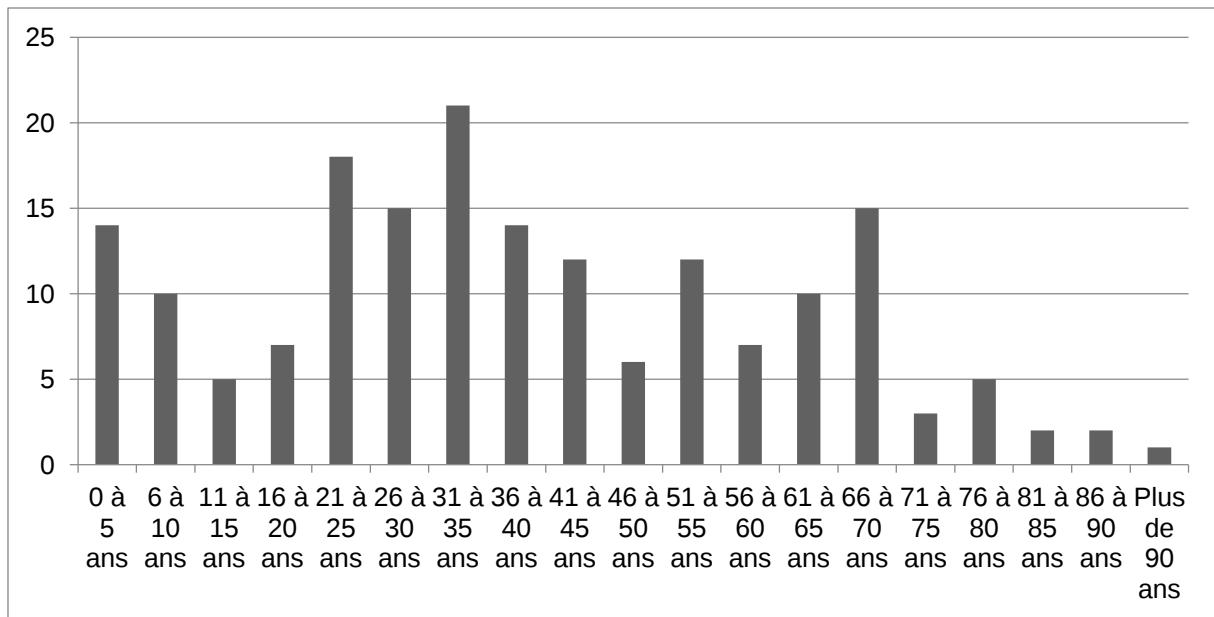


Figure 13 - Répartition du nombre de cas clinique par tranche d'âge.

c) Sexe des patients

En ce qui concerne le genre, la totalité des cas précisait si le patient était un homme ou une femme. Aucun des cas ne mentionnait un autre genre. On retrouve une proportion relativement équivalente de femmes et d'hommes dans la base de données, avec une légère prédominance masculine. La base comporte **83** cas concernant une femme (**environ 46%**) et **97** cas concernant un homme (**environ 54 %**).

d) Source des cas :

La figure 14 résume la répartition des cas selon le type de source.

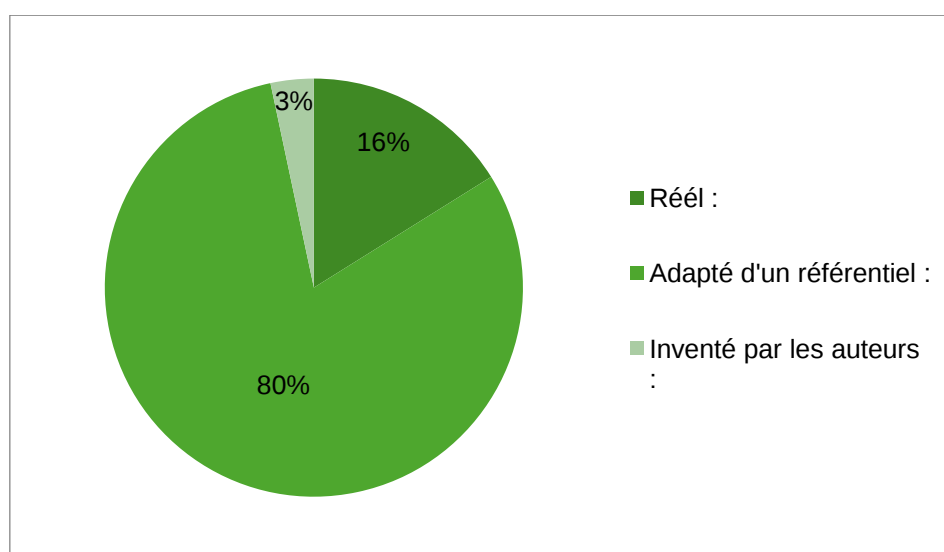


Figure 14 - Proportion de cas par type de source

e) Raisonnement clinique :

Chaque cas était associé à un raisonnement clinique proposé par son auteur. En ce qui concerne la nomenclature des types de raisonnements, nous avons choisis celle proposée par les auteurs du livre « 41 concepts clés pour la médecine général » (35). Ces derniers distinguent dans leur ouvrage 6 types de raisonnement diagnostique : le raisonnement en chainage avant, le raisonnement hypothético-déductif, le raisonnement exhaustif, le raisonnement par reconnaissance de cas (*ou raisonnement non analytique*), le raisonnement algorithmique et le raisonnement probabiliste (*pour une définition de chaque type de raisonnement, voir annexe 2*). La proportion de cas pour chaque type de raisonnement est présentée figure 15.

Aucun raisonnement proposé ne relevait d'un raisonnement exhaustif. Quand un raisonnement relevait de plusieurs « types » nous avons conservé celui jugé prédominant.

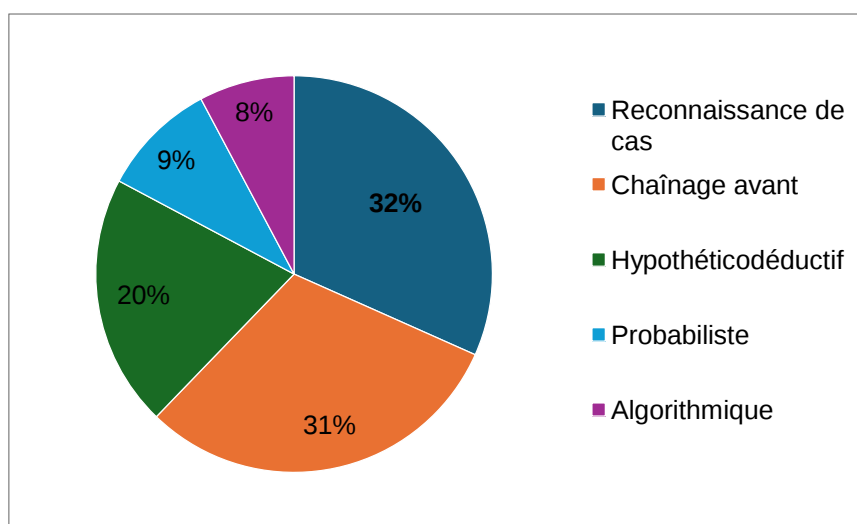


Figure 15 - Répartition des cas par type de raisonnement associé

Les cas ont été créés de sorte à comporter l'ensemble des données permettant de les résoudre. Cela explique certainement la majorité de raisonnements par reconnaissance de cas.

f) Déroulé des cas cliniques :

Certaines caractéristiques de nos cas cliniques nous ont semblé mériter une description détaillée. Cette classification nous a aussi permis de réaliser certaines analyses exploratoires.

Premièrement, nous avons précisé si le déroulé du cas correspondait à une situation de soin diachronique ou synchronique (*consultation unique ou rencontres de soins répétées*). Sans surprise, les situations synchroniques prédominent, représentant **76% du total, soit 137 cas**.

Il existe néanmoins une proportion non négligeable de situation diachroniques, **24% du total soit 43 cas**.

Nous avons également examiné la présence d'éléments paracliniques. La base de données comporte ainsi **71% de cas** où aucun examen paraclinique n'est mentionné (**128 cas**), tandis que 29% de cas en comportent (**52 cas**).

g) Type de diagnostic proposé :

Nous avons choisi de classer les diagnostics en seulement trois catégories, malgré les limites que cela peut impliquer. Ces catégories sont : étiologique, syndromique et paraclinique. La figure 16 montre la répartition des types de diagnostics attendus dans la base de cas.

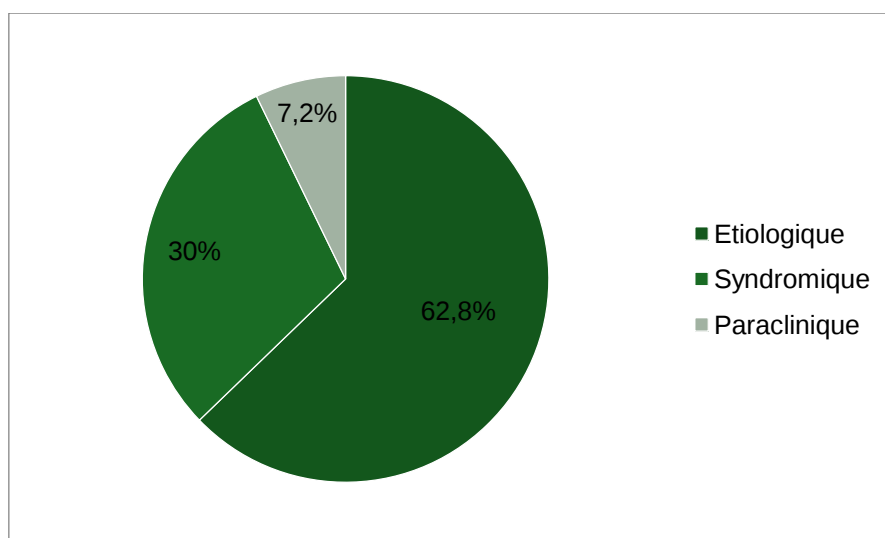


Figure 16 - Répartition des cas selon le type de diagnostic attendu

Les quelques cas correspondant à une condition non pathologique ont été classés dans la catégorie des diagnostics « étiologiques ».

h) Spécialités concernées :

La base de cas a été conçue en produisant un nombre équivalent de cas pour chaque spécialité. Néanmoins, nombre de pathologies et de situations cliniques peuvent relever de plusieurs spécialités différentes. Par exemple, une pyélonéphrite aiguë relève à la fois de l'urologie, de l'infectiologie et des urgences. Une tentative de suicide chez un jeune adolescent relève à la fois de l'urgence, de la pédiatrie et de la psychiatrie. Nous avons donc précisé pour chaque cas la ou les « co-discipline(s) » dont il pouvait relever.

Pour environ **29 %** des cas (**soit 52 cas**), le diagnostic ne relevait que d'une seule discipline (celle initialement attribuée). Pour les autres, le tableau 2 synthétise par spécialité le nombre de cas concernés. Les trois disciplines apparaissant le plus souvent comme « co-discipline »

étaient des spécialités très transversales. Il s'agissait de la médecine d'urgence (35 cas), de la pédiatrie (14 cas) et de l'infectiologie (11 cas).

Co-discipline	Nombre de cas concernés	Co-discipline	Nombre de cas concernés
Aucun	52	Médecine générale	9
Allergie	1	Néphrologie	3
Angiologie	1	Neurochirurgie	1
Cardiologie	4	Oncologie	7
Chirurgie viscérale	1	ORL	4
Dermatologie	7	Orthopédie	9
Endocrinologie	2	Pédiatrie	14
Gériatrie	2	Pneumologie	1
Gynécologie	2	Psychiatrie	2
Hématologie	2	Rhumatologie	4
HGE	2	Urgence	35
Infectiologie	11	Urologie	4
Médecine interne	7		

Tableau 2 - Nombre de cas concernés par chaque "co-discipline"

La totalité de ces données, cas par cas est proposé dans un tableau synthétique en annexe.

B. Choix de la métrique de test

Pour chaque cas clinique, il était demandé au modèle de fournir un diagnostic final et d'explicitier le raisonnement permettant d'y aboutir. Nos recherches documentaires n'ont pas permis d'identifier une méthode d'évaluation satisfaisante des réponses fournies par l'intelligence artificielle. Dans la littérature, les modes d'évaluation sont variées et ne sont pas toujours mentionnés. Il peut s'agir d'avis d'experts, dont les méthodes précises ne sont pas détaillées, de questions fermées ou à choix multiples dont l'analyse est faite de façon automatique. Par conséquent, nous avons fait le choix de proposer notre propre grille d'évaluation. L'objectif était d'évaluer avant tout la qualité du raisonnement proposé. Cette grille se voulait simple, reproductible et objective, avec une note finale sur dix points. Bien qu'il s'agisse de noter le raisonnement clinique, un des items de notation concernait le diagnostic final proposé par l'IA.

Bien que nous ayons initialement envisagé de faire deux notes séparées, il nous a semblé plus pertinent de ne faire qu'une seule note comportant l'évaluation du diagnostic et celle du raisonnement. Les autres items de la grille sont détaillés ci-après.

Plusieurs versions successives de la grille ont été proposées, puis modifiées après avoir été testées sur des premiers échantillons de cas. Les auteurs ont partagé leurs résultats à chaque étape afin d'affiner les règles de notation. La grille finale a été testée également sur un plus grand échantillon de réponses (20 cas), notées en aveugle. La comparabilité des résultats entre les deux auteurs a été jugée suffisante pour valider la grille.

Ces phases de test ont été réalisées sur les premières réponses obtenues, à savoir celles générées par ChatGPT-4. Les exemples mentionnés ci-dessous sont issus de réponses de ce modèle.

1. Grille de cotation finale :

Validité interne :

Dans cet item, nous évaluons la capacité de l'IA à extraire et interpréter les données pertinentes. L'IA doit être capable de mentionner tous les éléments du cas utiles au diagnostic. Par ailleurs, un certain nombre d'éléments, sans constituer le diagnostic à proprement parlé, peuvent être interprétés pour contribuer au raisonnement.

Nous évaluons aussi la capacité à retranscrire certaines descriptions sémiologiques en termes médicaux. Une « lésion cutanée violacée ne s'effaçant pas à la vitro pression » doit faire évoquer un purpura, ce qui représente un premier niveau de lecture. Il en est de même pour les constantes et données cliniques chiffrées, par exemple parler de fièvre en cas de température à 38.5°C, d'hypertension au-delà d'un certain seuil de pression artérielle systolique, ou de surpoids lorsque les données de poids et de taille permettent d'obtenir un $IMC \geq 25 \text{ kg/m}^2$. Enfin, l'interprétation s'étend aussi aux données paracliniques. Lorsque ces données ne sont pas déjà interprétées dans le cas, il paraît justifié d'attendre de l'IA qu'elle puisse le faire lorsque cela est utile au diagnostic.

- + 3 pts : tous les éléments du dossier utiles sont mentionnés et sont interprétés de façon correcte.
- + 2 pts : tous les éléments du dossier ne sont pas analysés de façon suffisante ou avec une interprétation médicale insuffisante
- + 1 pt : les éléments du dossier sont insuffisants pour parvenir au diagnostic
- 0 pt : la lecture du dossier est fausse, l'interprétation des éléments cliniques ou paracliniques est fausse.

Exemple :

Cas S13 : l'IA n'évoque pas certains facteurs de risque présents dans le cas clinique, pourtant utiles pour renforcer le diagnostic, + 2 points. Cas S14 : sur des céphalées intenses, l'IA ne mentionne pas le signe négatif pourtant très important qui est l'apyrexie, +1 pt. Ces éléments sont sanctionnés sur la validité interne.

Validité externe :

Lorsque l'IA fait appel dans son raisonnement à des éléments de connaissance médicale, non mentionnés dans le cas d'entrée, il s'agit ici d'évaluer la validité de ces connaissances sur le plan purement médico-scientifique. A noter que si l'IA ne fait appel à aucun élément « externe », une note de 2 points sera attribuée par défaut, afin de ne pas pénaliser l'absence de tels arguments qui ne sont pas toujours nécessaires.

+ 2 pts : connaissances externes mentionnées valides et actualisées, ou pas de connaissance externe mentionnée
+ 1 pt : connaissances médicales partiellement correctes ou obsolètes, discutables
0 pt : connaissances médicales fausses, ou nombreuses données discutables

Exemple :

Cas J12 : « la prise en charge devra inclure des prélèvements microbiologiques de la cicatrice ». Dans ce cas d'infection de prothèse, les recommandations s'accordent à dire qu'il ne faut pas faire de prélèvement local de la cicatrice. Ici, l'assertion de l'IA est fausse, la note est de 0.

Hypothèses / Diagnostics différentiels :

Nous évaluons dans cet item la capacité à formuler des hypothèses et à évoquer des diagnostics différentiels pertinents. Ceux-ci seraient idéalement hiérarchisés (en termes de probabilité et ou de gravité). Même si le diagnostic différentiel peut être facilement éliminé, il est souvent utile qu'il soit mentionné dans un raisonnement clinique. Si des diagnostics différentiels sont mentionnés dans le raisonnement des auteurs mais que le modèle ne l'évoque pas, la note sera de 0. En revanche, si le cas ne se prête à aucun diagnostic différentiel (cas « évident ») et que l'IA n'en mentionne pas, le point sera attribué par défaut afin de ne pas pénaliser ce raisonnement qui resterait donc valide.

+ 1 pt : le modèle évoque des diagnostics différentiels pertinents (ou n'en évoque pas lorsque cela n'est pas contributif)
0 pt : pas de diagnostics différentiels pertinents évoqués lorsque cela est nécessaire ou propositions non pertinentes ou fausses

Exemple :

Cas S19 : le raisonnement de référence proposait bien l'hypothèse à ne pas négliger de métastases osseuses. Le raisonnement proposé par l'IA ne le mentionne pas, la note est donc 0. Cas J12 : le cas est suffisamment explicite pour faire de la reconnaissance de cas, la note est donc 1.

Logique :

Cet item évalue de façon globale le déroulement du raisonnement, indépendamment de la validité des arguments avancés. Nous serons notamment attentifs à l'ordre logique des éléments mentionnés (les éléments cliniques devant généralement précéder les éléments paracliniques, par exemple), à l'absence de contradictions, ainsi que l'absence d'assertions parasites n'apportant rien au raisonnement.

+ 2 pts : tous les éléments évoqués par le modèle sont connectés de façon logique et organisés de manière à aboutir au diagnostic final
+ 1 pt : présence d'incohérences dans le déroulé du raisonnement
0 pt : les éléments du dossier sont évoqués de façon incohérente sans lien avec le diagnostic final, présence de contradictions ou ne sont pas proposés dans un ordre logique

Exemple :

Cas J16 : « Bien que le patient ait un antécédent familial de maladie de Parkinson, ce qui pourrait indiquer une prédisposition génétique, les symptômes cliniques et l'examen neurologique sont suffisamment évocateurs pour poser le diagnostic. » Cette assertion n'apporte aucun élément, voir semble sous-entendre que le diagnostic serait moins certain. De plus l'IA décrit les éléments du syndrome Parkinsonien (akinésie/hypokinésie, hypertonie extra pyramidale et tremblement de repos) et fait le raccourci de parler d'emblée de Maladie de Parkinson, qui est plutôt une des causes du syndrome Parkinsonien. La logique est ici sanctionnée.

Expression :

Nous sanctionnons dans cet item la présence de phrases présentant des erreurs en termes d'expression, de sens et/ou de syntaxe.

Il ne s'agit pas de sanctionner les fautes d'orthographe. Les réponses fournies en anglais alors que le français est attendu étaient sanctionnées.

0 pt : le texte est compréhensible, sans erreurs de syntaxe, ni phrases dont le sens serait douteux

-1 pt : il existe des phrases dont la syntaxe ou la tournure est de qualité insuffisante, voir gêne la bonne compréhension du texte, pouvant prêter à confusion. Certaines expressions sont mal formulées. En cas de mots inventés, de mélange anglais/français ou de phrases incompréhensibles l'expression sera aussi sanctionnée

Exemple :

Cas S3 : « infection par le paludisme » ne se dit pas. Le paludisme constitue l'infection et non l'agent infectieux. Il est courant de dire que l'on « attrape le paludisme » mais l'on n'est pas infecté par le paludisme. Dans ce cas -1 pt pour cette réponse. Cas J2 : « douleur [...] décrite comme rapide à l'apparition et non liée à l'effort », faute de syntaxe donc -1 pt aussi dans ce cas.

Précision diagnostique finale :

Dans cet item, l'objectif est d'évaluer la concordance entre le diagnostic final et celui proposé par les auteurs, considéré comme la référence. Si le diagnostic proposé par l'IA est incomplet ou insuffisant, ou que la nomenclature utilisée n'est pas jugée correcte (ou obsolète) il sera noté 1 point, s'il est tout à fait complet 2 points, s'il est faux 0.

De plus, si la réponse du modèle omet trop d'éléments jugés importants pour le diagnostic, la note peut également tomber à 0.

+2 pts : diagnostic exact

+1 pt : diagnostic incomplet ou nomenclature non appropriée

0 pt : diagnostic faux ou plusieurs éléments essentiels manquants

Exemples :

Cas J17, le modèle ne mentionne ni le sepsis, ni l'origine sexuellement transmissibles de l'infection. Deux informations clés. Même si le diagnostic d'IGH est correct, il manque ici deux éléments majeurs, la note est de 0.

Cas J14 : le modèle répond « érythème migrans » au lieu de « maladie de Lyme ». L'IA fait bien le diagnostic sémiologique de l'atteinte cutanée, mais pas le diagnostic étiologique attendu. L'érythème migrans étant pathognomonique de la maladie de Lyme, le diagnostic est jugé incomplet, la note est de 1.

Cas J11 : le modèle répond « adénite mésentérique » au lieu de « maladie des griffes du chat ». On peut considérer qu'il s'agit d'une hallucination. La réponse est fausse, la note est de 0.

Autres règles :

Lorsqu'un élément incorrect est sujet à caution quant à l'item concerné, c'est-à-dire lorsqu'il pourrait être pris en compte dans plusieurs items de notation, nous avons fait le choix de ne sanctionner l'erreur qu'une seule fois.

Exemple :

Cas J16 : l'IA parle de maladie de Parkinson sur un tableau de syndrome Parkinsonien. Elle ne mentionne aucun diagnostic différentiel, en particulier une origine médicamenteuse, qui pouvait être éliminée sur la base des données du cas clinique. On pourrait à la fois considérer que la validité interne est en cause, puisqu'il manque un élément important du cas clinique, mais aussi l'item « hypothèse » puisque l'IA n'en propose pas. Dans ce cas, on ne sanctionnera que l'un ou l'autre de ces deux items. Si une erreur peut être sanctionnée dans 3 items différents, il peut se discuter le fait de sanctionner deux items, mais pas les 3. Le but est de ne pas sous évaluer les modèles en sanctionnant trop fortement une seule erreur. Ces considérations donnent une certaine latitude aux auteurs.

Grille d'évaluation du raisonnement diagnostique		
Item évalué	Détail	Cotation
Validité interne <i>Retranscription et interprétation des informations utiles du cas</i>		0 pt +1 pt +2 pts +3 pts
Validité externe <i>Utilisation de connaissances médicales valides, actualisées</i>		0 pt +1 pt +2 pts
Hypothèse <i>Formulation de diagnostics différentiels pertinents lorsque cela est utile</i>	Pas de diagnostic différentiel Diagnostic différentiel	0 pt +1 pt
Logique <i>Construction d'un raisonnement logique en cohérent avec le diagnostic</i>		0 pt +1 pt +2 pts
Expression <i>Expression correcte et compréhensible</i>	Texte compréhensible, sans erreur de syntaxe Erreurs significatives dans la formulation et ou la syntaxe	0 pt - 1 pt
Précision diagnostique <i>Concordance entre le diagnostic proposé et le diagnostic attendu</i>	Diagnostic faux Diagnostic incomplet Diagnostic exact	0 pt +1 pt +2 pts
Score maximal théorique : 10 pts Score minimum théorique : - 1 pt		

Tableau 3 - Grille d'évaluation des raisonnements diagnostiques des LLM, version finale.

C. Choix des modèles testés et méthodologie des tests

Le choix des modèles de langue testés a été guidé par les hypothèses de travail. Ces hypothèses impliquaient de comparer des modèles ouverts (*open-source*) à des modèles fermés (privés), et des modèles généralistes à modèles ajustés (*fine-tuned*) sur le biomédical. Les contraintes temporelles et d'accès aux modèles nous ont fait retenir les trois modèles suivants :

- GPT-4 Turbo (modèle **privé, généraliste**) : fondé par OpenAI, il s'agit d'une version optimisée de GPT-4 conçue pour fournir des réponses à la fois rapides et précises.
- Llama3.1 405b (modèle **open-source, généraliste**) : fondé par Meta, c'est un modèle de langue très large avec un nombre de paramètres impressionnant.
- BioMistral (modèle **open-source, ajusté**) : modèle de langue open-source ajusté sur le domaine biomédical. Il possède beaucoup moins de paramètres que les deux cités précédemment (7 milliards).

Pour chaque cas, le modèle était interrogé par un *prompt* unique. Pour rappel, le prompt correspond à la requête envoyée au modèle. Il contient à la fois l'instruction et les données à évaluer à partir desquels l'IA va répondre. Le prompt que nous avons utilisé est illustré ci-dessous.

Pour les cas soumis en anglais, la réponse du modèle était faite en anglais. Pour les cas en français, la réponse était donnée en français. Les cas cliniques ont été soumis aux différents modèles via leurs APIs avec l'utilisation du langage Python.

Exemple de prompt en anglais :

Prompt: "Act as a doctor taking a medical history from a patient. Try your best to give your clinical reasoning in order to make an accurate final diagnosis. Format the output as JSON with 2 fields: 'diagnostic' with your final diagnosis and 'clinical reasoning' with your clinical reasoning that led to this diagnosis."

A 23-year-old man, weighing 60 kg, living in the Paris region, with no particular history, presented on his return from a stay in the southwest of France, as a summer camp monitor, with erythematous pharyngitis, without tonsillitis, with a temperature of 38°C, treated symptomatically in the city with paracetamol. Three days later, the temperature was 39°C. He was admitted to the hospital emergency room. He was confused, obtunded, and drowsy. The Glasgow score was 10/15, the heart rate was 100/min, the respiratory rate was 16 cycles per minute, and the blood pressure was 100/60mmHg. The clinical examination revealed purplish macules on the feet that did not disappear with pressure and meningeal stiffness.

Exemple de prompt en français :

Prompt: "Act as a doctor taking a medical history from a patient. Try your best to give your clinical reasoning in order to make an accurate final diagnosis in French. Format the output as JSON with 2 fields: 'diagnostic' with your final diagnosis and 'clinical reasoning' with your clinical reasoning that led to this diagnosis."

Un homme de 23 ans, 60 kg, vivant en Île-de-France, sans antécédents particuliers, a présenté au retour d'un séjour dans le sud-ouest de la France, comme moniteur de colonie de vacances, une pharyngite érythémateuse, sans amygdalite, avec 38°C de température, traitée en ville de manière symptomatique, par paracétamol. Trois jours plus tard, la température est à 39°C. Il est admis aux urgences de l'hôpital. Il est confus, obnubilé, somnolent. Le score de Glasgow est à 10/15, la fréquence cardiaque est à 100/min, la fréquence respiratoire est à 16 cycles par minute, la tension artérielle est à 100/60mmHg. L'examen clinique met en évidence au niveau des pieds des macules violacées qui ne s'effacent pas à la pression et une raideur méningée.

Les réponses des modèles ont été stockées dans un fichier Excel dans deux champs « diagnostic » et « clinical reasoning ». Une fois récupérées, elles ont été transmises à chacun des auteurs qui a noté chaque dossier à l'aide de la grille de notation décrite plus haut. La notation de chacun des auteurs s'est faite en aveugle de l'autre. La figure 17 résume cette méthodologie.

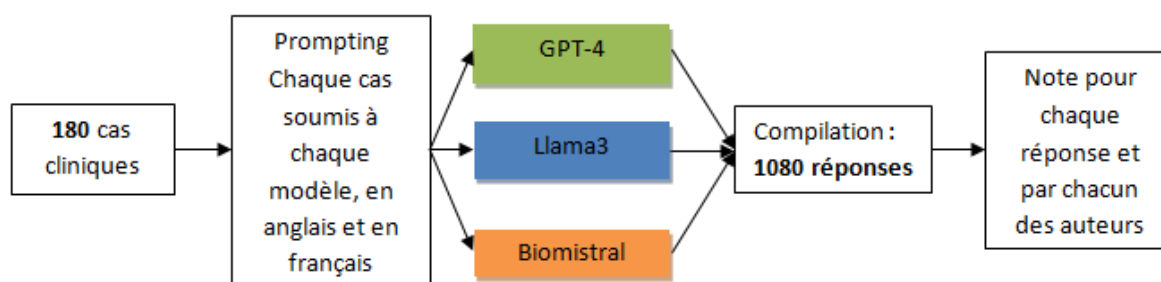


Figure 17 - Mise en place des tests et d'évaluation des modèles

En notant chaque cas clinique, les auteurs ont obtenu un total de 2160 notes sur 10.

Résultats

I. Enquête auprès des médecins généralistes

A. Données sociodémographiques

Nous proposons dans le texte suivant une analyse descriptive des résultats du questionnaire. Nous commençons par décrire l'échantillon des 87 participants que nous avons obtenus.

1. Sexe

L'échantillon comprend 50 femmes (57.5%) et 37 hommes (42.5%). 1 personne n'a pas répondu à la question sur le sexe (1.1%).

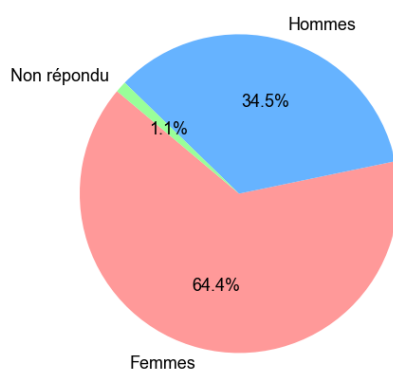


Figure 18 - Répartition des participants par sexe

2. Âge

La majorité des participants appartenaient à la tranche d'âge des 30 à 40 ans (49%).

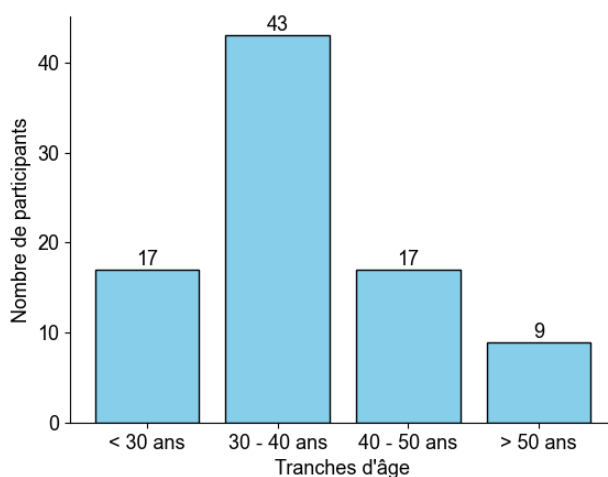


Figure 19 - Répartition des participants par tranche d'âge

3. Secteur d'activité

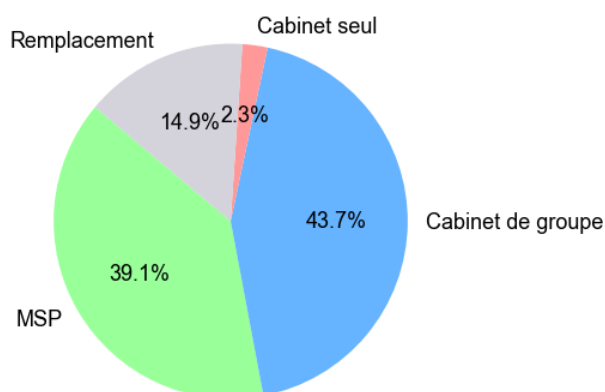


Figure 20 - Répartition des secteurs d'activité des participants

Nous constatons que la majorité des médecins étaient installés (15% de remplaçants), principalement dans des cabinets de groupe ou dans des MSP⁴ (82.8%).

4. Milieu d'exercice

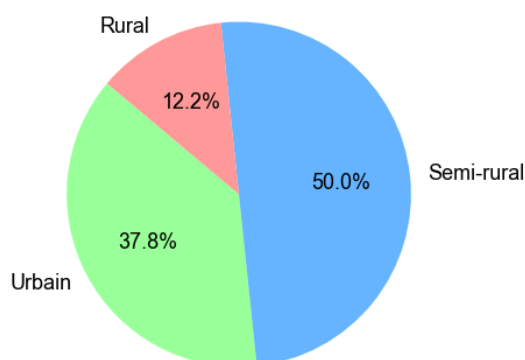


Figure 21 - Répartition des milieux d'exercice des participants

La majorité des réponses provenaient d'un milieu semi-rural (50%). Le milieu rural était sous-représenté dans l'échantillon (12.22%).

5. Département de réponse

La majorité des participants provenaient de Loire-Atlantique et Vendée (80%). L'autre département relativement représenté était le Maine-et-Loire avec 9 participants (10%). Les autres départements ayant participé étaient plus disparates.

⁴ MSP : Maison de santé pluridisciplinaire

B. Réponses au questionnaire

1. Compétence en informatique

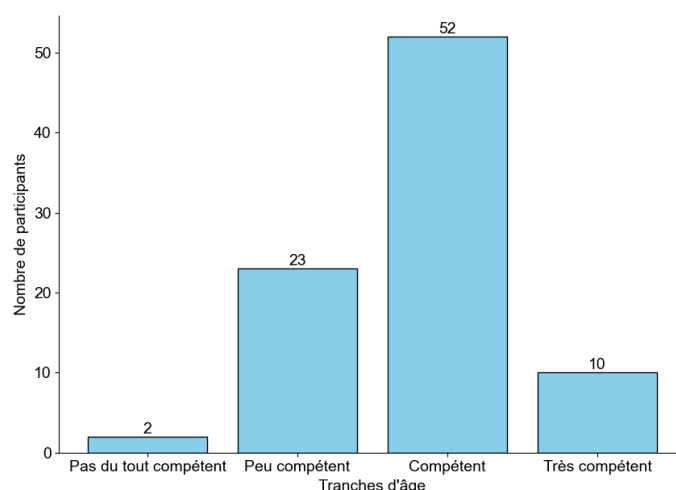


Figure 22 - Répartition des participants par compétence en informatique

La très grande majorité des participants se considéraient comme **peu compétent** à **compétent** (86.21%). Cela témoigne d'un certain niveau d'aisance de la part des médecins participants avec l'outil informatique. Néanmoins, cette perception reste subjective.

2. Contact avec l'intelligence artificielle

Dans notre échantillon, la très large majorité des participants avaient déjà été en contact, à leur connaissance, avec une intelligence artificielle (71.26%). Néanmoins, seulement 31 participants avaient déjà interagi avec une IA dans le cadre professionnel (35.63%). Nous préciserons cette tendance pour les intelligences artificielles génératives dans les réponses suivantes. Environ la moitié des participants avaient déjà interagi avec une intelligence artificielle générative (49.4%).

On peut constater que la plupart des médecins ont donc déjà été en contact avec ces technologies, à leur connaissance. Dans la proportion des médecins qui ont déjà été en contact avec une intelligence artificielle, environ 63% ont répondu qu'ils ont déjà été en contact avec une IA générative (n = 39, soit 62.90%).

3. Impact de l'intelligence artificielle générative

Sur une échelle de 1 à 10, les participants au questionnaire ont jugé que les intelligences artificielles ont un potentiel à transformer leur activité au quotidien, avec une moyenne de 6.15. Notez que nous ne sous-entendons aucune connotation positive ou négative dans la question. Le diagramme ci-dessous représente la répartition des notes.

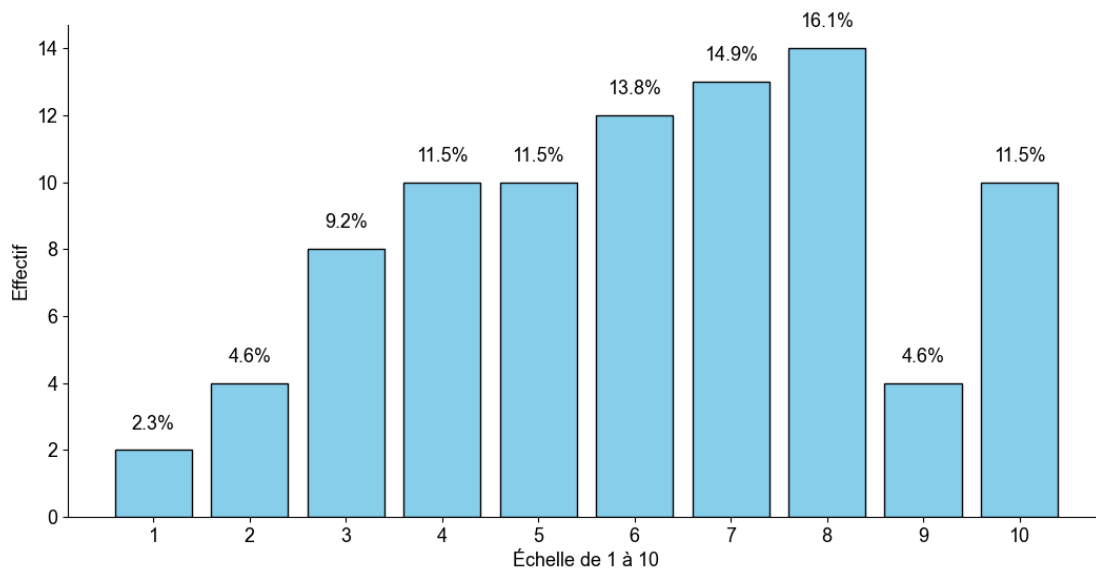


Figure 23 - Répartition des notes à la question "Sur une échelle de 1 à 10, quel serait selon vous le potentiel des IA génératives à transformer votre activité professionnelle au quotidien ?"

Les résultats montraient une distribution variée des notes attribuées par les participants. La note la plus fréquente est 8, avec 14 occurrences, représentant 16.10% des réponses. Les notes 6 et 7 étaient également bien représentées, avec respectivement 12 et 13 occurrences, soit 13.80% et 14.90% des réponses. Les notes extrêmes (1 et 10) sont moins fréquentes, avec seulement 2 et 10 occurrences respectivement, représentant 2.30% et 11.50% des réponses.

L'analyse de la répartition des notes montrait une perception variée du potentiel des IA génératives parmi les participants. Les notes les plus fréquentes (6, 7 et 8) suggéraient que les participants s'attendent à un **fort impact de l'intelligence artificielle dans leur pratique**.

4. Fréquence d'utilisation des intelligences artificielles génératives

a) Utilisation quotidienne des IA génératives

Dans le cadre de notre étude, nous avons analysé les réponses des participants à une question portant sur la fréquence d'utilisation des outils d'IA générative dans leur quotidien.

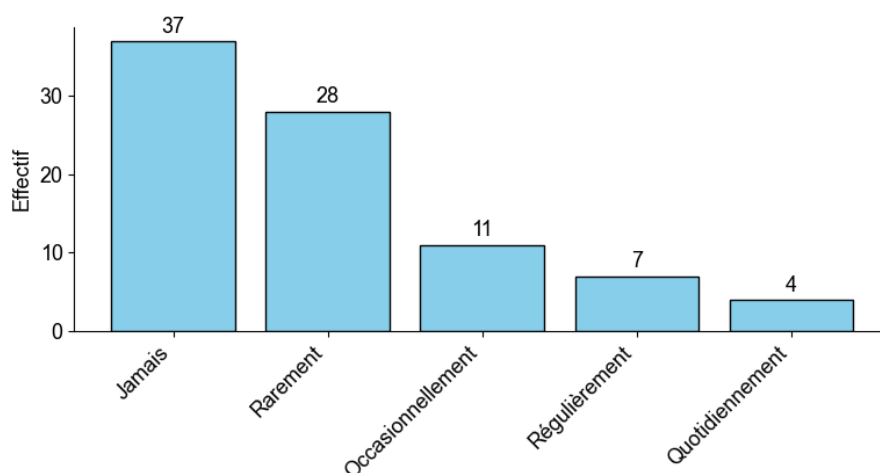


Figure 24 - Répartition des participants sur la fréquence d'utilisation des outils d'IA générative au quotidien

La réponse la plus fréquente était **Jamais** avec 42.53% des participants. Les réponses **Rarement** et **Occasionnellement** étaient également bien représentées, avec respectivement 28 et 11 occurrences, soit 32.18% et 12.64% des réponses.

Une majorité des participants (42.53%) n'utilise donc jamais ces outils, ce qui pourrait indiquer une méconnaissance ou une réticence à adopter ces technologies. Une proportion significative l'utilise rarement ou occasionnellement (44.82%) tandis qu'une minorité en fait un usage régulier voir quotidien (12.64%). Ces résultats pourraient refléter une adoption progressive des outils d'IA générative dans le quotidien.

b) Utilisation professionnelle des IA génératives

Nous avons cherché à distinguer, parmi les utilisateurs, ceux qui en faisaient une utilisation dans leur pratique professionnelle.

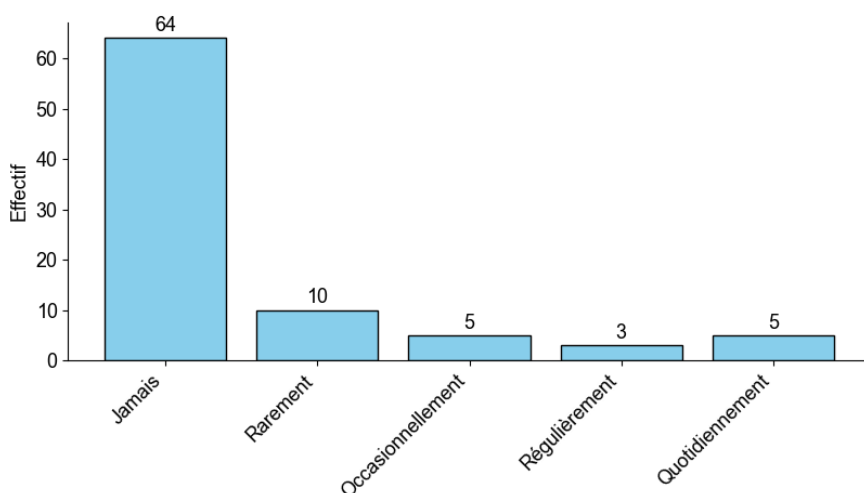


Figure 25 - Répartition des participants sur la fréquence d'utilisation professionnelle des outils d'IA générative

La réponse la plus fréquente était de loin **Jamais**, avec 64 occurrences, représentant 73.56% des réponses. On retrouvait 10 occurrences de la réponse **Rarement**. Les réponses **Occasionnellement**, **Régulièrement** et **Quotidiennement** étaient moins fréquentes, avec respectivement 5, 3 et 5 occurrences, soit 5.75%, 3.45% et 5.75% des réponses.

La majorité des médecins n'utilisaient jamais ces outils dans la pratique professionnelle, et ce malgré une adoption plutôt progressive dans leur quotidien comme le montrait le graphique précédent. Cela montrait une réelle réticence des médecins à incorporer les IA génératives dans le milieu médical.

5. Tâches intéressantes

Nous avons demandé aux participants de choisir les tâches pour lesquelles les outils d'intelligence artificielle générative étaient ou seraient les plus utiles. 3 choix étaient possibles par participant.

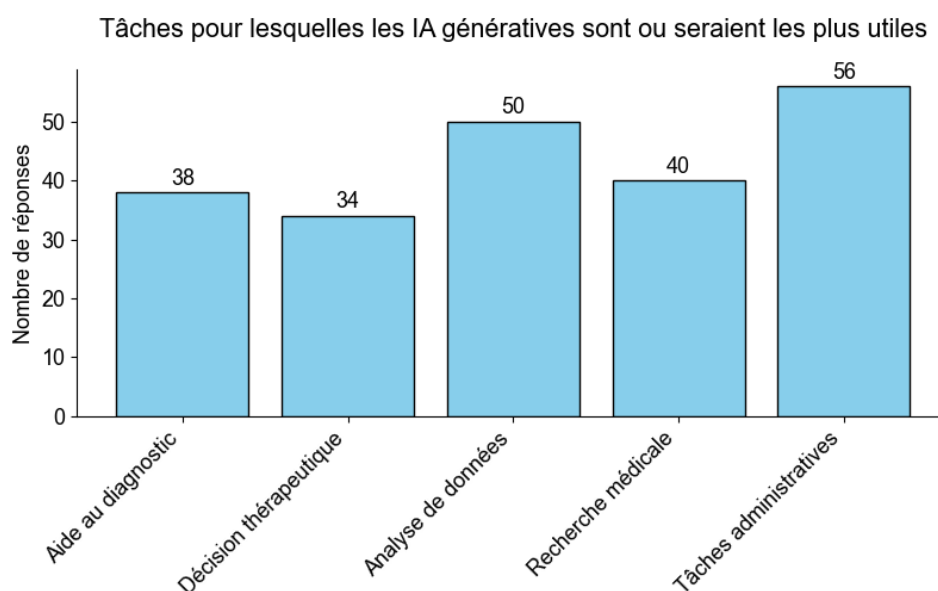


Figure 26 - Choix des participants sur les tâches dans lesquelles l'IA générative serait la plus utile

Les domaines dans lesquels l'IA générative aurait le plus d'intérêt étaient les **tâches administratives** (25.69%) et **l'analyse de données** (22.94%). Celle pour laquelle on retrouvait le moins de réponses était la décision thérapeutique (15.60%).

6. Principaux freins à l'utilisation de l'intelligence artificielle générative

Les médecins étaient invités à sélectionner les raisons qui pouvaient les dissuader de recourir aux IA génératives dans le cadre de leur pratique. 3 choix étaient possibles par participant.

Les **préoccupations éthiques ou de confidentialité des données** étaient les principaux freins à l'utilisation de l'IA générative par les médecins (21.68%).

Dans des proportions relativement identiques, on retrouvait également le **manque de confiance dans les résultats** (21.43%), le **manque de formation et de connaissance** (19.20%), ou encore le **défait d'intégration aux logiciels de travail** (18.75%).

Les motifs les moins représentés étaient la crainte de la perte d'emploi (4.02%) et la perception négative par les patients (3.12%).

7. Degré de favorabilité des médecins

Nous avons recueilli le degré de favorabilité des médecins participant au questionnaire à l'arrivée des intelligences artificielles génératives dans leur pratique.

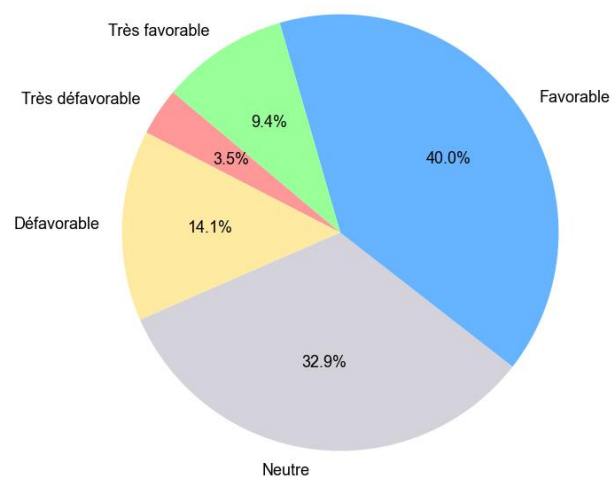


Figure 27 - Degré de favorabilité des médecins à l'arrivée des intelligences artificielles génératives dans leur pratique

Nous avons pu constater que près de la moitié des médecins (48.3%) étaient favorables ou très favorables au développement de l'IA générative dans leur pratique. Un tiers des médecins étaient neutres (32.2%) tandis que 17% étaient défavorables voir très défavorables.

8. Place de l'IA générative dans la pratique quotidienne

Nous avons interrogé les médecins sur la proportion de leurs consultations dans lesquelles les intelligences artificielles génératives pourraient leur être utiles.

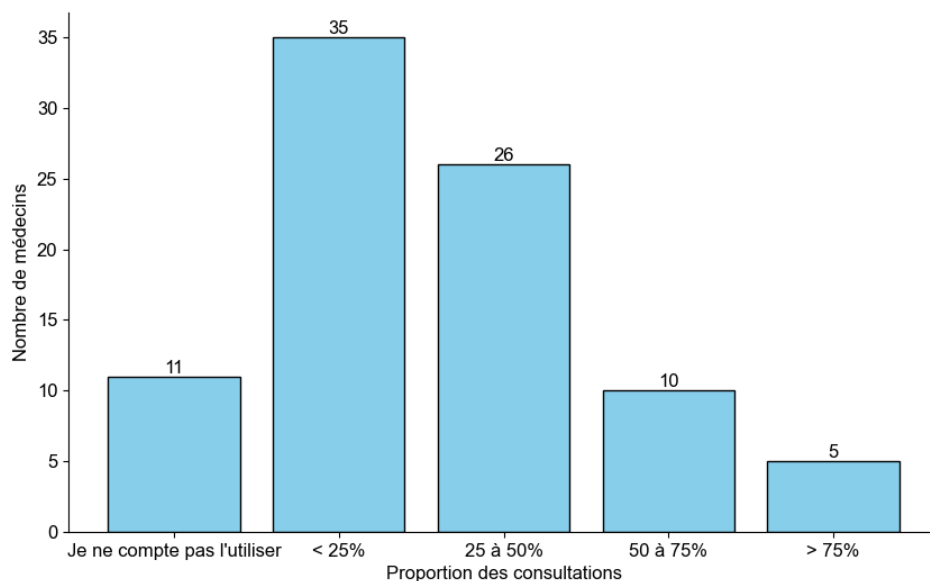


Figure 28 - Répartition des réponses à la question "Dans quelle proportion de vos consultations pensez-vous que l'utilisation de l'IA générative pourrait vous aider de façon significative ?"

On constatait des avis divergent dans ces réponses. La majorité des médecins semblait néanmoins penser que l'intelligence artificielle générative pourra être utile dans moins de 25% de leurs consultations. Près de 30% des médecins estimaient que l'IA générative pourrait être utile dans 25 à 50% des cas. Une très petite minorité de médecins, 5% environ, estimait qu'elle prendra 75% de leurs consultations. Un certain nombre de médecins, environ 13%, restaient contre l'idée de l'utiliser.

Bien que la majorité des médecins voient une utilité limitée de l'IA, une proportion significative reconnaît son potentiel à jouer un rôle dans une part notable de leurs consultations.

9. *Souhait de formation*

Dans le cadre de PENSO (Programme d'Enseignement du Numérique en Santé de l'Ouest), nous avons interrogé les médecins sur leur souhait d'en apprendre plus sur cette technologie.

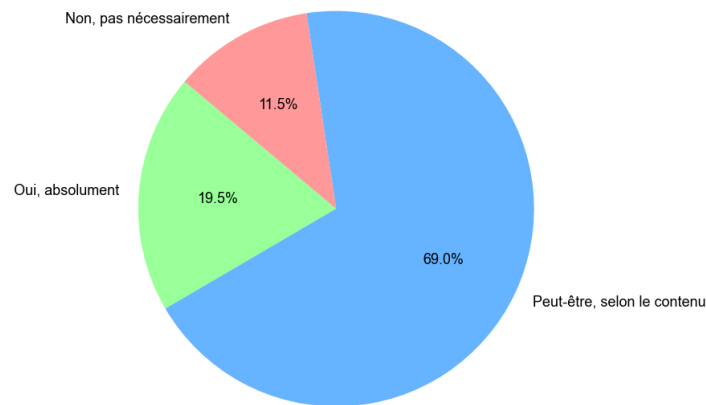


Figure 29 - Répartition des réponses sur le besoin de suivre une formation.

Nous pouvons observer qu'une minorité de médecins (11.5%) ne ressent pas le besoin de suivre une formation. Cela peut refléter soit une perception négative envers l'IA, soit une confiance dans les propres compétences actuelles pour gérer cette technologie. La majorité des médecins (69%) est ouverte à l'idée de suivre une formation selon son contenu. Enfin, on constate qu'une minorité significative (19.5%) ressent un besoin d'apprendre plus sur cette technologie.

10. *Mode de formation*

Nous avons souhaité évaluer le mode d'apprentissage préféré par les médecins pour se former dans ce domaine.

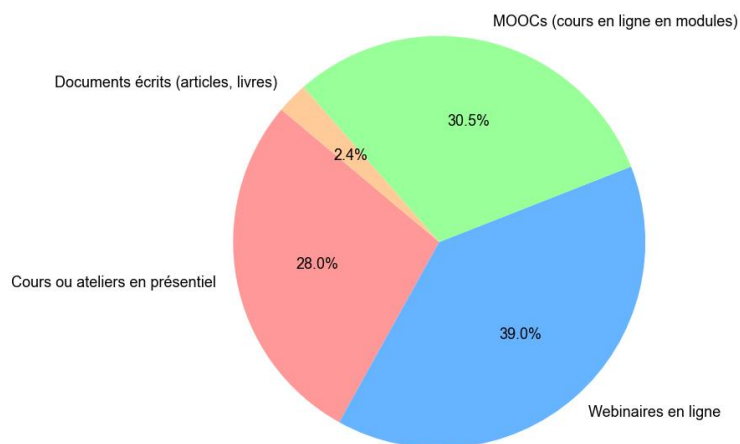


Figure 30 - Répartition des réponses sur le format souhaité pour la formation

Les webinaires en ligne (39%) et les MOOCs (30.5%) sont les formats les plus populaires, suivis de près par les cours ou ateliers en présentiel (28%). Les documents écrits sont le format le moins préféré (2.4%). Les options en ligne et en présentiel semblent privilégiées. Les articles ou livres semblent susciter peu d'intérêt auprès des médecins.

Variable	Catégorie	Effectif	Pourcentage	Intervalle de confiance
Sexe	Femme	56	64.4%	[0.542, 0.746]
	Homme	30	34.5%	[0.249, 0.441]
	Non répondu	1	1.1%	N/A
Âge	< 30 ans	17	19.5%	[11.4%, 27.6%]
	30 – 40 ans	43	49.4%	[39%, 59.8%]
	40 – 50 ans	17	19.5%	[11.4%, 27.6%]
	> 50 ans	9	10.3%	[4.0%, 16.6%]
	Non répondu	1	1.3%	N/A
Secteur d'activité	MSP	34	39.08%	[28.28%, 49.88%]
	Cabinet de groupe	38	43.68%	[32.88%, 54.48%]
	Cabinet seul	2	2.30%	[0.00%, 5.20%]
	Remplacement	13	14.94%	[7.44%, 22.44%]
Milieu d'exercice	Urbain	34	37.78%	[27.58%, 47.98%]
	Semi-rural	45	50.00%	[39.60%, 60.40%]
	Rural	11	12.22%	[5.32%, 19.12%]
Département	Loire-Atlantique	49	56.32 %	N/A
	Vendée	21	24.14%	N/A
	Maine-et-Loire	9	10.34%	N/A
	Seine-et-Marne	2	2.30%	N/A
	Morbihan	1	1.15%	N/A
	Nord	1	1.15%	N/A
	Oise	1	1.15%	N/A
	Sarthe	1	1.15%	N/A
	Mayenne	1	1.15%	N/A
	Seine-Maritime	1	1.15%	N/A
Compétence	Pas du tout compétent	2	2.30%	[0%, 5.30%]
	Peu compétent	23	26.44%	[17.24%, 35.64%]
	Compétent	52	59.77%	[49.57%, 69.97%]
	Très compétent	10	11.49%	[4.79%, 18.19%]
Avez-vous déjà interagi, à votre connaissance, avec une intelligence artificielle ?	Oui	62	71.26%	[62.05%, 80.46%]
	Non	25	28.74%	[19.54%, 37.94%]
Avez-vous déjà interagi avec une IA dans le cadre professionnel ?	Oui	31	35.63%	[25.63%, 45.63%]
	Non	56	64.37%	[54.37%, 74.37%]
Avez-vous déjà interagi avec une IA générative ?	Oui	43	49.43%	[39.03%, 59.83%]
	Non	44	50.57%	[40.17%, 60.97%]
Sur une échelle de 1 à 10, quel serait selon vous le potentiel des IA génératives à transformer votre activité professionnelle au quotidien ?	1	2	2.30%	N/A
	2	4	4.60%	N/A
	3	8	9.20%	N/A
	4	10	11.50%	N/A
	5	10	11.50%	N/A
	6	12	13.80%	N/A
	7	13	14.90%	N/A
	8	14	16.10%	N/A
	9	4	4.60%	N/A

	10	10	11.50%	N/A
Fréquence d'utilisation quotidienne	Jamais	37	42.53%	[32.14%, 52.92%]
	Rarement	28	32.18%	[22.37%, 42.00%]
	Occasionnellement	11	12.64%	[5.66%, 19.63%]
	Régulièrement	7	8.05%	[2.33%, 13.76%]
	Quotidiennement	4	4.60%	[0.20%, 9.00%]
Fréquence d'utilisation professionnelle	Jamais	64	73.56%	[64.30%, 82.83%]
	Rarement	10	11.49%	[4.79%, 18.20%]
	Occasionnellement	5	5.75%	[0.86%, 10.64%]
	Régulièrement	3	3.45%	[0%, 7.28%]
	Quotidiennement	5	5.75%	[0.86%, 10.64%]
Tâches les plus utiles pour l'IA générative	Aide au diagnostic	38	17.43%	[12.40%, 22.47%]
	Décision thérapeutique	34	15.60%	[10.78%, 20.41%]
	Analyse de données	50	22.94%	[17.35%, 28.52%]
	Recherche médicale	40	18.35%	[13.21%, 23.49%]
	Tâches administratives	56	25.69%	[19.89%, 31.49%]
Principaux freins à l'utilisation de l'IA générative	Crainte d'une perte d'emploi	9	4.02%	[1.45%, 6.59%]
	Manque de formation ou de connaissances sur l'outil	43	19.20%	[14.04%, 24.35%]
	Manque de confiance dans les résultats de l'IA	48	21.43%	[16.06%, 26.80%]
	Défaut d'intégration de l'IA aux logiciels de travail	42	18.75%	[13.64%, 23.86%]
	Préoccupations éthiques ou de confidentialité des données	62	27.68%	[21.82%, 33.54%]
	Perception négative par les patients	7	3.12%	[0.85%, 5.40%]
	Problèmes de coût	13	5.80%	[2.74%, 8.87%]
Favorabilité des médecins au développement de l'IA générative dans leur pratique	Très défavorable	3	3.45%	[0%, 7.28%]
	Défavorable	12	13.79%	[6.55%, 21.04%]
	Neutre	28	32.18%	[22.37%, 42.00%]
	Favorable	34	39.08%	[28.83%, 49.33%]
	Très favorable	8	9.20%	[3.12%, 15.27%]
Proportion de l'IA dans la pratique professionnelle	> 75%	5	5.57%	[0%, 7.28%]
	50 à 75%	10	11.49%	[6.55%, 21.04%]
	25 à 50%	26	29.89%	[22.37%, 42.00%]
	< 25%	35	40.23%	[28.83%, 49.33%]
	Je ne compte pas l'utiliser	11	12.64%	[3.12%, 15.27%]
Souhait de suivre une formation	Oui, absolument	17	19.54%	[11.21%, 27.87%]
	Peut-être, selon le contenu	60	68.97%	[59.24%, 78.69%]
	Non, pas nécessairement	10	11.49%	[4.79%, 18.20%]
Mode de formation	Cours ou ateliers (présentiel)	23	28.05%	[18.33%, 37.77%]
	Webinaires en ligne	32	39.02%	[28.47%, 49.58%]
	MOOCs (cours en ligne)	25	30.49%	[20.52%, 40.45%]
	Documents écrits (articles)	2	2.44%	[0%, 5.78%]

Tableau 4 - Tableau récapitulatif des réponses

C. Question ouverte

Pour rappel, nous proposons en fin de formulaire une question ouverte qui était : « Vos suggestions ou commentaires sont les bienvenus ». Cette question a reçu 13 réponses, pour 87 répondants (taux de réponse de 15%). Ci-après, les réponses proposées :

- Attente d'une IA faisant gagner du temps sur la rédaction d'observation médicale. Par l'écoute de la consultation en synthétisant et en ordonnant les éléments pertinents ? Serait déjà existant peut être, mais je n'ai jamais utilisé, et je n'en ai pas trouvé la trace. Questionne sur la confidentialité des informations cependant.
- Utilisation de nablacopilot parfois pour synthétiser et écrire à ma place. J'attends l'intégration des outils notamment dans le logiciel doctolib pour pouvoir utiliser en pratique ! J'attends un gain de temps dans les consultations avec l'administratif fait immédiatement, Exemple : certificat de sport ou enfant malade (qui sort tout seul après notre interrogatoire et examen clinique) ou même ordonnance et prise de sang, on dit à l'oral nos changements et on nous prépare l'ordonnance, les bons de transports et on a juste besoin de vérifier, valider et gagner du temps pour tout le monde. Avec une vérification et alerte sur les contre-indications + par exemple vérifier dans les données de la science quel traitement serait le plus adapté à la situation (et le must adapté à ce patient là d'après les consultations perçues). Je ne sais pas si c'est clair. Mais merci de cette thèse.
- je n'utilise pas ces outils à ce jour, je ne sais pas les utiliser et me méfie du résultat. J'utilise les algorithmes score, antibioclic.... le dernier mot doit être réservé à l'intelligence humaine.
- Je pense très utile pour les synthèses de dossier médicaux et consultations, les rappels (vaccins, rdv à prendre sur des suivis type diabète ou HTA). Sûrement gros gain de temps à l'avenir mais réticences sur l'éthique et le secret médical notamment.
- IA pour remplir les données administratives des dossiers: INS ...etc ...Pour coder les diagnostics (qu'elle se débrouille avec ce que j'ai écrit), remplir les dossiers divers et variés (MDPH, VIATRAJECTOIRE ...) à partir de mon dossier ... bref qu'elle fasse le travail inutile qu'on m'oblige à faire aux dépens du temps médical !!! Et temps qu'à faire: qu'elle le fasse depuis chez sa maman la Sécu !
- La question 9 est difficile à répondre puisqu'il n'y a actuellement pas ou peu de propositions...
- en soit cela me paraît une bonne idée pour améliorer nos compétences diagnostiques mais l'évolution en général vers l'IA m'angoisse beaucoup puisque l'on ne saura plus ce qui est humain ou tiré de l'IA et cela nous rends "paresseux" de ne plus avoir besoin de réfléchir/chercher/apprendre puisqu'une machine le fera pour nous et donc cela risque de changer +++ notre façon de soigner.

- J'utilise Nabla au quotidien. ChatGPT a remplacé avantageusement Doocteur et autres moteurs de recherche chronophage. J'attends une plus grande intégration avec les logiciels médicaux
- Je n'y avais jamais songé, je n'ai aucune idée de l'impact que cela peut avoir
- Je souhaite continuer à faire mes propres démarches diagnostiques, à réfléchir lors de mes consultations. Je ne souhaite pas me reposer sur un outil informatique avec le risque de perdre toutes mes compétences cliniques diagnostiques et thérapeutiques
- Actuellement le logiciel doctolib utilise une IA lors de l'intégration de courrier / CR afin de repérer : date/provenance. Autres usages administratifs seraient les bienvenus +

Deux réponses apportaient peu d'éléments et n'ont pas été reproduites ici. Les réponses reprennent globalement ce qui a déjà été mis en évidence par les questions. Cinq d'entre elles expriment le souhait d'une IA offrant une aide aux tâches administratives. Parmi les intelligences artificielles utilisées nommées on retrouve l'assistant Nabla et Doctolib, utilisés pour la rédaction automatique. Deux participants expliquent n'utiliser aucune IA. Parmi les craintes évoquées, plusieurs répondant évoquent le risque de perte de compétences ou la volonté de maintenir ces compétences.

D. Résumé des résultats

Les répondants à notre questionnaire étaient majoritairement des femmes (64.4%), pour la plupart de moins de 40 ans (69%), exerçant le plus souvent en groupe (82.8%) et en semi-rural (50%). Les réponses obtenues proviennent majoritairement des départements de Loire-Atlantique et de Vendée.

La plupart des participants étaient familiers avec l'intelligence artificielle « en général » (71%) et moins de la moitié déclaraient déjà avoir eu affaire à l'IA générative (49%). L'utilisation dans le cadre professionnel était encore inférieure (35%). Les répondants envisageaient un fort impact de l'IA générative dans leur métier. En ce qui concerne la fréquence d'utilisation, celle-ci était limitée avec 11 répondants rapportant une utilisation régulière ou quotidienne de l'IA générative.

Les tâches pour lesquelles les médecins envisageaient l'utilisation de l'IA générative étaient en premier lieu l'analyse des données, et les tâches administratives. L'aide au diagnostic arrivait en quatrième position sur cinq propositions. Plus de la moitié des médecins n'envisageaient un intérêt à l'utilisation de l'IA générative que dans moins de 25% de leurs consultations.

La moitié des répondants étaient favorables ou très favorables à l'arrivée de l'IA générative dans leur profession. 1/3 des répondants se disaient neutres sur cette question. Les deux principaux freins à l'utilisation de l'IA étaient les considérations éthiques et le manque de confiance dans les résultats.

II. Évaluation des grands modèles de langue

A. Caractéristiques des réponses

1. Taux de réponse

Comme expliqué dans la section méthodologie, chaque cas clinique était soumis via un prompt à chaque modèle de langue, en anglais et en français. Il leur était demandé de fournir une réponse au format JSON, contenant à la fois le diagnostic clinique, puis le raisonnement. Le taux de réponse a atteint les **100% pour GPT et Llama**. Seul BioMistral a retourné des réponses nulles. En français, cela s'est produit pour 32 cas, soit un taux de réponse de **82%**. En anglais, BioMistral a envoyé une réponse nulle pour 38 cas, soit un taux de réponse comparable de **78%**.

2. Longueur des raisonnements

En dehors du système de notation, déjà décrit ci-dessus, nous disposons de peu de moyens mathématiques pour évaluer un raisonnement clinique. Nous nous sommes néanmoins intéressés à la longueur des raisonnements (mesurée en nombre de caractères) fournis dans les réponses des différents modèles. En voici un tableau descriptif :

	Minimum	Maximum	Moyenne	Écart-type
GPT (FR)	577	1469	960	163,93
GPT (EN)	552	1769	965	190,09
Llama (FR)	546	1804	920	197,89
Llama (EN)	669	1673	998	166,59
BioMistral (FR)	202	3735	1075	1003,26
BioMistral (EN)	126	4867	788	924,13

Tableau 5 - Minimums, maximums, et moyennes des longueurs des réponses fournies par les LLM

De prime abord, les modèles GPT et Llama présentent des moyennes identiques, quel que soit le langage.

En revanche, l'analyse des variables descriptives de BioMistral révèle des moyennes de longueur de réponse nettement supérieures à celles des autres modèles, avec des valeurs extrêmes aberrantes. De plus, l'écart-type plus élevé indique une grande variabilité des réponses au sein du modèle. Ces observations concordent avec les impressions que nous avons eu du modèle, que nous verrons plus loin.

En effet, nous avons pu constater au cours de la notation une fréquence importante d'hallucinations et de paraphrases des dossiers cliniques pour ce modèle.

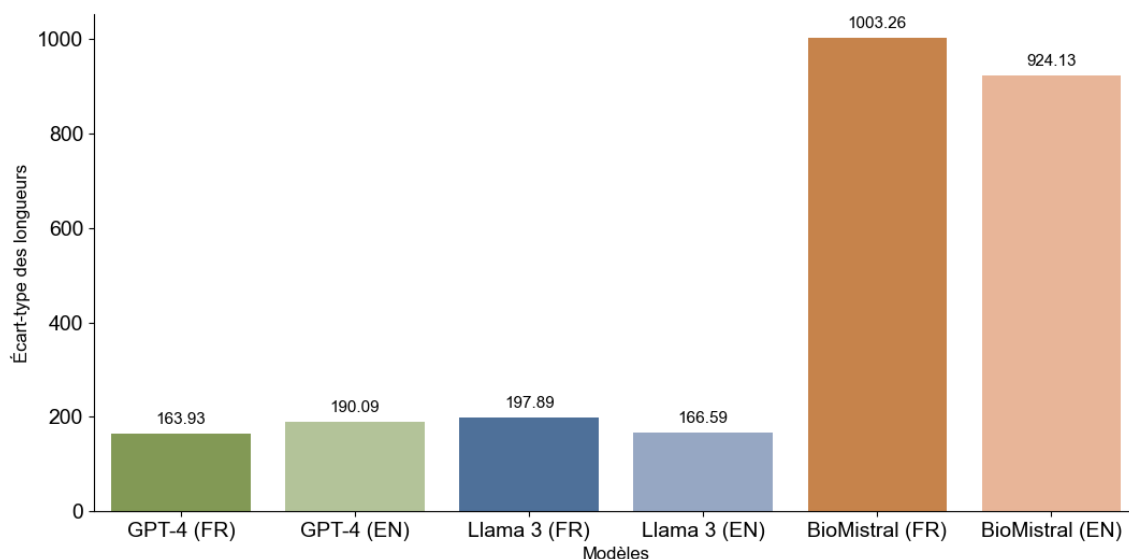


Figure 31 - Comparaison des écarts-types des longueurs des réponses, par modèle et par langue.

Bien que cette différence soit significative, il est important de noter que le nombre élevé de réponses nulles fournies par BioMistral a réduit considérablement la taille l'échantillon analysé (120 dossiers au lieu de 180), ce qui a pu augmenter l'écart-type.

3. Indice de lecture

Au-delà de la longueur des raisonnements, mesurée en nombre de caractères, il existe des indices basés sur des algorithmes mathématiques, permettant d'évaluer la complexité et la lisibilité d'un texte. Nous avons utilisé l'un de ces indices, valide uniquement pour l'anglais, et l'avons calculé pour chaque réponse fournie par les modèles dans cette langue. Les résultats sont présentés dans le tableau suivant.

L'indice de SMOG (Simple Measure of Gobbledygook) est une mesure de lisibilité qui estime le nombre d'années d'étude nécessaire pour comprendre un texte sans difficultés. Cet indice est particulièrement utilisé dans le domaine de la santé, notamment pour évaluer la lisibilité des documents destinés au grand public. Une étude réalisée en 2013 par l'Université de l'Oklahoma recommande l'utilisation de cet indice par rapport aux autres indices pour les évaluations médicales, en raison de sa fiabilité et de sa constance (60).

Modèle	Moyenne	Écart-type
GPT (EN)	16,4	1,57
Llama (EN)	14,96	1,83
BioMistral (EN)	11,35	3,14

Tableau 6 - Moyennes et écarts-type des indices de SMOG selon le modèle, en anglais.

Nous pouvons observer que l'indice de SMOG est plus élevé pour GPT et Llama.

B. Moyennes, écarts types, médianes, percentiles

Dans le cadre de l'évaluation des performances des modèles, nous affichons ci-dessous les variables associées à ces modèles, en fonction de la langue. Ces données offrent une perspective globale de la performance des LLMs.

	GPT-4		Llama 3		BioMistral	
	FR	EN	FR	EN	FR	EN
Moyenne	8,18	8,57	7,65	8,44	2,56	3,63
Écart type	1,61	1,63	1,92	1,63	1,94	2,39
Médiane	9	9	8	9	2	3
C75	9	10	9	10	4	5
C25	7	8	6	8	1	2

Tableau 7 - Tableau des moyennes, écarts types, médianes et percentiles par modèle selon la langue

C. Distribution des notes par modèle

1. GPT-4

Le graphique suivant présente la distribution des notes attribuées aux réponses de GPT-4, en anglais et en français. Les résultats sont représentés par deux histogrammes accompagnés de courbes de densité. L'objectif est d'avoir un premier aperçu de la performance des modèles en fonction de la langue d'évaluation.

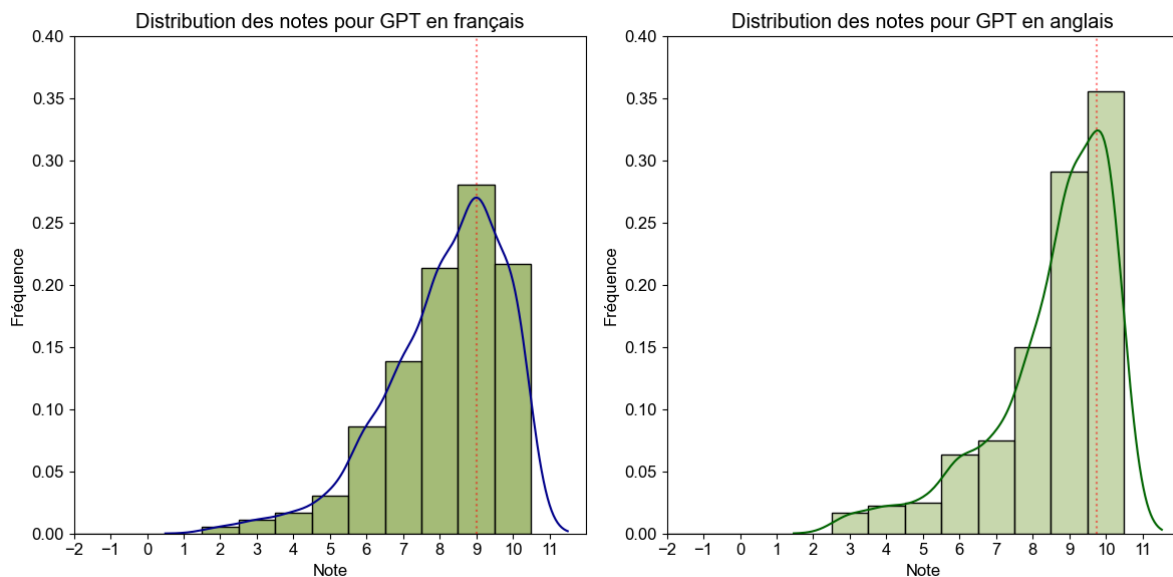


Figure 32 - Distribution des notes pour GPT-4, en français (à gauche) et anglais (à droite).

GPT-4 en français

L'histogramme de gauche illustre la distribution des évaluations attribuées aux réponses de GPT-4 sur les cas cliniques en français. Les notes s'étendent de 2 à 10, avec une concentration notable autour des valeurs élevées, particulièrement entre 8 et 10. Cependant, on observe un étalement des notes vers les valeurs plus faibles. La courbe de densité (KDE) superposée montre une asymétrie positive, avec une montée progressive vers un pic prononcé et relativement large autour de la note de 9. La pente moyenne sur les notes faibles reflète l'étalement des notes suggéré par l'histogramme.

GPT-4 en anglais

L'histogramme situé à droite représente la répartition des évaluations de GPT-4 pour les mêmes dossiers, mais présentés en anglais. Les notes varient de 3 à 10 avec une forte concentration sur les notes maximales. La courbe de densité montre un pic important et étroit sur la note de 10, indiquant une concentration élevée des notes sur cette plage. La pente de la courbe, plus raide, reflète une plus faible dispersion des notes basses.

Dans l'ensemble, ces deux graphiques mettent en évidence une tendance similaire, à savoir une performance globalement bonne de GPT, que ce soit en anglais et en français, pour répondre à des cas cliniques en français. Toutefois, la courbe de densité plus étroite et décalée vers la droite pour les réponses en anglais suggère une performance légèrement supérieure dans cette langue, avec une plus grande homogénéité des évaluations et une moindre dispersion.

2. Llama-3

Nous visualisons les résultats de Llama-3 à travers deux histogrammes accompagnés de courbes de densité, offrant ainsi un premier aperçu des performances du modèle selon la langue d'évaluation.

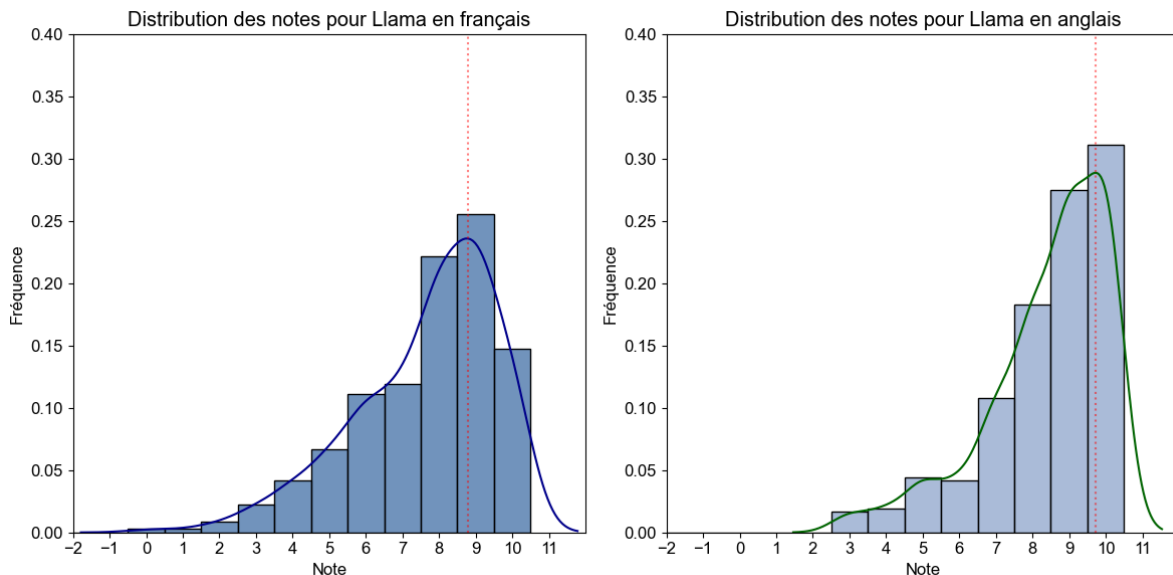


Figure 33 - Distribution des notes pour Llama, en français (à gauche) et anglais (à droite).

Llama-3 en français

L'histogramme de gauche représente la distribution des notes attribuées à Llama-3 sur les dossiers cliniques en langue française. Les notes s'étendent de 0 à 10, avec une concentration des valeurs entre 8 et 10. La courbe de densité superposée présente une asymétrie positive, avec une montée progressive des notes à partir des valeurs vers un pic large à environ 8.8 (8.78). On remarque un étalement des notes vers les valeurs inférieures, ce qui suggère une dispersion modérée des performances. La pente avant le pic est relativement douce, indiquant que bien que la majorité des notes soient élevées, certaines réponses sont encore évaluées dans les gammes de notes plus faibles, en particulier entre 6 et 7.

Llama-3 en anglais

L'histogramme de droite présente la répartition des notes pour les mêmes cas cliniques, évalués cette fois en anglais. Les notes varient de 3 à 10, avec une montée progressive vers les notes maximales et pic net à 10. La courbe de densité associée révèle une asymétrie positive marquée, avec une pente forte et un pic étroit autour de la note maximale, témoignant d'une faible dispersion des notes en amont de la valeur maximale.

Cela suggère qu'une grande partie des réponses de Llama a obtenu la note maximale et que peu d'entre elles ont obtenu des évaluations inférieures à 7.

3. BioMistral

Nous visualisons les résultats de BioMistral à travers deux histogrammes accompagnés de courbes de densité, offrant ainsi un premier aperçu des performances du modèle selon la langue d'évaluation.

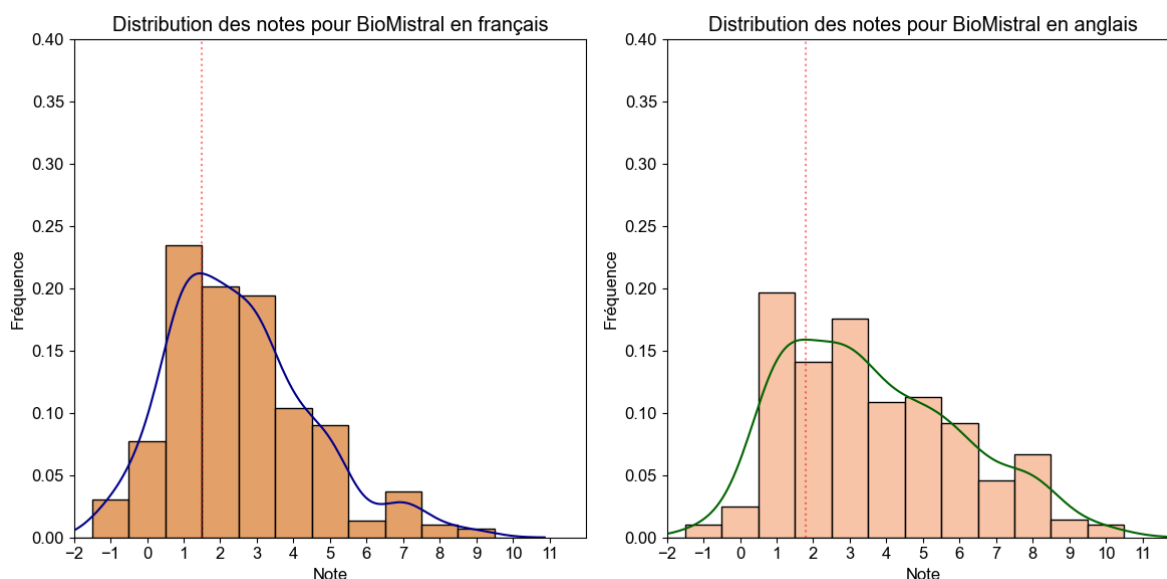


Figure 34 - Distribution des notes pour BioMistral, en français (à gauche) et anglais (à droite).

BioMistral en français

L'histogramme de gauche illustre la distribution des notes attribuées aux réponses de BioMistral sur les cas cliniques en français. Les notes s'étendent de -1 à 9 avec un pic notable à 1. La majorité des notes sont concentrées entre 1 et 3, ce qui témoigne d'une performance majoritairement faible. La courbe de densité superposée présente une légère asymétrie négative, avec un pic à 1.54 avec une pente douce de part et d'autre, suggérant une distribution hétérogène des notes. On constate également un léger pic secondaire à 7, ce qui pourrait refléter une variabilité dans la qualité des réponses.

BioMistral en anglais

L'histogramme de droite représente la distribution des notes attribuées aux réponses de BioMistral sur les cas cliniques en anglais. Les notes s'étendent de -1 à 10 avec un pic à 1. Les notes semblent concentrées sur une plage plus étendue, entre 1 et 6, démontrant une performance faible à moyenne, mais aussi une plus grande hétérogénéité dans la qualité des réponses. La courbe de densité renforce ces observations.

Elle ne présente pas de franche asymétrie. Nous n’observons pas de pic marqué mais un plateau large, et sa pente est plus douce en direction des valeurs moyennes, montrant une dispersion des notes assez importante dans cet intervalle.

En observant ces deux graphiques, on constate une performance de BioMistral faible dans la résolution de cas cliniques médicaux en anglais comme en français. En anglais, la courbe de densité montre une plus grande dispersion des notes vers les valeurs élevées, ce qui suggère que BioMistral est plus capable de produire des réponses pertinentes, bien que de manière sporadique.

4. Conclusions sur la distribution des notes

L’analyse des différents graphiques de distribution des notes révèle des différences notables en termes de performance et d’homogénéité des réponses.

1. En anglais, GPT-4 et Llama semblent présenter de très bonnes performances, comparables et homogènes. Leurs moyennes sont proches (respectivement 8.57 et 8.44) et leurs écarts types identiques (1.63) suggérant une homogénéité équivalente. Les courbes de densité, décrites ci-dessus, corroborent ces observations. BioMistral, de son côté, est en dessous en termes de performances, avec une moyenne à 3.63. Son écart type à 2.39 confirme l’hétérogénéité des notes mise en évidence par sa courbe.
2. En français, GPT-4 et Llama présentent de bonnes performances, avec des moyennes (respectivement 8.18 et 7.65) et des écarts types (1.61 et 1.92) sensiblement différents. Ces valeurs semblent en accord avec les courbes de densité, notablement décalées à gauche par rapport aux évaluations en anglais, et plus plates. BioMistral reste à nouveau très en dessous en termes de performances, avec une moyenne à 2.56 et un écart type à 1.94.

D. Profil des notes par item et par modèle

Comme expliqué en méthodologie, la note attribuée à chaque cas clinique est divisée en différents items, chacun évaluant un aspect de la réponse. Nous cherchons à observer et à comparer les performances des modèles pour chaque item.

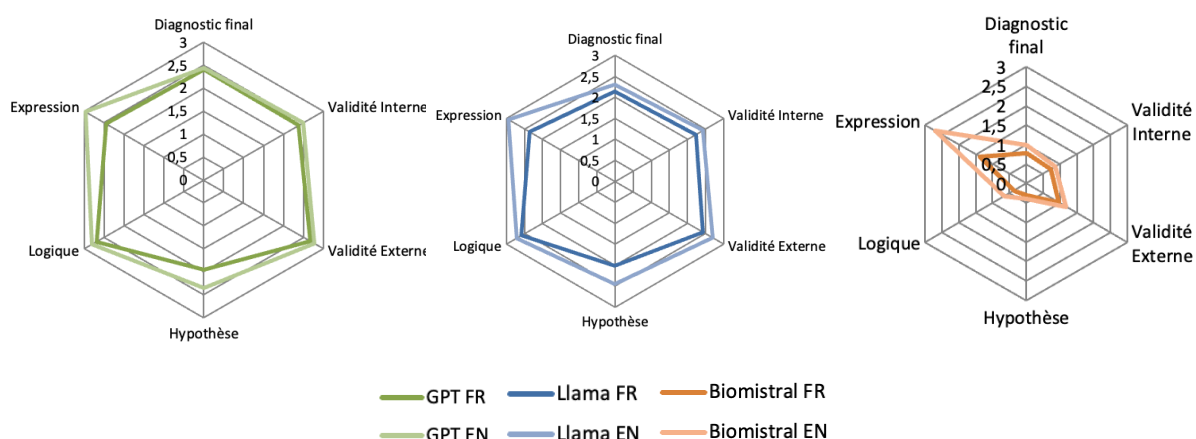


Figure 35 - Notes moyennes obtenues par item selon le modèle et la langue

Les diagrammes ci-dessus représentent des graphiques en toile d'araignée, utilisés pour comparer les performances des différents modèles selon les critères d'évaluation. Pour rappel, les items additionnés constituent la note finale sur 10 et sont : le Diagnostic final, la Validité interne, la Validité externe, l'Hypothèse, la Logique et l'Expression.

Pour rendre les diagrammes lisibles et cohérents, les notes moyennes pour chaque modèle ont été ramenées sur trois. Les notes sur l'expression ont été inversées et multipliées par dix pour la même raison.

Les résultats suggèrent que GPT en anglais et en français sont globalement équilibrés sur les dimensions évaluées. Les principales différences peuvent être observées sur les items Hypothèse et Expression. Les performances sur les items Diagnostic final, Validité interne, Validité externe et Logique sont très proches. Il semble y avoir une légère supériorité des réponses en anglais. Ces mêmes tendances peuvent être constatées pour Llama.

E. Analyse comparative des performances

Nous avons cherché à établir l'existence d'une différence globale de performance significative entre les différents modèles (GPT en français, GPT en anglais, Llama en français, Llama en anglais). Pour des raisons de viabilité statistique, et en raison des preuves d'infériorité déjà apportées par les données observationnelles, nous avons décidé d'exclure BioMistral de l'analyse.

1. Analyse des performances par paire

Dans le texte suivant, nous cherchons à confirmer nos premières observations par des analyses statistiques plus poussées.

Nous voulons établir s'il existe une différence statistiquement significative de performance entre les modèles, et s'il existe une différence statistiquement significative en fonction de la langue pour un modèle donné.

a) Performance de GPT-4 entre Français et Anglais

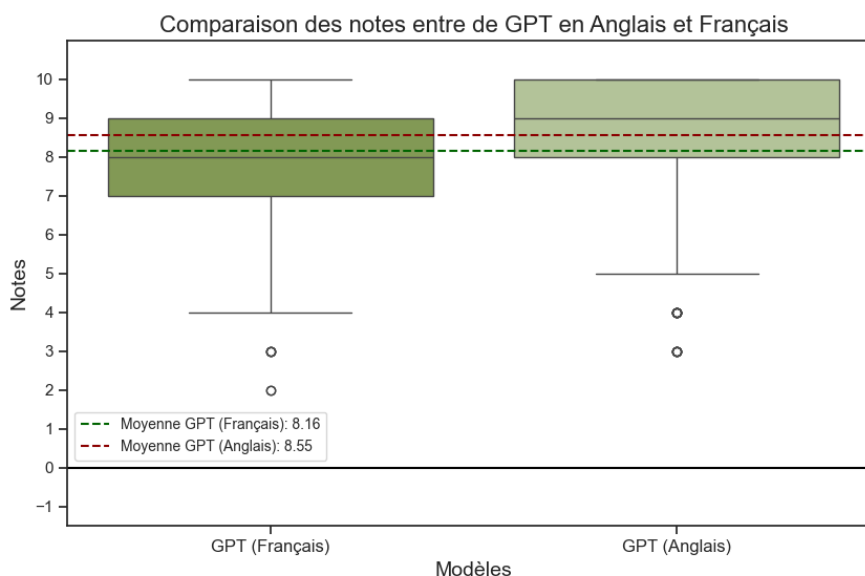


Figure 36 - Comparaison des notes entre GPT en français (à gauche) et en anglais (à droite) sous forme de boîtes à moustaches.

Le schéma ci-dessus présente une comparaison des notes attribuées par GPT en anglais et en français, sous forme de diagrammes en boîte. Nous pouvons en faire quelques observations.

En français, le modèle présente une médiane à 8, avec une moyenne de 8,16. La moitié des notes s'étend entre 7 et 9 (étendue interquartile). Les valeurs minimales et maximales sont respectivement 4 et 10. On retrouve des notes aberrantes. En effet, GPT a obtenu 4 fois la note de 3/10 et 2 fois la note de 2/10. En anglais, le modèle présente une médiane à 9, avec une moyenne à 8.55. La moitié des notes s'étend entre 8 et 10 (étendue interquartile). Les valeurs minimales et maximales sont respectivement 5 et 10. On retrouve plusieurs notes aberrantes, notamment 3/10 et 4/10 que le modèle a obtenu respectivement 6 et 8 fois.

En comparaison, on voit que les formes des boîtes sont globalement identiques, mais que celle en anglais est déviée vers le haut. La longueur des moustaches, plus longue en français, montre que le modèle tend à obtenir des notes plus dispersées dans cette langue. Nous pouvons conclure que, d'après ce diagramme, le modèle est sensiblement supérieur en anglais. Cela est en accord avec nos observations précédentes.

Les deux modèles présentent des valeurs aberrantes, ce qui montre que quel que soit la langue il existe des cas où GPT performe moins bien.

Nous avons souhaité établir cette supériorité statistiquement. Nous avons émis l'hypothèse nulle selon laquelle la distribution des différences de notes entre les deux langues suivrait une loi normale. Un test de normalité de Shapiro-Wilk a été réalisé et conduit toutefois au rejet de cette hypothèse ($p < 0.05$). La distribution ne suit pas une loi normale.

Face à cette non-normalité, nous avons recouru au test de Wilcoxon des rangs signés afin de tester l'hypothèse nulle selon laquelle il n'existe pas de différence significative entre les performances du modèle dans les deux langues. Les résultats de ce test ont permis de rejeter l'hypothèse nulle ($p < 0.0125$) suggérant qu'il existe une différence statistiquement significative entre GPT-4 en français et en anglais.

Ces résultats confirment statistiquement les observations faites à partir des graphiques de distribution des notes et des diagrammes en boîtes à moustaches. Les détails des analyses statistiques sont fournis en annexe. **En conclusion, nous pouvons affirmer qu'il existe, dans le cadre de notre étude, une différence significative de notes entre GPT-4 en anglais et en français, avec une supériorité marquée pour l'anglais.**

b) Performance de Llama-3 entre Français et Anglais

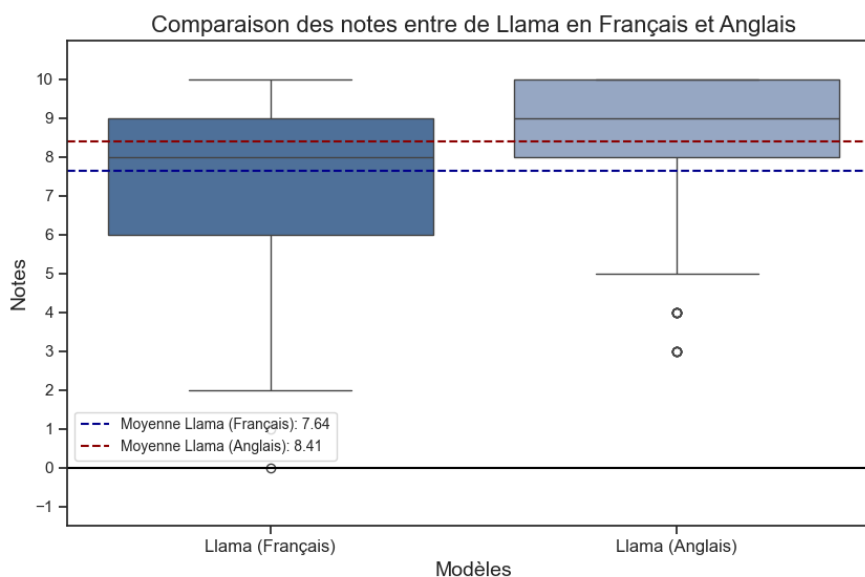


Figure 37 - Comparaison des notes entre Llama en français (à gauche) et en anglais (à droite) sous forme de boîtes à moustaches.

Le schéma ci-dessus présente une comparaison des notes attribuées par Llama-3 en anglais et en français, sous forme de diagrammes en boîte.

En français, le modèle présente une médiane à 8 avec une moyenne à 7.64/10. La moitié des notes s'étend de 6 à 9, ce qui représente une plage de 3 unités, plus large que celle observée dans les autres modèles. On observe des moustaches significativement étendues, avec des valeurs minimales et maximales respectivement à 2 et 10. Par ailleurs, la valeur minimale est nettement plus basse que celles des autres modèles étudiés. Deux valeurs aberrantes sont notables : des notes de 0/10 et 1/10, chacune obtenue une fois.

En anglais, le modèle présente une médiane à 9, avec une moyenne de 8.41/10. La moitié des notes s'étendent entre 8 et 10, soit une plage de 2 unités, plus restreinte. La valeur minimale est à 5, tandis qu'on retrouve deux valeurs aberrantes à 3 et 4, obtenues respectivement 6 et 7 fois.

En comparaison, on remarque une boîte plus étendue pour le modèle en langue française par rapport à l'anglais. Cela indique une dispersion plus importante des notes, témoignant d'une plus grande variabilité des performances du modèle en français. Cette impression est renforcée par l'aspect des moustaches, plus longue en langue française.

Le modèle présente quelques valeurs aberrantes sur les deux diagrammes, ce qui suggère qu'il existe des cas où Llama-3 performe moins bien quelle que soit la langue.

Nous avons émis l'hypothèse selon laquelle la distribution des différences de notes entre les deux langues suivrait une loi normale. Cependant, le test de normalité de Shapiro-Wilk a conduit au rejet de cette hypothèse ($p < 0.05$). Face à cette non-normalité, nous avons employé le test de Wilcoxon des rangs signés. Les résultats de ce test ont permis de rejeter l'hypothèse nulle ($p < 0.0125$), suggérant qu'il existe une différence statistiquement significative entre Llama-3 en français et en anglais dans la résolution de cas clinique médicaux.

Ces résultats confirment statistiquement les observations faites à partir des graphiques de distribution des notes et des diagrammes en boîte à moustaches. Les détails des analyses statistiques sont fournis en annexe. **En conclusion, nous pouvons affirmer qu'il existe, dans le cadre de notre étude, une différence significative de notes entre Llama-3 en anglais et en français, avec une supériorité marquée pour la version en anglais.**

c) Performance de GPT-4 versus Llama-3 en Français

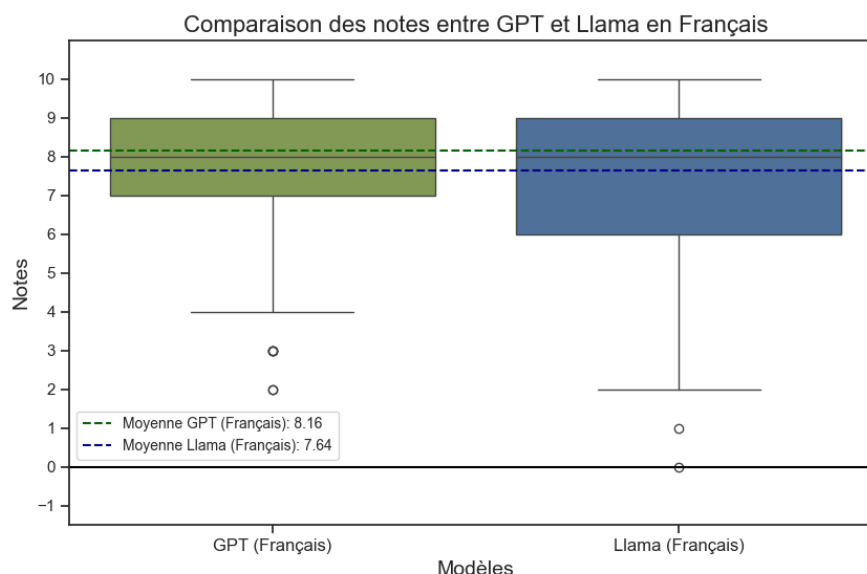


Figure 38 - Comparaison des notes entre GPT en français (à gauche) et Llama en français (à droite) sous forme de boîtes à moustaches.

Le schéma ci-dessus présente une comparaison des notes attribuées à GPT-4 et Llama-3 en français. Nous avons eu l'occasion précédemment de décrire les boîtes pour chacun de ces modèles.

En comparant les deux graphiques, on constate une boîte moins large et des moustaches plus courtes pour GPT-4 par rapport à Llama-3. Si la médiane est identique, on constate que la moyenne est plus élevée pour GPT-4 (8.16 contre 7.64). Les valeurs aberrantes de Llama-3 descendent jusqu'à 0 et 1, tandis que celles de GPT-4 sont moins extrêmes (2 et 3). Ces observations suggèrent que ce dernier présente des performances supérieures et plus homogènes que Llama-3 en français.

Nous avons cherché à valider cette supériorité de façon statistique. Le test de normalité de Shapiro-Wilk nous indique que la distribution ne suit pas une loi normale. Face à cette non-normalité, nous avons employé le test de Wilcoxon des rangs signés. Les résultats de ce test ont permis de rejeter l'hypothèse nulle ($p < 0.0125$), indiquant qu'il existe une différence statistiquement significative de performance entre GPT-4 et Llama-3 dans la résolution de cas cliniques médicaux en français.

Ces résultats confirment statistiquement les observations faites à partir des graphiques de distribution des notes et des diagrammes en boîte à moustaches.

En conclusion, nous pouvons affirmer qu'il existe, dans le cadre de notre étude, une différence significative de notes entre GPT-4 et Llama-3 dans le raisonnement diagnostique en français, avec une supériorité pour GPT-4.

d) Performance de GPT-4 versus Llama-3 en Anglais

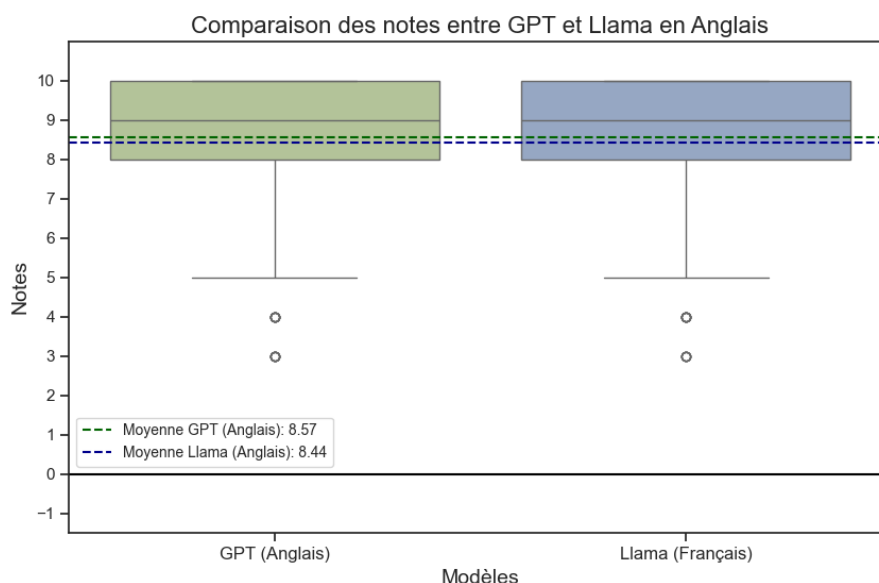


Figure 39 - Comparaison des notes entre GPT en anglais (à gauche) et Llama en anglais (à droite) sous forme de boîtes à moustaches.

La figure ci-dessus présente une comparaison des notes attribuées à GPT-4 et Llama-3 en langue anglaise, sous forme de diagramme à boîte. Nous avons eu l'occasion de décrire leur aspect précédemment.

En comparant les deux graphiques, on observe deux boîtes dont l'étendue est identique (8 à 10), avec la même médiane. Les moustaches sont elles aussi de la même longueur, avec des valeurs aberrantes similaires (4 et 3). Seule la moyenne diffère, légèrement en faveur de GPT-4 (8.57 contre 8.44). Nous pouvons donc suggérer une performance similaire entre les deux modèles en langue anglaise.

Nous avons procédé à des tests statistiques afin de valider cette observation. Un test de normalité de Shapiro-Wilk a été réalisé et conduit toutefois au rejet de cette hypothèse ($p < 0.05$). La distribution ne suit pas une loi normale. Face à cette non-normalité, nous avons recouru au test de Wilcoxon des rangs signés. Les résultats de ce test n'ont pas permis de rejeter l'hypothèse nulle ($p > 0.0125$), démontrant l'absence de différence statistiquement significative entre GPT-4 et Llama-3 en termes de performance.

Ces résultats confirment statistiquement l'impression visuelle donnée par les diagrammes à moustaches. **En conclusion, nous pouvons affirmer qu'il n'y a pas, dans notre étude, de preuve statistique indiquant une différence de notes entre GPT-4 et Llama-3 dans le raisonnement diagnostique en anglais.**

2. Conclusions de l'étude comparative

Notre étude nous permet à ce stade de tirer plusieurs conclusions :

1. Il n'existe pas de différence significative de notes entre GPT-4 et Llama-3 dans la résolution de cas cliniques en langue anglaise.
2. GPT-4 et Llama-3 a de meilleures notes en anglais qu'en français, avec des résultats plus homogènes.
3. GPT-4 présente des performances supérieures de manière statistiquement significative à celles de Llama-3 en français.
4. En langue française, les performances de GPT-4 et de Llama-3 sont non seulement inférieures à celles observées en anglais, mais également plus hétérogènes.
5. Il existe des situations où GPT-4 et Llama-3 produisent de mauvais résultats, quelle que soit la langue.

F. Performance des modèles selon la spécialité

Nous avons cherché à établir s'il existait une différence de performance des modèles en fonction de la spécialité. Le tableau ci-dessous récapitule les moyennes obtenues par chaque modèle selon la spécialité et la langue.

	Anglais			Français		
	GPT	Llama	BioMistral	GPT	Llama	BioMistral
Urgence-Réanimation	8.25	8.65	4.3	7.55	6.9	2.81
Endocrinologie	8.17	8.72	2.93	8.06	6.5	2.07
Gynécologie	9.17	8.06	3.88	8.67	8.17	3.5
Onco-Hématologie	8.72	8.17	3	8.44	6.78	1.64
Hépatogastro-entérologie	8.47	8.6	3.62	8.37	8.07	2.14
Maladies infectieuses	8.5	8.62	3.65	8	8.24	2.47
Médecine cardiovasculaire	8.71	8.42	3.55	8.83	7.71	2.6
Médecine générale	8.77	8.44	3.61	8.38	7.67	2.43
Médecine interne	8.75	8.33	3.33	8.17	7.42	1.9
Neurologie	8.45	7.95	2.75	8.36	7.41	2.75
Tête et cou	8.14	8.21	3.58	6.93	7.57	2.17
Pédiatrie	8.07	8.64	5	8.5	7.5	2.25
Pneumologie	9.4	9	5	8.7	9.15	3.94
Psychiatrie	8	8.19	5	7.62	8.12	3.38
Rhumatologie	8.17	8.22	2.29	7.78	7.06	2.5
Urologie-Néphrologie	8.39	7.94	4.25	7.06	6.94	1.67

Tableau 8 - Moyennes par spécialité selon le modèle et la langue

1. GPT-4

a) GPT-4 en français

Nous comparons dans le graphique suivant les moyennes obtenues par GPT, en langue française, selon la spécialité.

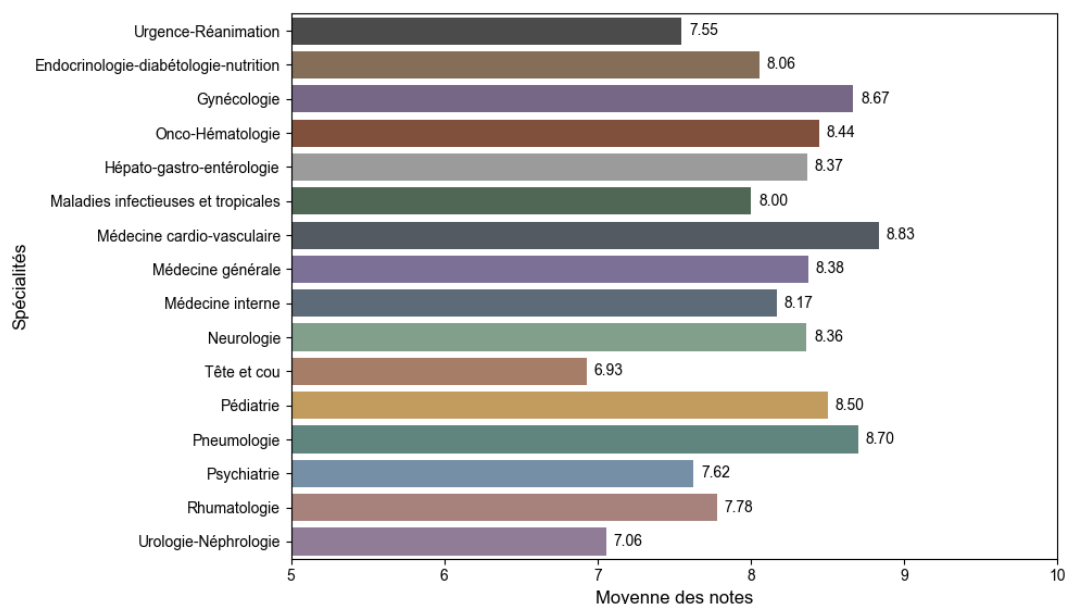


Figure 40 - Moyennes obtenues par GPT en français selon la spécialité

Sur ce graphique, nous observons des moyennes élevées pour des spécialités telles que la Médecine cardiovasculaire, la Pneumologie, la Pédiatrie et la Gynécologie. La Médecine cardiovasculaire se distingue avec la meilleure moyenne, atteignant 8.83. Il est également notable que quatre spécialités ont une moyenne inférieure à 8 : Urgence-Réanimation, Tête et cou, Psychiatrie, Rhumatologie, et Urologie-Néphrologie. Parmi elles, Tête et cou se positionne comme la spécialité la moins bien notée, avec une moyenne de 6.93.

La Médecine Générale présente une moyenne de 8.38, légèrement au-dessus de la performance globale du modèle (8.16).

Nous avons ensuite cherché à mettre en évidence une différence globale de performances entre les différentes spécialités. Pour cela, nous avons effectué un **test de Kruskal-Wallis** qui a confirmé l'existence d'une différence significative entre les spécialités dans notre étude ($p < 0.05$).

Nous avons ensuite procédé à une analyse post-hoc par un **test de Dunn** pour comparer les groupes et établir s'il existait une différence significative entre les spécialités. Les résultats sont présentés ci-dessous sous la forme d'une carte de chaleur.

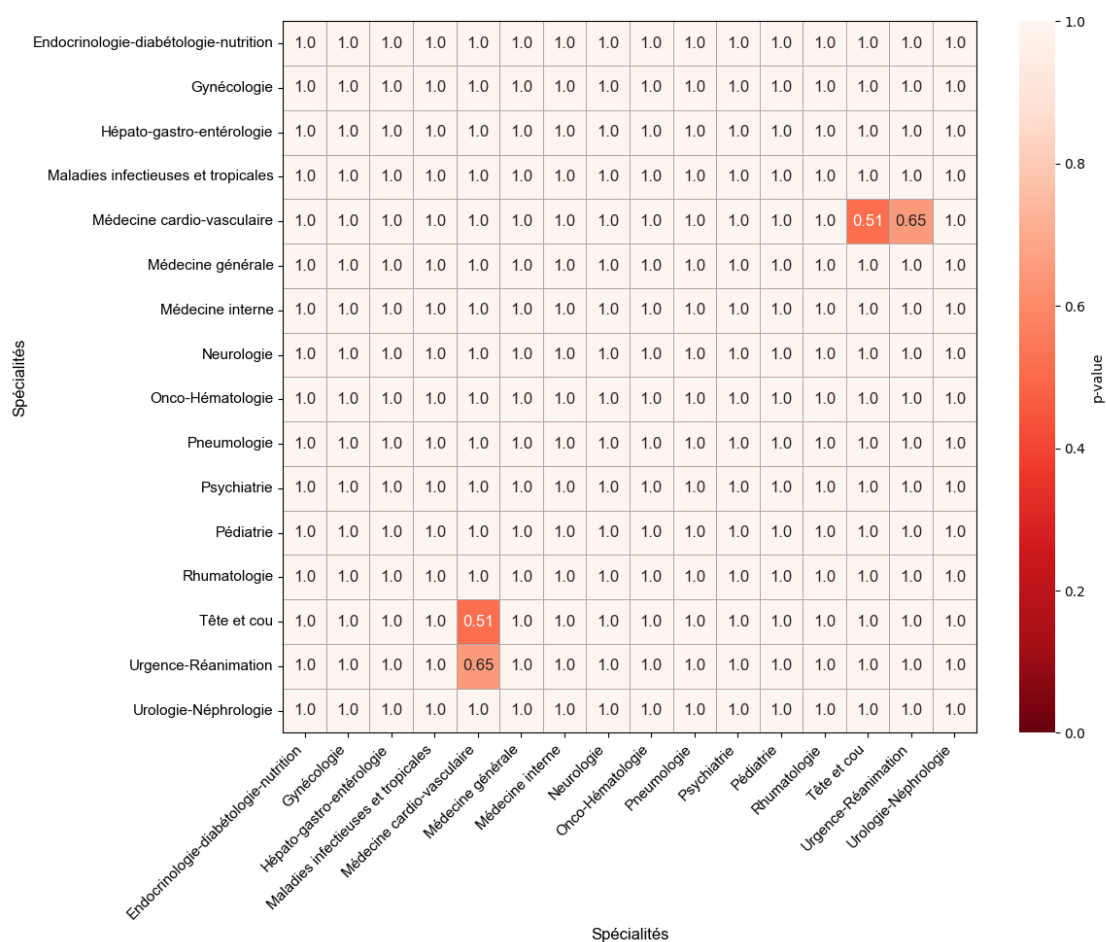


Figure 41 - Comparaison en sous-groupes des notes de GPT en français sous forme de carte de chaleur.

Nous pouvons observer des écarts de performance entre la Médecine cardiovasculaire et les spécialités Urgence-Réanimation et Tête et cou. **Néanmoins, ces différences ne sont pas statistiquement significatives ($p > 0.05$), et ne permettent pas de conclure.**

b) GPT-4 en anglais

Nous comparons dans le graphique suivant les moyennes obtenues par GPT, en anglais, selon la spécialité.

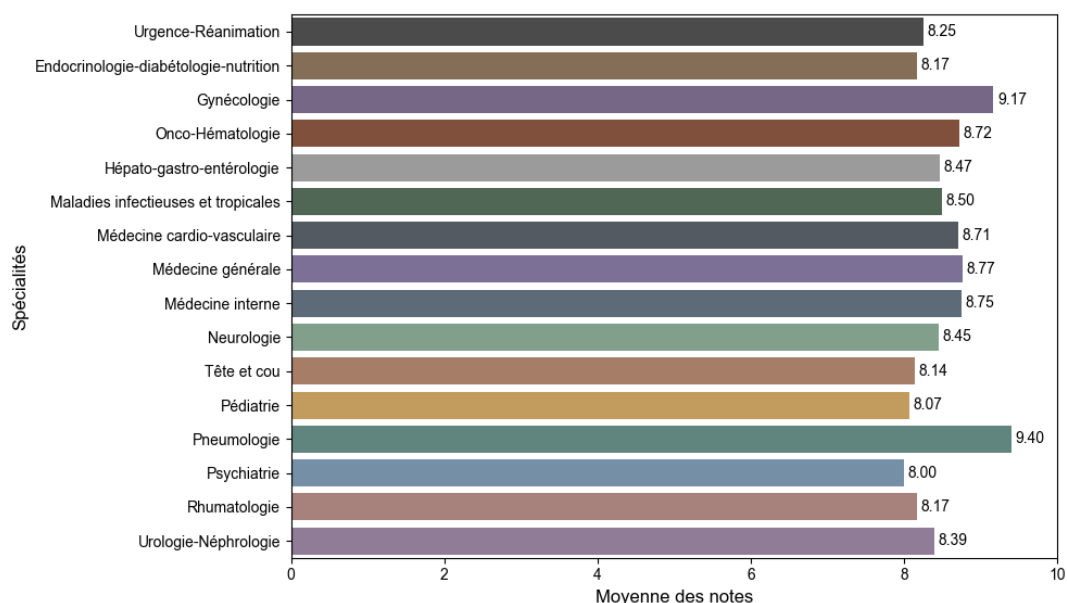


Figure 42 - Moyennes obtenues par GPT en anglais selon la spécialité

Dans ce graphique, nous constatons que la moyenne la plus faible est pour la Psychiatrie avec 8.0. La moyenne la plus élevée est pour la Pneumologie avec 9.40. En comparaison avec les résultats en français, les moyennes sont plus homogènes, ce qui est en accord avec les observations faites dans les précédentes analyses. Il n'existe aucune spécialité ayant eu une moyenne en dessous de 8.0. Nous pouvons noter que certaines spécialités semblent se détacher des autres avec des moyennes plus hautes, comme la Pneumologie ou la Gynécologie.

La Médecine Générale présente une moyenne de 8.77, légèrement supérieure à la moyenne générale du modèle (8.55).

Là encore, nous avons cherché à mettre en évidence une différence globale de performance entre les différentes spécialités. Pour cela, nous avons effectué un test de Kruskal-Wallis qui n'a **pas montré de différence significative entre les spécialités**. Nous avons donc interrompu l'analyse à ce stade.

2. Llama-3

a) Llama-3 en français

Nous comparons dans le graphique suivant les moyennes obtenues par Llama, en français, selon la spécialité.

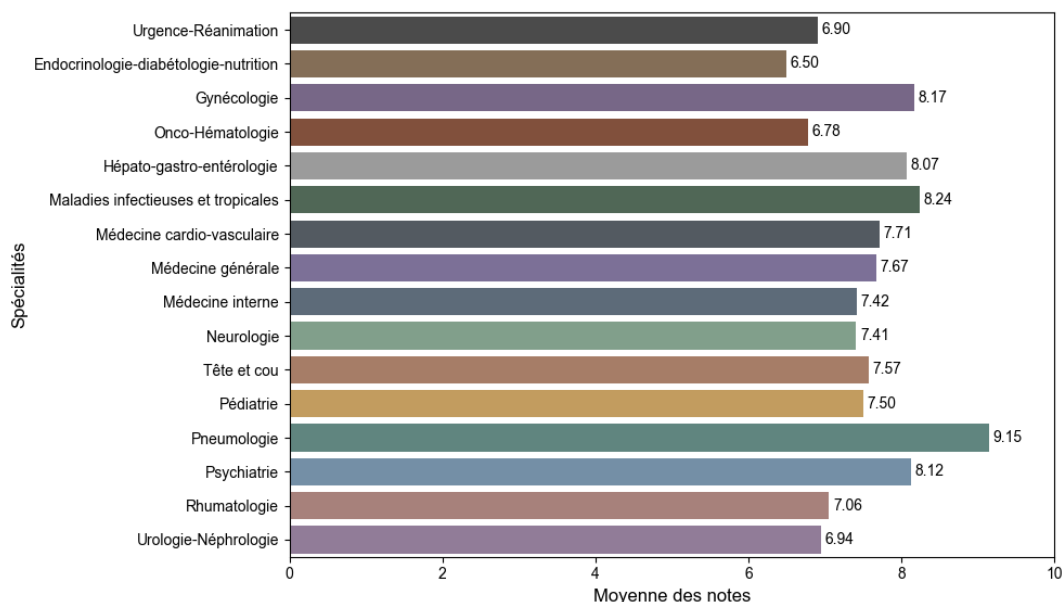


Figure 43 - Moyennes obtenues par Llama en français selon la spécialité

Nous pouvons constater que pour Llama-3, en français, la moyenne la plus faible est obtenue par l'Endocrinologie-Diabétologie-Nutrition à 6.50. La moyenne la plus élevée est, comme pour GPT en anglais, la Pneumologie avec 9.15. Tout comme GPT en français, on retrouve des moyennes hétérogènes, avec certaines spécialités descendant en dessous de la barre des 7.0, comme l'Urgence-Réanimation, l'Endocrinologie-Diabétologie-Nutrition, l'Onco-Hématologie, ou encore l'Urologie-Néphrologie.

La Médecine Générale présente une moyenne de 7.67, semblable à la moyenne générale du modèle (7.64).

Nous avons cherché à mettre en évidence une différence globale de performance entre les différentes spécialités. Nous avons réalisé un test de Kruskal-Wallis qui a montré l'existence d'une différence significative de performance entre les spécialités ($p < 0.05$). Pour cette raison, nous avons procédé à une analyse post-hoc par un test de Dunn pour comparer les groupes et établir s'il existait une différence significative entre les spécialités. Les résultats sont présentés ci-dessous sous la forme d'une carte de chaleur.

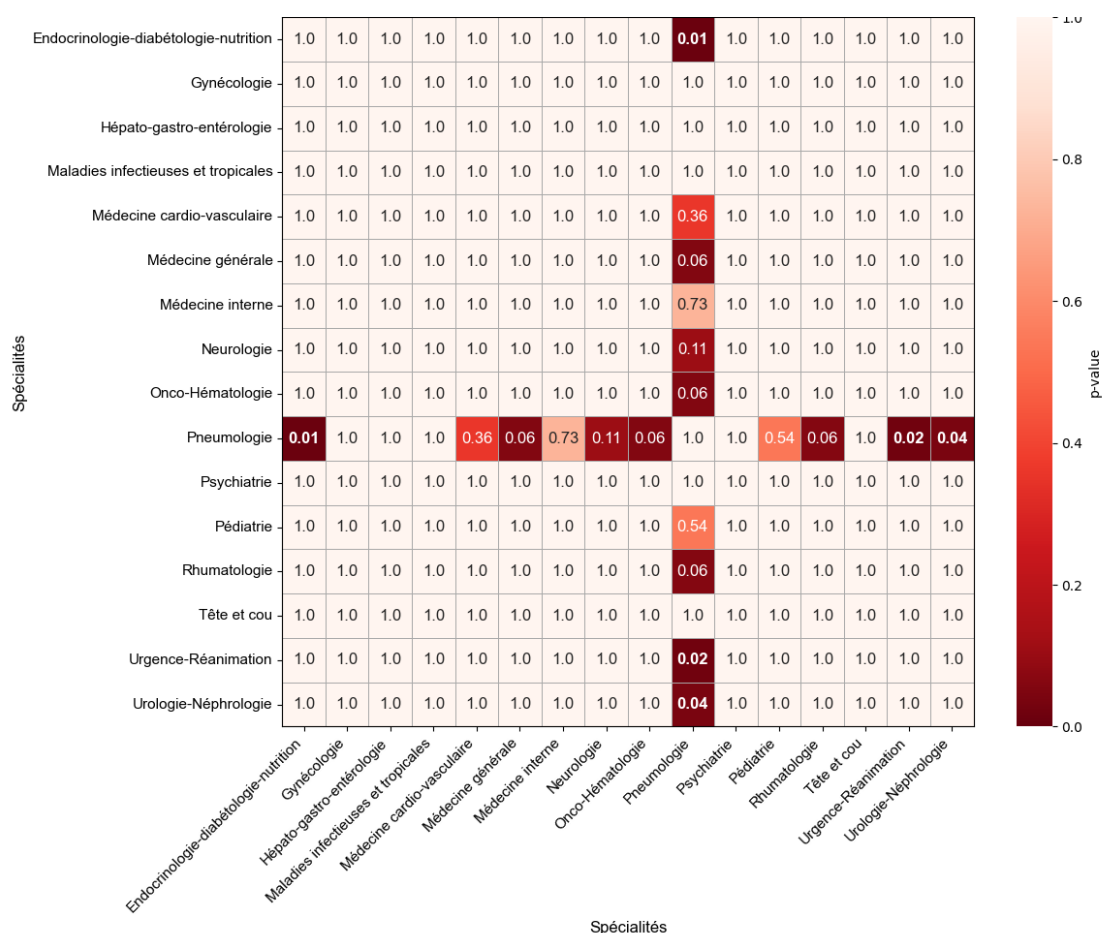


Figure 44 - Comparaison en sous-groupes des notes de Llama en français sous forme de carte de chaleur.

En observant cette carte, nous pouvons constater que la spécialité principalement responsable des différences est la Pneumologie. On peut observer que le seuil de significativité ($p < 0.05$) est atteint avec ce groupe en comparaison avec l'Endocrinologie-Diabétologie-Nutrition, l'Urgence-Réanimation, et l'Urologie-Néphrologie. **On pourra aussi faire remarquer qu'on est proche du seuil de significativité en comparaison avec la Médecine Générale et l'Onco-Hématologie (0.06).**

Nous pouvons conclure que, en français, dans le cadre de notre étude, Llama-3 est significativement plus performant avec la Pneumologie par rapport à l'Endocrinologie-Diabétologie-Nutrition, l'Urgence-Réanimation, et l'Urologie-Néphrologie.

b) Llama-3 en anglais

Nous comparons dans le graphique ci-dessous les moyennes obtenues par Llama, en anglais, selon la spécialité.

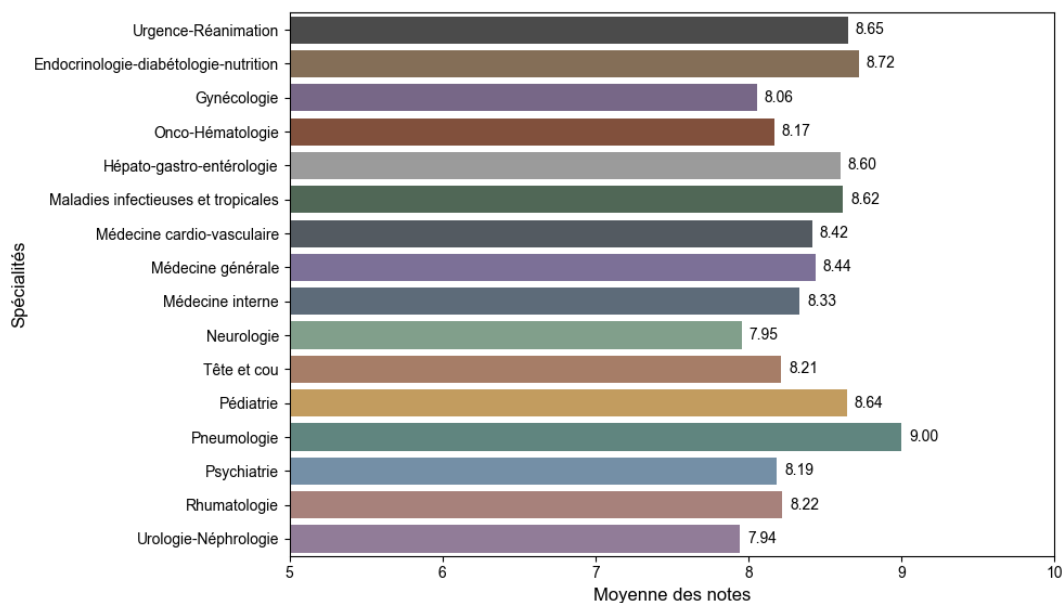


Figure 45 - Moyennes obtenues par Llama en anglais selon la spécialité

Dans ce graphique, nous constatons que la moyenne la plus faible est observée pour l'Urologie-Néphrologie avec 7.94, tandis que la moyenne la plus élevée revient à la Pneumologie avec 9.00. En comparaison avec les résultats en français, les moyennes sont moins hétérogènes. Il existe certaines spécialités ayant eu une moyenne en dessous de 8.0 notamment l'Urologie-Néphrologie et la Neurologie.

La Médecine Générale présente une moyenne de 8.44, très proche de la moyenne générale du modèle (8.55).

Nous avons effectué un test de Kruskal-Wallis, qui n'a pas montré de différence significative entre les spécialités. Nous avons donc interrompu l'analyse à ce stade.

3. Conclusions de l'analyse par spécialité

Au vu des analyses conduites entre les différents modèles, nous pouvons parvenir à certaines conclusions **statistiquement significatives** :

1. En anglais, nous n'avons pu mettre en évidence de différence statistiquement significative entre les spécialités pour GPT-4 et Llama-3.
2. En français, nous n'avons pas pu mettre en évidence de différence statistiquement significative entre les spécialités pour GPT-4. Néanmoins, pour Llama-3, nous avons retrouvé une différence significative pour Pneumologie par rapport à l'Endocrinologie-Diabétologie-Nutrition, l'Urgence-Réanimation et l'Urologie-Néphrologie.

3. En anglais et en français, quel que soit le modèle, il n'y a pas de différence significative de notes entre la Médecine Générale et l'ensemble des autres spécialités.

G. Comparabilité inter-auteurs

La définition d'une grille de notation commune, en amont de l'analyse des réponses, a permis d'assurer une comparabilité des notes entre les auteurs de la thèse. Sa conception est présentée dans la section méthodologie.

Afin d'objectiver cette comparabilité, nous avons calculé le coefficient de Spearman sur l'intégralité des notes. Le choix de ce test, en comparaison avec le coefficient de corrélation de Pearson, réside dans la distribution non normale des notes, et l'échelle ordinale utilisée (-1 à 10 sans valeurs décimales).

Nous obtenons un coefficient de Spearman à 0.71 ($p < 0.05$), ce qui témoigne d'une forte monotonie dans la distribution des notes.

H. Éléments d'analyse qualitative

Nous présentons quelques éléments d'analyse qualitative relevés par les auteurs lors de leur lecture des réponses des modèles. Les différents points relevés ci-dessous quant aux modes de réponses des modèles n'étaient pas évalués par la grille de cotation.

1. GPT-4

a) Schémas récurrents

Nous avons pu voir que plusieurs réponses suivent une structure récurrente. Par exemple, on observe fréquemment la configuration du type : **["symptômes A, B, C, ..." => "sont typiques/caractéristiques/classiques de" => "diagnostic final"]**, en particulier en français. Dans ce schéma de réponse, la valeur de chaque symptôme ou signe semble souvent équivalente. Un autre schéma de réponse récurrent est celui où GPT pose le diagnostic puis explicite différents éléments du cas pour le justifier avec une structure du type : **["diagnostic à priori" => "éléments concordants" => "diagnostic final"]**. Le modèle s'appuie fortement sur les données d'entrée, quitte à les citer de manière non pertinente ou à répéter les mêmes expressions plutôt que de reformuler. Il semble qu'il soit difficile pour le modèle de supposer des choses non explicitement dites en entrée.

b) Éléments non attendus.

Nous observons un diagnostic plus poussé que celui de référence dans environ **9% des cas** (19 en anglais, 15 en français).

Ces diagnostics plus détaillés qu'attendu ne semblent pas apporter de plus-value dans 58% des cas en anglais et 66% des cas en français. GPT-4 a tendance à faire des propositions de conduite à tenir à la suite de son raisonnement. Ces propositions sont présentes **dans 55% des réponses**. Elles sont d'ordre diagnostique ou thérapeutique (ex : J44 en français). Ces propositions sont parfois d'intérêt limité (ex : S44, J22) voir fausses (ex : S39, S40, S87). Les propositions semblent meilleures en anglais qu'en français (ex : J49, J22, J33). L'apparition de ces éléments ne semble pas répondre à une règle précise.

Contrairement à BioMistral, GPT-4 hallucine très peu. En ce qui concerne le diagnostic final on note pour le cas S88 en français le diagnostic de « panthénolite » qui ne semble correspondre à rien. Le cas concerné (*un cas de panaris*) n'était pas particulièrement difficile et GPT en anglais donne la bonne réponse. Si l'on considère pour GPT-4 ce seul cas d'hallucination, la fréquence de ces dernières pour le diagnostic final dans notre travail est de **1/360, soit 0.3% des réponses environ**.

2. Llama3

Les observations faites avec GPT-4 concernant les motifs récurrents, la tendance à se fonder beaucoup sur le texte d'entrée, la difficulté à reformuler et à faire des suppositions se retrouvent avec Llama3.

a) Diagnostics différentiels

Les réponses en anglais étaient souvent de meilleure qualité, nous pouvons le voir par exemple pour les cas S68, S79 ou S84. Il en va de même concernant les diagnostics différentiels, souvent meilleurs en anglais.

- S56 : « une maladie respiratoire chronique non infectieuse » en français contre « chronic obstructive pulmonary disease (COPD) or lung cancer » en anglais.
- S58 : « certaines causes infectieuses plus graves » en français contre « autres causes de douleur thoracique ou de fièvre » en « pharyngitis or pneumonia » en anglais.
- S69 : « causes vertébrales », « cause infectieuse » en français contre « pancreatic cancer or musculoskeletal disorders » en anglais
- S71 : aucun diagnostic différentiel proposé contre « pneumonia or cardiac disease » en anglais.

Nous avons constaté une tendance à mettre dans la même phrase des idées vraies et fausses. Par exemple pour le cas S4 : « *la paralysie faciale droite et l'abolition du réflexe nauséeux sont des signes de lésion du nerf facial* ».

b) Éléments non attendus

Comme pour GPT-4, Llama3 a tendance à fournir des diagnostics parfois plus détaillés que prévu, et cela se produit plus fréquemment que pour GPT-4, avec **18% des réponses concernées** (40 réponses en anglais et 26 en français). A l'instar de GPT-4, ces précisions supplémentaires n'apportent généralement aucune plus-value. En effet, nous n'avons pas retrouvé d'intérêt à ces précisions **pour 62% des réponses en anglais et 73% en français**. Les hallucinations sont également rares, du moins pour le diagnostic. On retrouve une seule assertion pouvant être considérée comme telle pour le cas J68 en français où Llama propose le diagnostic de « Panthère ».

Le taux d'hallucination est donc similaire à celui de GPT-4, **soit 1 pour 360 (environ 0.3 %)**.

3. BioMistral

La qualité des réponses de BioMistral est très limitée.

a) Éléments non attendus

BioMistral a tendance à inventer des éléments cliniques non présents dans le cas d'entrée. Les exemples en sont nombreux. Dans le cas J4, le modèle invente une prise médicamenteuse non mentionnée dans le cas clinique. Dans les cas J15 et J82, le modèle interprète un bilan sanguin qui n'existe pas. Dans le cas J34, il fait la même chose avec un électrocardiogramme.

Pour les diagnostics finals, on retrouve aussi des inventions de la part du modèle. Dans le cas S48 avec la « rétinopathie cytotatique-induite » ou encore le cas J73 avec la « testiculite aiguë », et la « prostatite bénigne » pour le cas S61. Si l'on considère ces éléments comme des hallucinations, et en prenant en compte seulement les cas pour lesquels BioMistral a fourni une réponse, on observe un **taux d'hallucination pour le diagnostic final de 3 pour 290, soit environ 1% des réponses**.

b) Schémas récurrents

Un schéma récurrent dans les réponses est celui du texte en boucle. Pour de nombreux cas, BioMistral termine sa réponse avec une phrase ou un mot répété en boucle. Ce genre de réponse aberrante se retrouve **dans 19% des cas en français et 7% en anglais, soit 13% de l'ensemble des réponses**. Dans de très nombreux cas, les réponses de BioMistral ne sont qu'une réécriture plus ou moins modifiée du cas clinique initial. Dans ces réponses, le modèle conclut parfois à un diagnostic mais pas toujours. Nous pouvons citer comme exemple le cas J72 en français.

c) Différences entre l'anglais et le français

On retrouve avec BioMistral des résultats qui semblent meilleurs en anglais qu'en français. BioMistral en français mélange souvent les deux langues dans ses réponses.

d) Format des réponses

Lorsque BioMistral formule une réponse de qualité, le modèle démontre une capacité à aller à l'essentiel, en fournissant une réponse plus concise.

I. Résumé des résultats

1. BioMistral est le seul modèle à ne pas avoir répondu à tous les cas cliniques. Il hallucine plus souvent et est plus instable.
2. En anglais, GPT-4 et Llama semblent présenter de très bonnes performances, comparables et homogènes. BioMistral, de son côté, est en dessous en termes de performances et une plus grande hétérogénéité des réponses.
3. En français, GPT-4 et Llama présentent de bonnes performances, avec des moyennes et écarts types sensiblement différents. BioMistral reste à nouveau très en dessous en termes de performances.
4. Les items de notation sur lesquels on observait le plus de différences étaient l'Hypothèse et l'Expression.
5. GPT-4 et Llama-3 obtient de meilleures notes en anglais par rapport au français.
6. En anglais, il n'existe pas de différence significative de notes entre GPT-4 et Llama-3.
7. En français, GPT-4 obtient de meilleures notes comparativement à Llama-3.
8. Il existe des situations où GPT-4 et Llama-3 produisent de mauvais résultats, quelle que soit la langue.
9. En anglais, nous n'avons pu mettre en évidence de différence statistiquement significative entre les spécialités pour GPT-4 et Llama-3.
10. En français, nous avons retrouvé une différence significative pour Pneumologie par rapport à l'Endocrinologie-Diabétologie-Nutrition, l'Urgence-Réanimation et l'Urologie-Néphrologie pour Llama-3.
11. En anglais et en français, quel que soit le modèle, il n'y a pas de différence significative de notes entre la Médecine Générale et l'ensemble des autres spécialités.
12. Les notes entre les auteurs étaient monotones avec un coefficient de Sperman à 0.71.
13. Llama-3 et GPT-4 présentaient des patterns de réponses récurrents. Les réponses de BioMistral étaient très variables.

Discussion

I. Enquête auprès des médecins généralistes

Les résultats de notre questionnaire ont permis d'obtenir un premier aperçu de la perception, des attentes, ainsi que des enjeux pour les médecins généralistes sur l'intégration des intelligences artificielles génératives dans la pratique médicale. Nous mettons dans le texte suivant plusieurs messages clés méritant d'être approfondis.

A. Discussion des résultats

1. Utilisation actuelle de l'intelligence artificielle

Dans notre étude, 29% des médecins rapportent n'avoir jamais été en contact avec un algorithme d'intelligence artificielle. Cette information est intéressante car elle montre la difficulté à percevoir ces technologies pourtant omniprésentes dans notre quotidien (smartphone, ordinateur). Une autre étude observationnelle faite chez les médecins généralistes en 2024, avec de plus grands échantillons, rapporte un taux identique (61).

Notre étude présente l'avantage qu'elle questionne les participants sur l'utilisation des intelligences artificielles **génératives** spécifiquement. S'il existe une réelle difficulté à percevoir l'IA dite « générale », ce n'est pas le cas pour les IA génératives, dont la production de texte ou d'image est souvent bien trop rapide pour être celle d'un être humain, et nécessite le plus souvent une démarche active de la part de l'utilisateur. Celui-ci est donc immédiatement au fait qu'un algorithme produit la réponse à la requête envoyée. On constate dans notre échantillon que 49% des médecins ont déjà été en contact avec ce type de technologie. Pour seule comparaison, l'enquête « Les Français et les IA génératives » par l'IFOP en 2024 rapporte que 78% de la population a déjà entendu parler des IA génératives (14).

Au sein de notre échantillon, environ 25% des médecins généralistes rapportent utiliser une intelligence artificielle générative occasionnellement (plus d'une fois par mois) à quotidiennement. Si ce chiffre est en-dessous de la population générale (32%) il est proche de ceux constatés pour les individus de 35 ans et plus (22%), tranche d'âge plus proche de notre échantillon (14).

Nous constatons que 26% des médecins généralistes indiquent avoir déjà utilisé une intelligence artificielle générative dans le cadre de leur travail. Il est difficile de comparer ce chiffre à la population générale, mais le baromètre de l'IFOP en 2024 rapporte que sur l'échantillon de population utilisant l'IA générative, 43% en font un usage professionnel.

Si on observe les autres métiers, comme l'enseignement, une enquête réalisée en 2023 auprès d'enseignants et étudiants d'universités françaises montrait que 65% des enseignants n'utilisaient pas les IA génératives, soit un taux d'utilisation de 35% (62). Ces chiffres sont légèrement supérieurs à ceux constatés chez les médecins généralistes. **Bien que des données plus précises soient nécessaires pour l'affirmer, nous pouvons constater dans notre étude que l'utilisation de l'IA générative en pratique courante pour les médecins généralistes est faible.**

2. Applications potentielles de l'intelligence artificielle générative

Dans notre étude, les médecins généralistes identifient principalement l'intelligence artificielle générative comme un outil complémentaire potentiellement utile surtout pour la **réalisation de tâches administratives et l'analyse de données**. Bien que ces applications soient effectivement parmi celles attendues et déjà explorées dans la littérature, l'influence de l'IA sur la question diagnostique ou la décision thérapeutique est moins souvent mentionnée dans les réponses. Pourtant, de nombreuses études et revues de la littérature mettent en avant le potentiel de ces technologies à contribuer à une médecine de précision, adaptée à chaque patient, grâce à la collecte et à l'analyse d'un grand nombre de données. **Cela peut indiquer une perception encore limitée parmi les médecins quant aux capacités réelles de ce type de technologie** (63). Un travail réalisé en 2024 a montré que, sur une grande quantité d'études, l'IA générative était principalement utilisée pour informer plutôt que pour assister ou automatiser certaines fonctions (64).

Ces réponses peuvent aussi être le témoin d'une offre de solutions en intelligences artificielles génératives limitée pour le moment. Il est probable qu'il existe un décalage entre les besoins réels des médecins et les offres disponibles. Bien qu'il existe de plus en plus de solutions disponibles sur le marché, elles sont encore loin de couvrir l'ensemble du travail (*workflow*) des médecins.

C'est pourquoi, au-delà des évaluations faites sur la performance de l'IA générative (ce que nous faisons dans la deuxième partie de notre travail), d'autres études sur son applicabilité dans les différents aspects du travail des médecins serait souhaitable.

Rappelons que l'utilisation professionnelle de l'IA dans notre échantillon était faible. Ceci indique que les utilisations mentionnées de l'IA sont des utilisations « attendues » et non pas des utilisations actuelles.

Parmi les freins identifiés dans notre étude, **les préoccupations éthiques et la confidentialité des données sont le plus souvent mentionnées**. Ces enjeux sont également soulignés dans des études précédentes (8,61).

Les médecins généralistes semblent donc eux aussi particulièrement préoccupés par ces questions, ce qui rend d'autant plus **crucial de proposer des solutions souveraines dans un cadre éthique et légal pour la gestion des données sensibles.**

Contrairement à certaines tendances observées dans d'autres métiers, peu de médecins perçoivent l'IA générative comme un risque par rapport à leur emploi. Cet aspect est intéressant car si plusieurs travaux déjà cités évoquent une crainte vis-à-vis de la perte de compétence ou la perte d'autonomie induite par ces outils, la perte de travail en lui-même est finalement peu présente. Cette impression est renforcée par certaines réponses à la question ouverte de notre questionnaire, qui évoquent cette crainte de « perdre la main » sur la décision médicale.

Par ailleurs, la forte majorité d'entre eux pensent qu'elle ne sera utile que dans moins de 25 % de leurs consultations. Ce constat est surprenant lorsqu'on regarde les réponses à la question de l'impact de l'IA sur leur pratique qui étaient majoritairement supérieures à 5/10. Quand on sait que cette technologie peut être intégrée à plusieurs étapes de la pratique, cette donnée peut suggérer un manque de perception de son utilité ou que celle-ci ne trouvera sa place que dans certains aspects spécifiques de la pratique médicale.

Le coût est rarement mentionné dans les réponses, contrairement à ce qui a pu être observé dans les études sur l'IA dite « générale » (47,61). **En effet, bien que certains produits dotés d'intelligence artificielle puissent être très onéreux, l'IA générative présente l'avantage d'être relativement accessible.** A l'heure de la rédaction de cette thèse, un modèle de ChatGPT est accessible gratuitement, tandis que la version professionnelle est à 20 euros par mois.

Les médecins généralistes sont globalement favorables à l'intégration de l'IA générative dans leur pratique, tandis d'un tiers d'entre eux restent neutres. Près de 90% d'entre eux ont montré un intérêt à suivre une formation selon le contenu, tandis qu'environ 20% identifient le manque de connaissance de l'outil comme un frein à son utilisation.

Ces résultats sont cohérents avec ceux présents dans la littérature (61). **Nous mettons en évidence l'importance de développer une offre de formation diversifiée pour les médecins, en mettant particulièrement en avant les solutions en ligne.**

B. Limites de l'étude

Il est important de garder en tête le faible taux de réponses à notre questionnaire. Ce faible taux de réponse rend la représentativité de notre échantillon discutable. Par ailleurs, en raison de l'actualité du sujet, il existe sans doute un biais de recrutement.

Certaines personnes intéressées par le sujet ou ayant un avis très tranché sur la question y sont peut-être surreprésentées. Les données socioéconomiques sont malgré tout assez similaires à ce que l'on observe dans des études semblables (61) et à l'échelle de la région, notamment sur la proportion femme/homme compte tenu de l'âge (dans notre travail 65% de femmes chez les répondants de moins de 40 ans, pour 72% chez les moins de 35 ans dans les PDL en 2021), concernant les modes d'exercice (70% d'exercice en MSP ou cabinet de groupe en PDL en 2021, 82% dans notre étude) et le nombre de remplaçants (14.9% dans notre travail, pour 16% dans les PDL en 2021) (65).

II. Évaluation des modèles de langue

A. Discussion des résultats

Pour rappel, les hypothèses de travail, reposant sur la base des données de la littérature évoquées plus haut, étaient les suivantes :

- Les modèles les plus importants en termes de corpus d'entraînement et de nombre de paramètres (GPT-4 et Llama3) sont probablement plus performants que les modèles plus petits.
- Les corpus d'entraînement des modèles de fondations étant le plus souvent en langue anglaise (22), l'hypothèse d'une plus grande performance des modèles en anglais par rapport au français a été proposée.
- La faible quantité de données propres à la médecine générale dans la littérature scientifique et les spécificités de cette dernière peuvent laisser supposer que les grands modèles de langue seraient moins performants pour cette discipline que pour les autres.

1. Différences entre modèles

Le modèle le plus petit dans notre étude, BioMistral, présentait un niveau de performance largement inférieur à ce qui était observé avec GPT-4 et Llama3, au regard du critère d'évaluation principal. BioMistral, le seul modèle ajusté sur le biomédical testé dans notre étude s'est donc avéré beaucoup moins bon que deux modèles généralistes. Ces résultats sont contradictoires avec l'état actuel de la science, qui suggère que les modèles ajustés peuvent avoir des performances meilleures dans des tâches spécifiques (66,67). Nous ne sommes néanmoins pas surpris du résultat au vu de la taille de BioMistral (7 milliards de paramètres), et il serait intéressant de faire d'autres études avec des modèles plus développés.

Notre analyse a aussi montré qu'il existait une différence statistiquement significative de performance dans la résolution de cas cliniques via notre grille de notation entre GPT et Llama, mais seulement en français. Il est possible que cette différence, bien que statistiquement significative, ne soit pas pertinente en pratique, car restant limitée. Étant donné l'absence de certitude quant au nombre de paramètres exacts de GPT-4, il est difficile de conclure sur l'influence exacte de ce nombre sur la performance diagnostique, mais il est très probable qu'il explique les différences observées.

2. Question de la langue

L'hypothèse de meilleures réponses en anglais a également semblé se confirmer dans notre travail. Notre analyse a montré une différence statistiquement significative entre les notes avec de meilleures réponses en anglais qu'en français, pour GPT-4 comme pour Llama3. Les notes en anglais sont non seulement plus élevées, mais semblent aussi moins dispersées et plus homogènes. Si cette tendance vis à vis de l'anglais n'a pas fait l'objet de tests statistiques pour BioMistral, elle a aussi été observée avec ce modèle. Ce constat renvoie aussi à l'importance du prompting, qui devrait donc prendre en compte la langue utilisée en entrée.

Notons que lorsque l'on parle de cas en anglais, il s'agit en fait de cas écrits au départ en français qui ont été traduits par les auteurs en anglais. La portée de ces résultats doit donc être nuancée par le fait que, d'une part les cas n'ont pas été écrits au départ en anglais, d'autre part que les réponses ont été produites et analysées en anglais. De plus, nous n'avons pas évalué la capacité des modèles à répondre directement en français. Les réponses obtenues auraient peut-être été différentes. On peut raisonnablement conclure malgré tout, qu'une traduction du texte d'entrée est suffisante pour obtenir des résultats de bon niveau, et même meilleurs qu'avec le texte en français. D'autres protocoles pourraient être envisagés pour explorer plus avant cette question. Il pourrait être intéressant par exemple de comparer les notes pour des cas écrits en français puis traduits en anglais, à des cas écrit initialement en anglais. Il pourrait aussi être intéressant de demander aux modèles sur une requête en anglais de répondre en français d'emblée.

La question d'entraîner des LLM avec plus de données en français reste entière. Des modèles entraînés avec plus de textes en français seraient peut-être meilleurs sur des cas en français. La principale conclusion que nous pouvons faire ici est qu'il semble que les modèles testés fassent de meilleures réponses en anglais.

3. Performance selon la discipline

Pour GPT-4 et Llama3 en anglais, l'analyse statistique ne permet pas de conclure à l'existence de différences de notes entre les spécialités. Il en va de même pour GPT-4 en français. Llama-3 en français quant à lui, est le seul à présenter des différences statistiquement significatives entre les disciplines. Les notes étaient meilleures en Pneumologie par rapport à l'Endocrinologie-Diabétologie-Nutrition, l'Urgence-Réanimation, et l'Urologie-Néphrologie. Ce résultat n'était pas attendu. La principale hypothèse que nous pouvons avancer pour l'expliquer est un niveau de difficulté pour les cas de pneumologie relativement faible, idée renforcée par une tendance similaire chez GPT, qui avait aussi une moyenne plus importante en pneumologie.

Il est probable que cette différence soit liée à la nature des cas plus qu'à une spécificité des modèles ou des spécialités. Nous reviendrons plus loin sur les limites concernant les cas utilisés.

En ce qui concerne la médecine générale, en anglais et en français, pour tous les modèles, aucune différence statistiquement significative de note n'a pu être mise en évidence vis à vis des autres spécialités. On observe néanmoins, et plutôt en contradiction avec notre hypothèse de travail, des moyennes relativement hautes pour la médecine générale, en tout cas plus hautes que la moyenne globale des modèles. La moyenne en médecine générale avec GPT était de 8.38 en français et de 8.77 en anglais, pour des performances globales du modèle dans ces langues de respectivement 8.16, et 8,55. Ces différences ne sont pas statistiquement significatives. Bien que le nombre de cas de médecine générale ait été plus grand que pour les autres disciplines, ce nombre était sans doute insuffisant pour conclure spécifiquement sur cette question. Il ne s'agissait pas de l'objectif principal de l'étude, qui n'avait pas été dimensionnée pour cette analyse. Une méthodologie similaire avec un plus grand nombre de cas de médecine générale testés contre les cas des autres disciplines serait souhaitable pour répondre à cette question. L'hypothèse selon laquelle les modèles testés seraient aussi bon en médecine générale que dans une autre discipline, du moins sur le plan diagnostique, ne peut être rejetée sur la base de nos résultats. La nature de nos cas, limite que nous détaillons plus bas, peut aussi expliquer cette absence de différence.

4. Éléments d'analyse qualitative

La méthodologie de ce travail n'a pas été conçue dans le sens d'une analyse qualitative des réponses de l'IA. Les éléments suivants ne sauraient donc constituer des conclusions définitives quant aux fonctionnements des modèles testés.

La lecture des réponses aux cas cliniques, a néanmoins permis aux auteurs d'appréhender cette manière de fonctionner, permettant une réflexion au-delà des seules notes.

Les réponses de l'IA, y compris pour les modèles les plus performants, laissent souvent entrevoir leur mode de fonctionnement stochastique. Il existe une forte tendance à lister des évidences, à faire des généralités et à proposer des listes d'arguments peu hiérarchisés. On retrouve de manière récurrente, chez GPT comme chez Llama, certains des patterns de réponse, avec en particulier des structures du type **diagnostic à priori => listing d'arguments issus du cas clinique => diagnostic final**. Les modèles répondent parfois comme s'ils voulaient « faire rentrer » leurs arguments pour aboutir à un diagnostic posé d'emblée quitte parfois à prendre certaines libertés. Il ne semble pas exister de véritable réflexion sous-jacente qui avancerait étape par étape, similaire à ce que produit l'humain.

On pourrait dire que les modèles présentent plutôt un raisonnement « de façade », souvent impressionnant en termes de qualité rédactionnelle, mais dont l'analyse plus poussée montre les failles. On retrouve un manque de nuances, des généralités de peu d'intérêt, voir des affirmations que les précautions langagières du modèle ne rendent pas tout à fait fausses, et pourtant sujettes à débat.

Sur le diagnostic seul, il est intéressant de voir comme l'IA parvient à bien se positionner quant à la « profondeur » du diagnostic, c'est à dire sur l'échelle allant de la simple traduction des symptômes en termes médicaux au diagnostic étiologique ultraprécis. Les diagnostics attendus, qui constituaient la référence, étaient variables selon les cas. Ils pouvaient être syndromique, étiologique ou purement paracliniques. La demande initiale était assez générale, il était demandé au modèle de porter un diagnostic, sans autres précisions. Sur ce point, l'IA se positionne très bien d'une manière générale. On pourrait dire que les modèles saisissent assez bien ce qui leur est demandé. Néanmoins, lorsqu'il dépasse la demande initiale, qu'il s'agisse de diagnostics plus poussés ou de propositions de conduite à tenir (très fréquent chez GPT), ces propositions étaient le plus souvent de faible qualité, sans plus-value et parfois même fausses.

Ces constats nous laissent penser que l'utilisation des réponses de l'IA dans le cadre du diagnostic clinique nécessite une certaine expertise médicale, du moins, un recul suffisant pour nuancer certaines affirmations. Au regard de ces erreurs, l'utilisation de ces modèles, de cette manière, par des personnes sans culture médicale (patients eux-mêmes ou soignants non-médecins par exemple) nous paraît d'un intérêt limité, voir potentiellement dangereuse. Une utilisation similaire à la nôtre à des fins d'enseignement pourrait aussi être sujette à caution.

Enfin, notons que les hallucinations existent, y compris chez les meilleurs modèles, et que par là même ceci pose la question du niveau de risque acceptable dans leur utilisation. On observe un taux d'environ 0.3% de diagnostic « hallucinés » pour GPT-4 et Llama3, et un taux de diagnostic insuffisants ou faux dans 20 à 30% des réponses, ce qui serait considérable en pratique courante. Les bonnes notes obtenues par les modèles ne doivent pas occulter leurs limites, qui restent non négligeables à la lumière de nos impressions. Nous encourageons la réalisation de plus d'analyses qualitatives.

5. Analyse des réponses de l'IA :

Notre utilisation de quelques indicateurs chiffrés (*indice de SMOG*) a permis de mettre en évidence des réponses qui semblent plus complexes chez GPT et Llama que chez BioMistral. Ces données objectivent l'impression des auteurs sur ce point.

Le grand nombre de réponses aberrantes de BioMistral explique aussi en partie ces chiffres. Il existe des écarts types beaucoup plus grands pour BioMistral, suggérant une plus grande variabilité du modèle dans la longueur de ses réponses. La longueur des réponses n'est pas révélatrice de leur niveau de complexité. Néanmoins, il existe quelques cas pour BioMistral où le modèle démontre une certaine capacité à être plus concis et direct dans ses réponses, ce qui pourrait être souhaitable pour certaines applications réelles. Cette capacité à fournir des réponses plus concises et directes pourrait être un item de notation des réponses des modèles et un critère de qualité dans de futurs travaux.

B. Limites de l'étude

1. Base de cas clinique

Notre base de cas clinique souffre de plusieurs limites. La constitution des cas s'est faite très largement sur la base de référentiels de médecine. Si les auteurs se sont aussi attachés à produire des cas inventés et des cas réels issus de leur pratique, ceux-ci restent largement minoritaires. Des analyses plus poussées en fonction du type de cas n'ont donc pas été jugées pertinentes.

L'utilisation de cas inspirés de la littérature scientifique a produit beaucoup de cas « classiques », qu'il s'agisse des pathologies abordées ou de leur présentation clinique. La nature des référentiels utilisés, pour beaucoup des ouvrages destinés aux étudiants en médecine, a très certainement donné une forme plutôt « scolaire » à nos cas clinique. La méthodologie initiale était d'apporter suffisamment d'éléments dans le cas pour permettre sa résolution. Ainsi, les cas étaient généralement bien renseignés et ne comportaient que peu voire pas de données inutiles ou perturbatrices. Par ailleurs, le format « cas clinique », qui

résume une situation en synthétisant les informations essentielles, constitue une première forme de sélection de l'information, rendant sa résolution bien plus simple qu'une anamnèse brute, rapportée par un tiers ou directement par le patient. La collecte de ces éléments, incluant l'interrogatoire et l'examen physique, est partie intégrante du raisonnement du clinicien. Les choix des examens paracliniques sont également orientés par cette première étape. Ces différentes étapes cognitives, pourtant essentielles, n'ont donc pas été réalisées par l'intelligence artificielle.

Il en résulte des cas que les auteurs jugent d'une manière générale d'un faible niveau de difficulté, et surtout assez éloignés de la pratique réelle. On peut raisonnablement supposer que nos cas cliniques ne présenteraient pas de difficultés majeures à des étudiants préparant l'EDN, ou à des cliniciens expérimentés. Il serait donc hasardeux d'extrapoler ces résultats en affirmant que ces modèles sont de bons « diagnosticiens » d'autant plus que leurs performances, bien qu'excellentes, restent imparfaites.

Une piste d'amélioration, serait de compléter la base existante avec des cas plus difficiles, plus souvent inspirés de situations réelles, rédigés par des experts de la discipline concernée. Des cas comportant plus d'éléments « perturbateurs » ou « trompeurs », comme cela arrive souvent en pratique clinique, seraient également intéressants.

Une caractérisation précise de chaque cas sur le plan de la difficulté, mais aussi du niveau d'incertitude, serait probablement très utile à de futurs travaux. Un effort approfondi serait nécessaire pour définir et grader ces deux concepts. Nous nous contenterons pour l'instant de faire l'hypothèse que la base de cas cliniques utilisée dans notre étude présente un faible niveau de difficulté et d'incertitude, ce qui constitue une des limites de nos travaux.

2. Notation des réponses de l'IA

Notre grille de notation s'est avérée limitée pour plusieurs raisons.

La grille a été conçue sur la base des premières réponses disponibles de l'IA, à savoir celles obtenues avec GPT-4. Cette méthodologie a posé problème vis-à-vis de BioMistral, dont les réponses, de qualité nettement inférieure, ont parfois été difficiles à évaluer. L'échelle aurait gagné à avoir des notes inférieures pour pouvoir mieux sanctionner certaines réponses aberrantes.

L'une des règles initiales était que, lorsqu'une erreur pouvait relever de plusieurs items, elle ne devait être sanctionnée que sur un seul d'entre eux. Cette règle a probablement contribué à augmenter artificiellement les notes de BioMistral. Cela s'applique également aux items « Validité externe » et « Diagnostics différentiels ». Pour la validité externe, la note était maximale par défaut en l'absence de connaissances externes rapportées. De même, pour

les diagnostics différentiels, la note était maximale si aucun diagnostic différentiel n'était proposé pour un cas clinique ne l'exigeant pas. Cette règle avait été instaurée pour éviter de pénaliser excessivement les modèles. Cependant, dans le cas de BioMistral qui citait souvent des connaissances externes de manière inadaptée ou n'en mentionnait aucune, l'item était fréquemment surévalué.

Diagnostic final

La qualité des diagnostics a été variable. Il s'est avéré difficile de mettre en évidence certaines différences, notamment entre l'anglais et le français, avec la notation proposée. Une notation plus développée serait donc intéressante.

Validité interne et Logique

Ces deux concepts sont fortement liés, ce qui a souvent rendu difficile la distinction entre ce qui relevait de l'un ou de l'autre. La règle fixée au départ pour la validité interne était d'évaluer la capacité du modèle à mentionner les éléments utiles du cas. Cette règle a peut-être conduit à surévaluer les performances des modèles lorsque ceux-ci mentionnaient les éléments utiles sans les hiérarchiser correctement. En effet, les réponses du type « listing » d'arguments, sans pondération selon leurs valeurs sémiologiques étaient fréquentes. Les choix sémantiques des modèles rendaient valide l'interprétation des données du cas, sans pour autant être optimale.

La tendance des modèles à fortement s'appuyer sur le texte d'entrée et à le réutiliser aboutissait aussi à une bonne validité interne, pourtant parfois discutable. Cet item ne prenait pas en compte la capacité à « sélectionner » l'information. La présence d'assertions sans intérêt était alors parfois sanctionnée sur l'item logique, sans que cela n'ait été une règle pré définie. L'évaluation de la capacité à « mettre de côté » les données inutiles pourrait être une piste d'amélioration de la notation de la validité interne. La question de la logique dans les modèles de langue est complexe. Ceux-ci ne raisonnent pas, ils n'ont donc pas vraiment de logique propre, au-delà du sens du texte. Il peut être discutable d'évaluer cet item. Quoi qu'il en soit, si l'item logique devait être réutilisé, peut être faudrait-il s'en tenir à l'évaluation de la logique du texte. Sanctionner surtout la présence de contradictions dans le raisonnement.

Validité externe

La question de la validité externe n'a pas posé de problème majeur. Les connaissances évoquées par l'IA étant faciles à vérifier. Cet aspect de la note est important, compte tenu des risques en cas de fausse information rapportée par l'IA et devrait être évalué systématiquement dans les réponses concernant le raisonnement diagnostique. Une limite à notre manière d'évaluer cet aspect, était que nous n'avons pas pris en compte la pertinence

des connaissances rapportées. Il est arrivé, surtout avec BioMistral, que le modèle rapporte des connaissances externes vraies, mais non à propos. Cet aspect de pertinence devrait être pris en compte. Enfin, nous avons noté la validité externe comme satisfaisante (2 points /2) lorsqu'aucune connaissance externe n'était mentionnée par l'IA. Ainsi, pour nombre de cas, la note finale était peut-être artificiellement augmentée en marquant des points sur cet item sans qu'il y ait eu démonstration de la capacité du modèle à mentionner des connaissances externes.

Diagnostics différentiels

La capacité à générer des hypothèses cohérentes est essentielle à une réflexion de qualité et présente une grande utilité en pratique courante. Sur les diagnostics différentiels, les modèles étaient souvent meilleurs en anglais par rapport au français. En anglais, les propositions étaient souvent plus pertinentes, plus précises et explicites. La notation à deux degrés, selon qu'il y ait ou pas de diagnostics différentiels paraît donc insuffisante. Une échelle plus nuancée serait souhaitable, permettant de différencier les diagnostics différentiels de qualité (explicitement nommés, pertinents et hiérarchisés) par rapport à la simple mention de diagnostics potentiels sans précisions.

Expression :

Les différences observées entre anglais et français pourraient être attribuées à une différence de familiarité des auteurs entre les deux langues. Il est probable que les auteurs aient été moins « sévères » sur cet item avec les réponses en anglais. On peut distinguer deux types d'erreur en termes d'expression. Celles affectant le sens du texte et celles ne l'affectant pas. Les erreurs affectant le sens du texte devraient être sanctionnées d'emblée, il est en effet légitime d'attendre d'un modèle de langue qu'il produise des réponses compréhensibles.

Pour ce qui est des erreurs n'affectant pas le sens, on peut distinguer plusieurs sous-types. Les erreurs de vocabulaire (utilisation de mots peu appropriés, comme « diagnose » chez Llama), les erreurs de langue (BioMistral mélange très souvent anglais et français dans ses réponses), les erreurs relevant de faute de français (par exemple quand GPT-4 dit « infecté par le paludisme » où le paludisme n'est pas l'agent infectant). Certaines de ces erreurs pourraient être considérées comme plus graves que d'autres.

Nous proposons ci-dessous une nouvelle grille de notation :

Validité interne : analyse des données (à la fois clinique et paraclinique) - 5 pts : interprétation exhaustive et correcte de toutes les données utiles au diagnostic. Les éléments du cas sont hiérarchisés. - 4 pts : analyse non exhaustive des données, mais suffisante pour poser le diagnostic. Des éléments non pertinents sont relevés. - 3 pts : analyse non exhaustive des données insuffisante pour poser le diagnostic - 2 pts : analyse des données orientant vers un diagnostic erroné - 1 pt : paraphrase le dossier sans analyse des données - 0 pt : hallucination ou bogue ou aucune réponse générée
Validité externe : - 4 pts : connaissances médicales bonnes et utiles au bon diagnostic - 3 pts : connaissances médicales bonnes mais peu ou pas utile au diagnostic. - 2 pts : connaissances médicales partiellement correctes ou obsolètes - 1 pt : connaissances médicalement fausses - 0 pt : aucune réponse générée
Hypothèses : - 2 pts : capable d'évoquer des diagnostics différentiels dans des cas pertinents. Les diagnostics sont explicites et hiérarchisés - 1 pt : capable d'évoquer des diagnostics différentiels corrects mais insuffisant (pas suffisamment explicites ou non hiérarchisés) - 0 pt : pas de diagnostics différentiels pertinents évoqués ou diagnostics évoqués non pertinents <i>Si dossier médical évident et pas de diagnostic différentiel pertinent à évoquer : 1 pt</i>
Logique : - 4 pts : connexion entre les données médicales cohérente, orientée vers le diagnostic correct - 3 pts : connexion entre les données médicales cohérentes, orientées vers un diagnostic erroné - 2 pts : connexion incohérente des données médicales. Contradictions. - 1 pt : pas de connexion entre les données médicales visible. - 0 pt : aucune réponse générée <i>Si tous les éléments ne sont pas évoqués par le dossier mais que le raisonnement est bon : 2 pts</i>
Précision diagnostique : - 3 pts : diagnostic parfait - 2 pts : diagnostic correct mais incomplet (diagnostic syndromique, non latéralisé, sans précision de la gravité) - 1 pt : diagnostic incorrect - 0 pt : aucune réponse générée
Expression : - 2 pts : absence d'erreur syntaxique ou de formulation médicale erronée - 1 pt : erreurs de langage ou de choix de langue inadapté n'altérant pas le sens de la phrase - 0 pt : erreur altérant le sens du texte ou aucune réponse générée

3. Autres limites méthodologiques :

Notre méthode d'évaluation des réponses des modèles, comme mentionné précédemment, présente une part de subjectivité. Si des spécialistes d'autres disciplines avaient évalué les réponses de l'IA concernant leur domaine de compétence, les notes auraient pu être différentes. Peut-être moins bonnes, car évaluées plus sévèrement par exemple. Une telle méthodologie, plus lourde, n'a pas été retenue dans notre étude. Le fait que les deux auteurs soient médecins généralistes a permis une certaine homogénéité dans les évaluations, mais constitue aussi une limite de ce point de vue. Cette subjectivité dans la notation des items de la grille nous a également conduits à ne pas faire d'analyse item par item, préférant nous focaliser sur la note globale.

Notre méthodologie évaluait peut-être d'avantage la capacité des modèles à « imiter » une réponse humaine plutôt qu'à produire un raisonnement opérationnel qui contournerait complètement l'effet boîte noire. La requête initiale était assez large : « *Try your best to give your clinical reasoning in order to make an accurate final diagnosis* ». L'importance du prompting, évoquée dans la littérature (25,57), nous laisse penser qu'une requête plus développée en insistant sur nos critères de qualité permettrait d'avoir de meilleures réponses.

C. Forces de notre étude

Une des plus importantes forces que nous pouvons citer, c'est le caractère unique de notre étude. Nous n'avons retrouvé dans la littérature aucune étude évaluant spécifiquement la performance des modèles de langues dans le raisonnement clinique en français. La majorité des études observent sa capacité à répondre principalement à des examens de médecine (68–70). Bien que notre étude reprenne un grand nombre de cas cliniques issus de programmes d'entraînement aux Épreuves Classantes Nationales, nous évaluons le raisonnement des modèles autour de situations médicales, et non pas leur capacité à répondre à des questions toutes faites issues de banque de données.

Secondairement, très peu d'études en médecine comparent les résultats des modèles de langue en anglais et en français. Nous n'avons retrouvé qu'une seule étude très récente qui comparait le score obtenu par ChatGPT sur le score d'hésitation vaccinal en anglais, français et espagnol (71). A l'instar de notre travail, celle-ci montrait que le modèle donnait des réponses variables d'une langue à l'autre.

La plupart des études existant dans la littérature évaluent les modèles de langues sur des questions fermées ou à choix multiples (72,73), principalement dans le but d'automatiser l'analyse finale. Nous avons utilisé un modèle de prompt unique avec une question ouverte, plus proche d'une situation réelle, où le modèle était invité à expliciter son raisonnement

sans restriction. Cela permet d'avoir une appréciation plus précise des capacités des modèles à raisonner mais aussi à déceler des erreurs qui ne seraient pas visibles dans des questions fermées ou à choix. Nous avons retrouvé peu d'études utilisant ce type de méthodologie (74).

L'analyse des réponses était rigoureuse et standardisée. Nous avons utilisé une échelle de notation unique, nouvelle, qui évalue différents critères dans le raisonnement clinique : l'analyse des symptômes, la logique, la validité des connaissances, la mention de diagnostics différentiels et enfin l'expression. L'utilisation des sous-critères apporte une granularité plus fine dans l'évaluation de la performance des modèles. Les deux auteurs ont évalué les réponses en aveugle avec leur propre expérience médicale.

Les tendances dans les notes sont comparables avec un coefficient de Spearman élevé. Cette méthodologie a permis la production d'un grand nombre de notes (près de 2000) et des résultats obtenus significatifs, renforçant la robustesse de notre étude.

L'inclusion de près de 200 cas cliniques dans cette étude constitue une force majeure, car elle garantit une évaluation représentative des modèles à raisonner autour de cas cliniques variés. Cela permet d'apprécier non seulement leur performance globale, mais également leur adaptabilité à certaines situations, de la médecine générale à des cas plus spécialisés.

D. Perspectives

Dans notre étude, nous évaluons le raisonnement clinique des modèles de langue à l'aide d'un prompt standard. Toutefois, de nombreuses recherches ont démontré qu'il était possible d'optimiser ces prompts pour améliorer la précision des réponses. Malgré l'utilisation d'un prompt constant, nous avons observé des variations dans la profondeur des réponses d'un cas à l'autre. Il est essentiel d'éviter cet écueil en explorant les stratégies permettant d'obtenir des réponses pertinentes des modèles.

Le prompt engineering, une discipline en plein essor, joue un rôle clé dans ce domaine y compris dans la médecine. Une étude a montré qu'il était possible d'améliorer les performances diagnostiques d'un modèle en le poussant à adopter une approche de raisonnement en chaîne (56), imitant l'interaction dynamique entre un médecin et son patient. Nous encourageons à la réalisation de plus d'études dans ce domaine. Nous suggérons l'utilisation de notre grille de notation améliorée pour comparer les performances d'un modèle en fonction de différents types de prompts expérimentés. Cette approche permettra d'identifier les prompts les plus efficaces pour faire « verbaliser » au mieux le raisonnement clinique par le modèle, éviter l'effet boîte noire et ainsi produire des réponses qui pourraient être plus utiles dans la pratique courante.

Nous devons nous interroger sur les raisons qui expliquent la différence significative de performance observée entre l'anglais et le français que nous avons retrouvé dans notre étude. Comme mentionné précédemment, les pays les plus avancés dans l'intelligence artificielle sont les États-Unis et la Chine. Bien que les données d'entraînement des modèles soient confidentielles, il est estimé que l'anglais et le chinois représentent chacun environ 30% de la taille des corpus utilisés, tandis que le français n'en représente que 13% (44). Cette disparité soulève des questions sur l'équité des langues dans la construction des modèles et sur l'impact de cette répartition sur leurs performances. Si cette répartition influe réellement sur les performances cliniques des modèles, il devient crucial que la France développe sa compétitivité dans le domaine et qu'elle investisse dans le développement de modèles intégrant une proportion plus importante de français dans leurs corpus d'entraînement.

Comme l'a déjà soulevé *Joshi S et al.* dans son étude (71), ce constat peut également avoir un impact sur la conception des modèles de langue, car il met en évidence le fait que les performances peuvent varier selon la langue, potentiellement en raison d'un cloisonnement linguistique (fonctionnement « en silos »). Il serait pertinent d'explorer cette hypothèse pour offrir une approche plus équitable, indifférenciée selon la langue, en particulier dans le domaine de la santé où ces différences peuvent soulever des questions éthiques et légales.

Dans son dernier rapport (44), l'Académie Française de Médecine souligne la nécessité de généraliser l'usage des systèmes d'intelligence artificielle générative, affirmant qu'il serait « contraire à l'éthique de se passer de l'aide de ces outils ». Bien que des axes d'amélioration existent (précédemment explicités), notre étude a cherché à simuler des situations cliniques réelles. Cependant, cette approche demeure très peu intuitive. En effet, il semble peu probable qu'un médecin interrompt sa consultation pour ouvrir ChatGPT et formuler des questions détaillées sur le cas clinique qu'il rencontre, après avoir dû expliciter ce dernier. Pour que ces recommandations soient suivies, il est essentiel que ces outils soient correctement intégrés dans le flux de travail (workflow) des médecins. Cela implique aussi, comme le recommande le Conseil National de l'Ordre des Médecins (75), que les médecins participent à la conception de ces systèmes afin de garantir leur pertinence non seulement sur le plan clinique, mais aussi en tenant compte des enjeux éthiques et légaux.

Conclusion

Nous avons pu montrer, au travers de notre travail, que l'utilisation de l'intelligence artificielle générative en médecine générale reste très limitée. Il nous apparaît essentiel de fournir une plus grande diversité de solutions, en tenant compte des besoins et modes d'utilisation envisagés par les médecins généralistes. Une plus grande offre de formation sur ce sujet semble aussi incontournable pour permettre aux professionnels de santé de se l'approprier.

Notre projet de recherche a été l'occasion pour des médecins de « faire de l'IA en vrai ». Nous avons interrogé et évalué de façon rigoureuse et concrète plusieurs intelligences artificielles génératives. Nous avons mis en évidence une réelle capacité à produire un bon raisonnement diagnostique pour les plus gros modèles, avec certaines limites qui restent encore à explorer. Par ailleurs, nous avons montré une réelle différence de performance entre l'anglais et le français, ce qui pose la question de notre compétitivité mais aussi interroge sur le plan éthique. Cela plaide pour le développement d'outils souverains puissants en français, afin de ne pas dépendre de solutions privées étrangères.

Au-delà de ces résultats, nous espérons avoir montré l'importance capitale de l'implication des médecins (généralistes) dans l'évaluation, le développement et le bon usage de ces nouveaux outils qui seront sans aucun doute une révolution dans nos vies. Ce travail, original, invite à d'autres évaluations plus poussées des modèles de langue dans des conditions se rapprochant toujours plus de la réalité de la pratique clinique.

Bibliographie

1. Norvig P, Russell S. Intelligence artificielle : une approche moderne. 4e éd. Pearson; 2021. 976 p.
2. Manning C. Artificial Intelligence : Definitions [Internet]. Stanford University; 2022. Disponible sur: <https://hai.stanford.edu/sites/default/files/2023-03/AI-Key-Terms-Glossary-Definition.pdf>
3. Searle JR. Minds, brains, and programs. Behav Brain Sci. 1980;(3):417- 57.
4. Pause Giant AI Experiments: An Open Letter [Internet]. Future of Life Institute. [cité 26 juin 2024]. Disponible sur: <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>
5. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. févr 2017;542(7639):115- 8.
6. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med. janv 2019;25(1):24- 9.
7. Our World in Data [Internet]. [cité 28 juin 2024]. Annual scholarly publications on artificial intelligence. Disponible sur: <https://ourworldindata.org/grapher/annual-scholarly-publications-on-artificial-intelligence?time=2010..2021>
8. Nordlinger B, Villani C, de Fresnoye, Olivier. Médecine et intelligence artificielle. CNRS éditions. 2022. 400 p.
9. Commission de l'Intelligence Artificielle. IA : Notre ambition pour la France [Internet]. 2024 [cité 28 juin 2024]. Disponible sur: <https://www.economie.gouv.fr/cge/commission-ia>
10. Maslej N, Fattorini L, Perrault R, Parli V, Reuel A, Brynjolfsson E, et al. The AI Index 2024 Annual Report. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA; 2024.
11. Loi sur l'intelligence artificielle de l'UE - Développements et analyses actualisés de la loi sur l'intelligence artificielle de l'UE [Internet]. [cité 28 juin 2024]. Disponible sur: <https://artificialintelligenceact.eu/fr/>
12. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Networks [Internet]. arXiv; 2014 [cité 1 juill 2024]. Disponible sur: <http://arxiv.org/abs/1406.2661>
13. Salmandjee Y. ChatGPT pour les Nuls. First Interactive. 2023. 320 p. (Pour les Nuls).
14. Pratiel E. Baromètre 2024 « Les Français et les IA génératives » Vague 2. 2024;
15. Katz DM, Bommarito MJ, Gao S, Arredondo P. GPT-4 passes the bar exam. Philos Trans R Soc Math Phys Eng Sci. 26 févr 2024;382(2270):20230254.
16. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. août 2021;596(7873):583- 9.
17. Merchant A, Bartzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. Nature. déc 2023;624(7990):80- 5.
18. Brynjolfsson E, Li D, Raymond L. Generative AI at Work [Internet]. arXiv; 2023 [cité 2 juill 2024]. Disponible sur: <http://arxiv.org/abs/2304.11771>
19. McCurry J. South Korean AI chatbot pulled from Facebook after hate speech towards minorities. The Guardian [Internet]. 14 janv 2021 [cité 17 juill 2024]; Disponible sur: <https://www.theguardian.com/world/2021/jan/14/time-to-properly-socialise-hate-speech-ai-chatbot-pulled-from-facebook>
20. Freedom House [Internet]. [cité 17 juill 2024]. The Repressive Power of Artificial Intelligence. Disponible sur: <https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence>
21. La Tribune [Internet]. 2023 [cité 2 juill 2024]. ChatGPT : deux plaintes en France pour non-respect du RGPD. Disponible sur: <https://www.latribune.fr/techno-medias/informatique/chatgpt-deux-plaintes-en-france-pour-non-respect-du-rgpd-958010.html>
22. Systèmes d'IA générative en santé : enjeux et perspectives.
23. FDA has now cleared 700 AI healthcare algorithms, more than 76% in radiology [Internet]. [cité 10 juill 2024]. Disponible sur: <https://healthimaging.com/topics/artificial->

intelligence/fda-has-now-cleared-700-ai-healthcare-algorithms-more-76-radiology

24. Teare P, Fishman M, Benzaquen O, Toledano E, Elnekave E. Malignancy Detection on Mammography Using Dual Deep Convolutional Neural Networks and Genetically Discovered False Color Input Enhancement. *J Digit Imaging*. août 2017;30(4):499- 505.
25. Molcard A. Médecine générale et intelligence artificielle : revue de la littérature.
26. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Broeders M, Gennaro G, Clauser P, et al. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. *JNCI J Natl Cancer Inst*. 5 mars 2019;111(9):916- 22.
27. Kalmet PHS, Sanduleanu S, Primakov S, Wu G, Jochems A, Refaee T, et al. Deep learning in fracture detection: a narrative review. *Acta Orthop*. 13 janv 2020;91(2):215- 20.
28. Hwang EJ, Park S, Jin KN, Kim JI, Choi SY, Lee JH, et al. Development and Validation of a Deep Learning–Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs. *JAMA Netw Open*. 22 mars 2019;2(3):e191095.
29. Yu KH, Zhang C, Berry GJ, Altman RB, Ré C, Rubin DL, et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat Commun*. 16 août 2016;7:12474.
30. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*. 12 déc 2017;318(22):2199- 210.
31. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*. 13 déc 2016;316(22):2402- 10.
32. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med*. 1 juin 2023;183(6):589- 96.
33. Koranteng E, Rao A, Flores E, Lev M, Landman A, Dreyer K, et al. Empathy and Equity: Key Considerations for Large Language Model Adoption in Health Care. *JMIR Med Educ*. 28 déc 2023;9:e51199.
34. Europe TLRH. Embracing generative AI in health care. *Lancet Reg Health – Eur* [Internet]. 1 juill 2023 [cité 19 mars 2024];30. Disponible sur: [https://www.thelancet.com/journals/lanep/article/PIIS2666-7762\(23\)00096-0/fulltext](https://www.thelancet.com/journals/lanep/article/PIIS2666-7762(23)00096-0/fulltext)
35. Kandel, Bousquet, Chouilly. Manuel théorique de médecine générale. 41 concepts nécessaires à l'exercice de la discipline - Olivier Kandel, Marie-Alice Bousquet, Julie Chouilly [Internet]. Global média santé. 2018 [cité 2 janv 2022]. 207 p. (Le plaisir de comprendre). Disponible sur: <https://www.decitre.fr/livres/manuel-theorique-de-medecine-generale-9782919616183.html>
36. WONCA Europe. LA DEFINITION EUROPEENNE DE LA MEDECINE GENERALE - MEDECINE DE FAMILLE [Internet]. 2002. Disponible sur: <https://www.woncaeurope.org/file/f82a02aa-4f12-447e-ae8f-31f6c9f66c7b/WONCA%20definition%20French%20version.pdf>
37. Katon WJ, Walker EA. Medically unexplained symptoms in primary care. *J Clin Psychiatry*. 1998;59 Suppl 20:15- 21.
38. Green LA, Fryer GE, Yawn BP, Lanier D, Dovey SM. The ecology of medical care revisited. *N Engl J Med*. 28 juin 2001;344(26):2021- 5.
39. Société Française de Médecine Générale : Démarche diagnostique [Internet]. 2022 [cité 27 févr 2022]. Disponible sur: http://www.sfm-g.org/theorie_pratique/demarche_diagnostique/
40. Tousseul S. Petite histoire conceptuelle de l'homosexualité. *Psychol Clin Proj*. 2016;22(1):47- 68.
41. Kueper JK, Terry AL, Zwarenstein M, Lizotte DJ. Artificial Intelligence and Primary Care Research: A Scoping Review. *Ann Fam Med*. mai 2020;18(3):250- 8.
42. Abbasgholizadeh Rahimi S, Légaré F, Sharma G, Archambault P, Zomahoun HTV, Chandavong S, et al. Application of Artificial Intelligence in Community-Based Primary Health Care: Systematic Scoping Review and Critical Appraisal. *J Med Internet Res*. 3 sept

2021;23(9):e29839.

43. PulseLife [Internet]. 2024 [cité 3 août 2024]. Baromètre IA en santé : alliée ou menace ? Disponible sur: <https://pulselife.com/fr-fr/blog/post/barometre-ia-en-sante-alliee-ou-menace>
44. Nordlinger B, Kirchner C, De Fresnoye O. Rapport 24-03. Systèmes d'IA générative en santé : enjeux et perspectives. Bull Académie Natl Médecine. mai 2024;208(5):536- 47.
45. Intégration des dispositifs médicaux numériques à usage professionnel dans la pratique. 2022;
46. Buck C, Doctor E, Hennrich J, Jöhnk J, Eymann T. General Practitioners' Attitudes Toward Artificial Intelligence–Enabled Systems: Interview Study. J Med Internet Res. 27 janv 2022;24(1):e28916.
47. Moukrim B. Intelligence artificielle en santé : espoirs et craintes des médecins généralistes. Sciences du Vivant. 2019;
48. Varghese J. Artificial Intelligence in Medicine: Chances and Challenges for Wide Clinical Adoption. Visc Med. déc 2020;36(6):443- 9.
49. Nabla · L'assistant médical qui vous libère du temps [Internet]. 2024 [cité 10 juill 2024]. Disponible sur: <https://www.nabla.com/fr/>
50. Doctolib Pro - France [Internet]. 2024 [cité 10 juill 2024]. Doctolib | Assistant de Consultation. Disponible sur: <https://info.doctolib.fr/doctolib-medecin/assistant-consultation/>
51. Liu J, Wang C, Liu S. Utility of ChatGPT in Clinical Practice. J Med Internet Res. 28 juin 2023;25:e48568.
52. Rao A, Pang M, Kim J, Kaminen M, Lie W, Prasad AK, et al. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study. J Med Internet Res. 22 août 2023;25:e48659.
53. Tabla S. Modules d'intelligence artificielle au cabinet de ville : qu'en attendent les médecins généralistes français ? Université de Lille; 2021.
54. Youssef DERRADJI et Marie LAFARGE. L'utilité de l'intelligence artificielle en soins primaires : Le point de vue des médecins généralistes Étude qualitative auprès de médecins généralistes exerçant en ambulatoire en Occitanie [Internet]. Toulouse; 2022. Disponible sur: <http://thesesante.ups-tlse.fr/4015/1/2022TOU31075-1076.pdf>
55. MARTIN-GAUTHIER J. Nantes Université. Joëlle MARTIN-GAUTHIER; [cité 5 nov 2023]. MALADES - Grands Modèles de Langue Adaptables et Souverains pour le Domaine Médical Français. Disponible sur: <https://www.univ-nantes.fr/universite/vision-strategie-et-grands-projets/malades-grands-modeles-de-langue-adaptables-et-souverains-pour-le-domaine-medical-francais>
56. Wu CK, Chen WL, Chen HH. arXiv.org. 2023 [cité 27 sept 2024]. Large Language Models Perform Diagnostic Reasoning. Disponible sur: <https://arxiv.org/abs/2307.08922v1>
57. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. NPJ Digit Med [Internet]. 2024 [cité 29 mai 2024];7. Disponible sur: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10808088/>
58. Grabar N, Claveau V, Dalloux C. CAS: French Corpus with Clinical Cases. In: Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis [Internet]. Brussels, Belgium: Association for Computational Linguistics; 2018 [cité 29 sept 2024]. p. 122- 8. Disponible sur: <http://aclweb.org/anthology/W18-5614>
59. Guide [Internet]. Futur Interne. [cité 19 mars 2024]. Disponible sur: <http://www.futur-interne.com/guide/>
60. Wang LW, Miller MJ, Schmitt MR, Wen FK. Assessing readability formula differences with written health information materials: Application, results, and recommendations. Res Soc Adm Pharm. 1 sept 2013;9(5):503- 16.
61. Beregi M, Magnaval J. L'intégration de l'Intelligence Artificielle dans la pratique des médecins généralistes : une exploration des pratiques actuelles et des perspectives futures. 18 avr 2024;52.

62. Compilatio [Internet]. [cité 27 sept 2024]. L'IA dans l'enseignement : résultats détaillés d'une enquête où étudiants et enseignants confrontent leurs regards. Disponible sur: <https://www.compilatio.net/blog/enquete-ia-enseignement-2023>
63. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare*. 19 mars 2023;11(6):887.
64. Yim D, Khuntia J, Parameswaran V, Meyers A. Preliminary Evidence of the Use of Generative AI in Health Care Clinical Services: Systematic Narrative Review. *JMIR Med Inform*. 20 mars 2024;12:e52073.
65. https://www.orspaysdelaloire.com/sites/default/files/pages/pdf/2021_PDF/2021_31_MED_SPE_generalistes_v3.pdf [Internet]. [cité 18 sept 2024]. Disponible sur: https://www.orspaysdelaloire.com/sites/default/files/pages/pdf/2021_PDF/2021_31_MED_SPE_generalistes_v3.pdf
66. Jang J, Kim S, Ye S, Kim D, Logeswaran L, Lee M, et al. Exploring the Benefits of Training Expert Language Models over Instruction Tuning [Internet]. *arXiv*; 2023 [cité 27 sept 2024]. Disponible sur: <http://arxiv.org/abs/2302.03202>
67. Yang H, Zhang Y, Xu J, Lu H, Heng PA, Lam W. Unveiling the Generalization Power of Fine-Tuned Large Language Models [Internet]. *arXiv*; 2024 [cité 27 sept 2024]. Disponible sur: <http://arxiv.org/abs/2403.09162>
68. Cherif H, Moussa C, Missaoui AM, Salouage I, Mokaddem S, Dhahri B. Appraisal of ChatGPT's Aptitude for Medical Education: Comparative Analysis With Third-Year Medical Students in a Pulmonology Examination. *JMIR Med Educ*. 23 juill 2024;10:e52818.
69. Panthier C, Gatinel D. Success of ChatGPT, an AI language model, in taking the French language version of the European Board of Ophthalmology examination: A novel approach to medical knowledge assessment. *J Fr Ophtalmol*. 1 sept 2023;46(7):706- 11.
70. Guigue PA, Meyer R, Thivolle-Lioux G, Brezinov Y, Levin G. Performance of ChatGPT in French language Parcours d'Accès Spécifique Santé test and in OBGYN. *Int J Gynecol Obstet*. 2024;164(3):959- 63.
71. Joshi S, Ha E, Rivera Y, Singh VK. ChatGPT and Vaccine Hesitancy: A Comparison of English, Spanish, and French Responses Using a Validated Scale. *AMIA Summits Transl Sci Proc*. 31 mai 2024;2024:266- 75.
72. Shieh A, Tran B, He G, Kumar M, Freed JA, Majety P. Assessing ChatGPT 4.0's test performance and clinical diagnostic accuracy on USMLE STEP 2 CK and clinical case reports. *Sci Rep*. 23 avr 2024;14(1):9330.
73. Hadi A, Tran E, Nagarajan B, Kirpalani A. Evaluation of ChatGPT as a diagnostic tool for medical learners and clinicians. *PLOS ONE*. 31 juill 2024;19(7):e0307383.
74. Warriar A, Singh R, Haleem A, Zaki H, Eloy JA. The Comparative Diagnostic Capability of Large Language Models in Otolaryngology. *The Laryngoscope*. 2024;134(9):3997- 4002.
75. Philippe CINQUIN, TIMC Imag ; Marie-Christine, JAULENT, LIMICS ; Jacques MARESCAUX, IRCAD ; Fabrice DENIS, Institut interrégional, de cancérologie Jean-Bernard ; Guy, FAGHERAZZI, Institut Gustave Roussy ; Raja, CHATILA, ISIR ; Roman ROUZIER, Institut Curie, et al. Médecins et patients dans le monde des DATA, des algorithmes et de l'intelligence artificielle - Analyse et recommandations du CNOM [Internet]. 2018. Disponible sur: https://www.conseil-national.medecin.fr/sites/default/files/cnomdata_algorithmes_ia_0.pdf

Annexes

ANNEXE 1 – QUESTIONNAIRE

Question 1 : Avez-vous déjà, à votre connaissance, interagi avec un algorithme d'IA ?

Réponse à choix unique : Oui / Non.

Question 2 : Avez-vous déjà interagi avec un algorithme d'IA dans le cadre professionnel ?

Réponse à choix unique : Oui / Non.

Question 3 : Avez-vous déjà utilisé une IA générative ?

Réponse à choix unique : Oui / Non.

(Exemple : ChatGPT)

Question 4 : Sur une échelle de 1 à 10, quel serait selon vous le potentiel des IA génératives à transformer votre activité professionnelle au quotidien ?

Réponse à choix unique : 1 à 10.

(1 : aucun changement ; 10 : modification radicale)

Question 5a : A quelle fréquence utilisez-vous des outils d'IA générative (ex : ChatGPT) dans votre quotidien ?

Réponse à choix unique : Jamais / Rarement (quelques fois par an) / Occasionnellement (quelques fois par mois) / Régulièrement (au moins une fois par semaine) / Quotidiennement

Question 5b : A quelle fréquence utilisez-vous des outils d'IA générative (ex : ChatGPT) dans votre pratique professionnelle ?

Réponse à choix unique : Jamais / Rarement (quelques fois par an) / Occasionnellement (quelques fois par mois) / Régulièrement (au moins une fois par semaine) / Quotidiennement

Question 6 : Selon vous, pour quelles tâches les outils d'IA génératives sont ou seraient-elles le plus utiles dans votre pratique ?

Réponse à choix multiple (**3 réponses max**) : Aide au diagnostic / Décision thérapeutique / Analyse des données (examens paracliniques, résumés médicaux, etc.) / Recherche médicale (rédaction scientifique, synthèse de recommandations, etc.) / Tâches administratives

Question 7 : Quels sont, selon vous, les principaux freins à l'utilisation de l'IA générative dans votre pratique ?

Réponse à choix multiple (**3 réponses max**) : Crainte d'un remplacement ou perte d'emploi / Manque de formation ou de connaissances sur l'outil / Manque de confiance dans les résultats de l'IA / Défaut d'intégration de l'IA aux logiciels métiers / Préoccupations éthiques ou de confidentialité des données / Perception négative par les patients / Problèmes de coût

Question 8 : Êtes-vous favorable au développement de l'IA générative dans votre pratique ?

Réponse à choix unique : Très favorable / Favorable / Neutre / Défavorable / Très défavorable

Question 9 : Dans quelle proportion de vos consultations pensez-vous que l'utilisation de l'IA générative pourrait vous aider de façon significative ?

Réponse à choix unique : Je ne compte pas l'utiliser / < 25% / 25 à 50% / 50% à 75% / > 75%

Question 10 : Ressentez-vous le besoin de suivre une formation spécifique sur les IA génératives ?

Réponse à choix unique : Oui, absolument / Peut-être, selon le contenu / Non, pas nécessairement

Question 11 : Dans quel format de formation préférez-vous apprendre sur les IA génératives ?

Réponse à choix unique : Cours ou ateliers physiques / Webinaires en ligne / MOOCs (cours en ligne en modules) / Documents écrits (articles, livres)

Votre âge

Réponse à choix unique : < 30 ans / 30 - 40 ans / 40 - 50 ans / > 50 ans / Très défavorable

Votre sexe

Réponse à choix unique : Homme / Femme

Type d'exercice principal

Liste à choix : Maison de santé pluridisciplinaire / Cabinet de groupe / Cabinet seul

Département d'exercice

Liste à choix contenant l'ensemble des départements de France

Votre milieu d'exercice principal

Réponse à choix unique : Urbain / Semi-rural / Rural

Vos compétences en informatique

Réponse à choix unique : Pas du tout compétent / Peu compétent / Compétent / Très compétent

Commentaire additionnel

Commentaire libre : Vous pouvez nous donner plus d'informations sur vos utilisations actuelles en IA et vos attentes autour des IA génératives.

ANNEXE 2 – TYPES DE RAISONNEMENT DIAGNOSTIQUES

Les types de raisonnements diagnostiques sont nombreux et nombreuses sont les classifications et dénominations utilisées. Nous avons fait le choix d'utiliser les dénominations et définitions mentionnés par Kandel, Bousquet et Chouilly dans leur manuel théorique de médecine générale.

- Les processus de raisonnement « non analytiques » ou « intuitifs » correspondent aux processus de raisonnement qui reposent sur une reconnaissance sans efforts conscients d'une configuration caractéristique de signes (reconnaissance de forme ou *pattern recognition*) ou une situation clinique déjà vécue dans le passé qui évoque très fortement un ou plusieurs diagnostics. Ces processus d'identification induite par analogie s'appliquent à des cas simples et typiques (*purpura fulminans*) et nécessitent une culture médicale.

- Le raisonnement hypothético-déductif : il s'agit d'une démarche analytique de vérification systématique des hypothèses diagnostiques (souvent générées elles-mêmes intuitivement par processus non analytique). Le praticien recherche consciemment (par l'interrogatoire, l'examen clinique, les examens complémentaires) à confirmer ou à rejeter les hypothèses diagnostiques à envisager.

- Le raisonnement en chaînage avant : il s'agit d'une démarche analytique. Elle est mise en œuvre quand un médecin ne parvient pas à identifier rapidement des formes ou des exemples concrets, en raison de son manque d'expérience, de la complexité du cas, ou de sa rareté. Le cheminement va se faire consciemment des données cliniques et paracliniques vers la solution, grâce à l'application de règles causales ou conditionnelles (mobilisation des connaissances, physiopathologie...).

- L'approche de type interniste par réalisation d'un dossier d'évaluation exhaustif : cette approche consiste à faire l'inventaire de tous les éléments susceptibles d'aboutir à un diagnostic, avec tentative d'exploration exhaustive de toutes les hypothèses possibles, même les plus rares et les plus improbables. C'est une démarche que le médecin peut être tenté d'avoir face à des plaintes multiples et complexes. Elle peut être utile dans le cas de maladies univoques et est utilisée dans la formation des médecins. Cependant, en pratique courante son coût en temps et en examens complémentaires multiples la rendent moins utile. De plus, elle expose à la découverte de nouvelles données parasites qui peuvent occulter le diagnostic ou orienter vers de nouvelles affections ce qui entraîne de nouveaux examens complémentaires et de l'anxiété chez le patient.

- L'utilisation d'arbres décisionnels ou d'algorithme : le cheminement est de type binaire, amenant le médecin à proposer un diagnostic par éliminations successives, en fonction de la présence ou non d'un signe, ou de la positivité ou non d'un test. Il y a peu d'arbres diagnostiques disponibles et ces outils sont inefficaces dès qu'il s'agit d'une situation complexe.

- La démarche probabiliste : cette démarche est fondée sur les prévalences connues ou estimées et les VPN / VPP des signes et tests. Elle nécessite la connaissance de la prévalence de l'affection envisagée dans la population du patient (ou probabilité pré test), et des valeurs du RV+ et RV- de chaque élément d'information clinique ou paraclinique. L'objectif de la démarche diagnostique probabiliste est ainsi d'estimer la probabilité d'un diagnostic à partir des infos recueillies : il s'agit de la probabilité post test. Cette démarche peut être utile en situation de grande incertitude. Elle invite alors à se doter de quelques repères sur les performances diagnostiques des données cliniques et paracliniques, afin de rechercher tel ou tel signe discriminant ou de réaliser tel examen complémentaire dont le résultat influencera de façon importante la probabilité post test de la maladie.

ANNEXE 3 – RESULTATS DES ANALYSES COMPARATIVES

Les analyses comparatives ont été effectuées en utilisant le langage Python et les bibliothèques Numpy, Pandas, SciPy. Les graphiques permettant de visualiser les données ont été faites en utilisant Matplotlib et Seaborn. Nous présentons dans cette annexe les résultats des tests statistiques. Les valeurs sont arrondies au centième. Nous avons cherché à comparer à faire les comparaisons suivantes :

- GPT en anglais contre Llama en anglais
- GPT en anglais contre GPT en français
- Llama en anglais contre Llama en français
- GPT en français contre Llama en français

Avant d'utiliser les tests de Wilcoxon, nous avons vérifié si les notes pour chaque test effectué suivaient une distribution normale en utilisant le test de Shapiro :

Test effectué	Shapiro	Valeur de p
GPT anglais vs Llama anglais	0.921299	$p = 8.34 \times 10^{-13}$
GPT anglais vs GPT français	0.919125	$p = 5.23 \times 10^{-13}$
Llama anglais vs Llama français	0.942725	$p = 1.42 \times 10^{-10}$
GPT français vs Llama français	0.937389	$p = 3.56 \times 10^{-11}$

Chaque test ne suivant pas une distribution normale, nous avons réalisé les 4 tests de Wilcoxon que nous présentons dans le tableau suivant. Nous avons adapté le seuil de significativité α à 0.0125.

Test effectué	Wilcoxon	Valeur de p
GPT anglais vs Llama anglais	10337	$p = 1.26 \times 10^{-1}$
GPT anglais vs GPT français	9046	$p = 2.16 \times 10^{-6}$
Llama anglais vs Llama français	7607	$p = 7.68 \times 10^{-13}$
GPT français vs Llama français	10618.5	$p = 9.81 \times 10^{-6}$

Pour comparer les notes entre les spécialités, nous avons réalisé un test de Kruskal-Wallis pour chaque modèle en fonction de la langue, que nous présentons dans le tableau suivant. Chaque analyse étant distincte l'une de l'autre (notes indépendantes isolées par modèle et langue), nous avons conservé un seuil de significativité α à 0.05.

Test effectué	Kruskal-Wallis	Valeur de p
GPT anglais	15.9391	$p = 0.3861$
Llama anglais	9.10036	$p = 0.8722$
GPT français	28.2417	$p = 0.0201$
Llama français	31.9362	$p = 0.0066$

Les tests post-hoc de Dunn sont présentés sous forme de cartes de chaleur avec les valeurs de p et sont disponibles dans la section **Résultats**.

Nous mettons à disposition les codes des analyses, des tableurs descriptifs des cas et les notes des auteurs dans un dossier GitHub accessible grâce au QR Code ci-dessous.



NOM : SID-AHMED
NOM : CORVELLEC

PRENOM : Sofiane-Djillali
PRENOM : Josselin

Titre de Thèse : IA générative en médecine générale, attentes des professionnels et évaluation de grands modèles de langue sur le raisonnement diagnostique

RESUME

L'Intelligence Artificielle générative est en pleine expansion et la santé est un champ privilégié pour ses applications. Les attentes des médecins généralistes autour de ces technologies semblent mal cernées. Les travaux d'évaluation produits par ces mêmes médecins sont peu nombreux. Notre travail tente de répondre à cette double problématique. D'une part, nous avons réalisé une enquête auprès des médecins généralistes quant à leurs utilisations et attentes vis-à-vis de l'IA générative. D'autre part, nous avons testés des grands modèles de langues sur leur capacité à produire un raisonnement diagnostique de qualité.

Un questionnaire a été diffusé auprès des médecins généralistes de la région Pays de La Loire durant l'été 2024. En parallèle, nous avons élaboré une banque de cas clinique qui ont été soumis à différents LLM (GPT-4, Llama3 et BioMistral) afin de les évaluer sur le raisonnement diagnostique.

Nos résultats montrent une utilisation encore limitée des outils d'IA générative en pratique courante par les médecins généralistes. Les médecins interrogés attendent surtout de l'IA générative une aide quant aux tâches administratives et à l'analyse de données. Ils sont également demandeurs de formation sur le sujet. L'évaluation des réponses des modèles de langue, quant à elle, a montré une excellente performance de ces derniers sur la capacité à produire un raisonnement de qualité. Les modèles les plus grands semblent meilleurs que les autres. Nous avons aussi montré que les modèles étaient plus performants en anglais par rapport au français.

D'autres travaux seront nécessaires afin d'évaluer les modèles de langue dans des situations se rapprochant plus de la pratique réelle. Notre méthodologie d'évaluation pourra être réutilisée pour d'autres travaux.

MOTS-CLES

Médecine, Médecine générale, raisonnement diagnostique, diagnostic, intelligence artificielle, IA, AI, intelligence artificielle générative, grands modèles de langue, large language model, LLM, ChatGPT, GPT, Llama, Biomistral, Mistral