Étude quantitative de la fiabilité de ChatGPT sur des questions médicales : focus sur l'asthme

Pedro Marcal Lopes

UNIVERSITE CÔTE D'AZUR

ANNEE 2024

THESE D'EXERCICE DE MEDECINE

Pour l'obtention du diplôme d'Etat de Docteur en Médecine

Etude quantitative de la fiabilité de ChatGPT sur des questions médicales : focus sur l'asthme

Présentée et soutenue le 18 octobre 2024

A la faculté de Médecine de Nice

Par **Pedro MARCAL LOPES**

Né le 15 février 1995 à Lisbonne, (99) Portugal

MEMBRES DU JURY

Président :

Monsieur le Professeur David DARMON

Assesseurs :

Monsieur le Docteur Armand BISROR

Directeur de thèse :

Monsieur le Docteur Fabien ROLLAND

UNIVERSITÉ CÔTE D'AZUR
FACULTÉ DE MÉDECINE

Liste des enseignants au 1er septembre 2023 à l' U. F. R. Médecine de Nice

Doyen

Pr. Jean DELLAMONICA

Vice-doyennes

Pédagogie	**Pr. Véronique ALUNNI**
Recherche	**Pr. Barbara SEITZ-POLSKI**
Relations internationales	**Pr Fanny BUREL-VANDENBOS**

Conservateur de la bibliothèque	Mme Danièle AMSELLE
Directrice administrative des services	Mme Isabelle CALLEA

Doyens Honoraires	M. Patrick RAMPAL
	M. Daniel BENCHIMOL
	M. Patrick BAQUÉ

UNIVERSITÉ CÔTE D'AZUR
FACULTÉ DE MÉDECINE

Liste des enseignants au 1er septembre 2023 à l' U. F. R. Médecine de Nice

PROFESSEURS DES UNIVERSITES-PRATICIENS HOSPITALIERS

Mme	Véronique	ALUNNI	Médecine Légale et Droit de la Santé (46.03)
M	Nicolas	AMORETTI	Radiologie et Imagerie Médicale (43.02)
M.	Rodolphe	ANTY	Gastro-entérologie (52.01)
Mme	Florence	ASKENAZY-GITTARD	Pédopsychiatrie (49.04)
M.	Philippe	BAHADORAN	Cytologie et Histologie (42.02)
Mme	Stéphanie	BAILLIF	Ophtalmologie (55.02)
Mme	Sylvie	BANNWARTH	Génétique (47.04)
M.	Patrick	BAQUÉ	Anatomie - Chirurgie Générale (42.01)
M.	Emmanuel	BARRANGER	Gynécologie Obstétrique (54.03)
M.	Emmanuel	BENIZRI	Chirurgie Générale (53.02)
M.	Michel	BENOIT	Psychiatrie (49.03)
M.	Gilles	BERNARDIN	Réanimation Médicale (48.02)
M.	J-Philippe	BERTHET	Chirurgie Thoracique (51-03)
M.	André	BONGAIN	Gynécologie-Obstétrique (54.03)
M.	Alexandre	BOZEC	ORL- Cancérologie (47.02)
M.	Jean	BREAUD	Chirurgie Infantile (54-02)
Mme	Véronique	BREUIL	Rhumatologie (50.01)
M.	Nicolas	BRONSARD	Anatomie Chir Ortho et Traumato (42.01)
M.	Olivier	CAMUZARD	Chirurgie Plastique (50-04)
Mme	Fanny	BUREL-VANDENBOS	Anat. cytolo. path. (42.03)
M.	Michel	CARLES	Mal. infect. ; trop. (45.03)
M.	Laurent	CASTILLO	O.R.L. (55.01)
M.	Nicolas	CHEVALIER	Endo.diab.mal. métab (54.04)
M.	Patrick	CHEVALLIER	Radiologie et Imagerie Médicale (43.02)
Mme	Giulia	CHINETTI	Biochimie-Biologie Moléculaire (44.01)
M.	Thomas	CLUZEAU	Hématologie (47.01)
M.	Jacques	DARCOURT	Biophysique et Médecine Nucléaire (43.01)
M.	David	DARMON	Médecine Générale (53,03)
M.	Jean	DELLAMONICA	Réanimation médicale (48.02)
M.	Jérôme	DELOTTE	Gynécologie-obstétrique (54.03)
M.	Jérôme	DOYEN	Radiothérapie (47.02)
M.	Milou-Daniel	DRICI	Pharmacologie Clinique (48.03)
M.	Matthieu	DURAND	Urologie (52.04)

UNIVERSITÉ CÔTE D'AZUR
FACULTÉ DE MÉDECINE

Liste des enseignants au 1er septembre 2023 à l' U. F. R. Médecine de Nice

PROFESSEURS DES UNIVERSITES-PRATICIENS HOSPITALIERS

M.	Vincent	ESNAULT	Néphrologie (52-03)
Mme	Christelle	ESTRAN-POMARES	Parasitologie et mycologie (45.02)
M	Guillaume	FAVRE	Physiologie (44.02)
M.	Emile	FERRARI	Cardiologie (51.02)
M.	J-Marc	FERRERO	Cancérologie ; Radiothérapie (47.02)
M.	Denys	FONTAINE	Neurochirurgie (49.02)
M.	J-Paul	FOURNIER	Thérapeutique (48-04)
M.	Eric	GILSON	Biologie Cellulaire (44.03)
Mme	Valérie	GIORDANENGO	Bactériologie-Virologie (45.01)
Mme	Lisa	GIOVANNINI-CHAMI	Pédiatrie (54.01)
M.	Olivier	GUERIN	Méd. In ; Gériatrie (53.01)
M.	Nicolas	GUEVARA	Oto-Rhino-laryngologie (55.01)
M.	Jean	GUGENHEIM	Chirurgie Digestive (52.02)
M.	J-Michel	HANNOUN-LEVI	Cancérologie ; Radiothérapie (47.02)
M.	Reda	HASSEN KHODJA	Chirurgie Vasculaire (51.04)
M.	Xavier	HÉBUTERNE	Nutrition (44.04)
M.	Paul	HOFMAN	Anat. cytolo. path. (42.03)
M.	Olivier	HUMBERT	Biophysique et Médecine Nucléaire (43.01)
M.	Antonio	IANNELLI	Chirurgie Digestive (52.02)
Mme	Carole	ICHAI	Anesth. réa. (48.01)
M.	Marius	ILIÉ	Anat. cytolo. path. (42.03)
M	Elixène	JEAN-BAPTISTE	Chirurgie vasculaire (51.04)
M.	Georges	LEFTHERIOTIS	Physiologie ; médecine vasculaire (51.04)
Mme	Sylvie	LEROY	Pneumologie-Addictologie (51.01)
M.	Jacques	LEVRAUT	Médecine d'urgence (48.05)
M.	Michel	LONJON	Neurochirurgie (49.02)

UNIVERSITÉ CÔTE D'AZUR
FACULTÉ DE MÉDECINE

Liste des enseignants au 1er septembre 2023 à l' U. F. R. Médecine de Nice

PROFESSEURS DES UNIVERSITES-PRATICIENS HOSPITALIERS

M.	Charles	MARQUETTE	Pneumologie (51,01)
M.	J-François	MICHIELS	Anat. cytolo. path. (42.03)
Mme	Pamela	MOCERI	Cardiologie (51.02)
M.	Henri	MONTAUDIÉ	Dermatologie (50.03)
M.	Nicolas	MOUNIER	Cancérologie, Radiothérapie (47.02)
M.	J-Christophe	ORBAN	Anesth. réa. (48.01)
M.	Bernard	PADOVANI	Radiologie et Imagerie Médicale (43.02)
M.	Philippe	PAQUIS	Neurochirurgie (49.02)
Mme	Véronique	PAQUIS	Génétique (47.04)
M.	Thierry	PASSERON	Dermato-Vénéréologie (50-03)
M.	Thierry	PICHE	Gastro-entérologie (52.01)
M.	Christian	PRADIER	Epid., éco. santé (46.01)
Mme	Virginie	RAMPAL	Chirurgie Infantile (54-02)
M.	Pierre	ROHRLICH	Pédiatrie (54.01)
M.	Eric	ROSENTHAL	Médecine Interne (53.01)
M.	Christian	ROUX	Rhumatologie (50.01)
M.	Raymond	RUIMY	Bactériologie-virologie (45.01)
Mme	Sabrina	SACCONI	Neurologie (49.01)
Mme	Nirvana	SADAGHIANLOO	Chirurgie vasculaire (51.04)
M.	Stéphane	SCHNEIDER	Nutrition (44.04)
Mme	Barbara	SEITZ-POLSKI	Immunologie (47.03)
M.	Antoine	SICARD	Néphrologie (52.03)
M.	Pascal	STACCINI	Biostat. inf.méd. TC (46.04)
M.	Pierre	THOMAS	Neurologie (49.01)
M.	Albert	TRAN	Hépato Gastro-entérologie (52.01)
M.	Geoffroy	VANBIERVLIET	Gastro-entérologie (52.01)

UNIVERSITÉ CÔTE D'AZUR
FACULTÉ DE MÉDECINE

Liste des enseignants au 1er septembre 2023 à l' U. F. R. Médecine de Nice

MAITRES DE CONFÉRENCES DES UNIVERSITÉS - PRATICIENS HOSPITALIERS

M.	Damien	AMBROSETTI	Cytologie et Histologie (42.02)
Mme	Caroline	BERNARDI	Médecine légale et droit de la Santé (46.03)
Mme	Julie	BERNARDOR	Pédiatrie (54.01)
Mme	Ghislaine	BERNARD-POMIER	Immunologie (47.03)
Mme	Tiphanie	BOUCHEZ	Médecine Générale (53.03)
Mme	Julie	CONTENTI-LIPRANDI	Médecine d'urgence (48-04)
M.	Johan	COURJON	Mal. infect. ; trop. (45.03)
Mme	Bérengère	DADONE-MONTAUDIÉ	Cancérologie-radiothérapie (47.02)
M.	Alain	DOGLIO	Bactériologie-Virologie (45.01)
M.	Arnaud	FERNANDEZ	Pédopsychiatrie (49-04)
Mme	Charlotte	HINAULT	Biochimie et biologie moléculaire (44.01)
M.	Mathieu	JOZWIAK	Médecine intensive-Réanimation (48.02)
Mme	Brigitte	LAMY	Bactérilogie-virologie (45.01)
Mme	Elodie	LONG-MIRA	Cytologie et Histologie (42.02)
M.	Michaël	LOSCHI	Hématologie et Transfusion (47.01)
M.	Romain	LOTTE	Bact-vir ; Hyg.hosp. (45.01)
Mme	Marie-Noëlle	MAGNIÉ	Physiologie (44.02)
M.	Arnaud	MARTEL	Ophtalmologie (55.02)
M.	Nihal	MARTIS	Méd int. ; gériatrie (53.01)
M.	Damien	MASSALOU	Chirurgie Viscérale (52-02)
Mme	Sandra	MUSSO-LASSALLE	Anat. cytolo. path. (42.03)
M.	Mourad	NAÏMI	Biochimie et Biologie moléculaire (44.01)
Mme	Céline	OCELLI	Médecine d'urgence (48-04)
M.	Charles	SAVOLDELLI	Chir. maxill. & stom (55.03)
M.	Fabien	SQUARA	Cardiologie (51.02)
Mme	Susanne	THÜMMLER	Pédopsychiatrie (49-04)
M.	Antoine	TRAN	Pédiatrie (54.01)

UNIVERSITÉ CÔTE D'AZUR
FACULTÉ DE MÉDECINE

Liste des enseignants au 1er septembre 2023 à l' U. F. R. Médecine de Nice

MAITRE DE CONFÉRENCES DES UNIVERSITÉS

Mme	Auriane	GROS	Orthophonie (69)

PROFESSEURS AGRÉGÉS

Mme	Rebecca	LANDI	Anglais

PRATICIEN HOSPITALIER UNIVERSITAIRE

Mme	Emeline	MICHEL	Médecine interne-Gériatrie (53.01)

PROFESSEURS ASSOCIÉS

Mme	Christine	LEBRUN-FRENAY	Neurologie (49.01)
Mme	Brigitte	MONNIER	Médecine Générale (53.03)
Mme	Flora	TREMELLAT-FALIERE	Médecine palliative (46.05)

MAITRES DE CONFÉRENCES ASSOCIÉS

Mme	Céline	CASTA	Médecine Générale (53.03)
M.	Fabrice	GASPERINI	Médecine Générale (53.03)
M.	Marc-André	GUERVILLE	Médecine Générale (53.03)
Mme	Maud	RAQUIN-POUILLON	Médecine Générale (53.03)

UNIVERSITÉ CÔTE D'AZUR
FACULTÉ DE MÉDECINE

Liste des enseignants au 1er septembre 2023 à l' U. F. R. Médecine de Nice

Constitution du jury en qualité de 4ème membre

Professeurs Honoraires

M.	Marc	ALBERTINI	M.	Pierre	GIBELIN
M.	Jean	AMIEL	M.	J-Yves	GILLET
M.	Daniel	BALAS	M.	Patrick	GRELLIER
M.	Michel	BATT	M.	Dominique	GRIMAUD
M.	Etienne	BÉRARD	M.	Philippe	HOFLIGER
M.	Bruno	BLAIVE	M.	Jacques	JOURDAN
Mme	Florence	BLANC-PEDEUTOUR	M.	J-Philippe	LACOUR
M.	Patrice	BOQUET	M.	J-Claude	LAMBERT
M.	André	BOURGEON	M.	Michel	LAZDUNSKI
M.	Patrick	BOUTTÉ	M.	Yves	LE_FICHOUX
M.	J-Noël	BRUNETON	M.	J-Claude	LEFEBVRE
Mme	Françoise	BUSSIERE	M.	Roger	MARIANI
M.	J-Pierre	CAMOUS	M.	Pierre	MARTY
M.	Bertrand	CANIVET	M.	René	MASSEYEFF
M.	Jill-patrice	CASSUTO	M.	Mathieu	MATTEI
M.	Marcel	CHATEL	M.	Jean	MOUIEL
M.	Alain	COUSSEMENT	M.	Jérôme	MOUROUX
Mme	Dominique	CRENESSE	Mme	Martine	MYQUEL
M.	Guy	DARCOURT	M.	Dominique	PRINGUEY
M.	Fernand	DE_PERETTI	M.	Gérald	QUATREHOMME
M.	Pierre	DELLAMONICA	M.	Marc	RAUCOULES-AIMÉ
M.	Jean	DELMONT	Mme	Dominique	RAYNAUD
M.	François	DEMARD	M.	Philippe	ROBERT
M.	Claude	DESNUELLE	M.	Joseph	SANTINI
M.	Claude	DOLISI	M.	J- Baptiste	SAUTRON
M.	Patrick	FENICHEL	M.	Maurice	SCHNEIDER
M.	Alain	FRANCO	M.	Antoine	THYSS
M.	Pierre	FREYCHET	M.	Jacques	TOUBOL
M.	J-Gabriel	FUZIBET	M.	Dinh Khiem	TRAN
M.	Pierre	GASTAUD	M.	Emmanuel	VAN OBBERGHEN
M.	J-Pierre	GÉRARD			

UNIVERSITÉ CÔTE D'AZUR
FACULTÉ DE MÉDECINE

Liste des enseignants au 1er septembre 2023 à l' U. F. R. Médecine de Nice

Constitution du jury en qualité de 4ème membre

M.C.U. Honoraires

M.	Jacques	ARNOLD	M.	Marcel	GASTAUD
M.	Bernard	BASTERIS	M.	Jean	GIUDICELLI
M.	José	BENOLIEL	M.	Jacques	MAGNÉ
Mlle	Rose-Marie	CHICHMANIAN	Mme	Nadine	MEMRAN
Mme	Michèle	DONZEAU	M.	Raymond	MENGUAL
M.	Roméo	EMILIOZZI	M.	Patrick	PHILIP
M.	Thierry	FOSSE	M.	J-Claude	POIRÉE
M.	Philippe	FRANKEN	Mme	Marie-Claire	ROURE
M.	Rodolphe	GARRAFFO	M.	Jean	TESTA
			M.	Pierre	TOULON

UNIVERSITÉ CÔTE D'AZUR
FACULTÉ DE MÉDECINE

Liste des enseignants au 1er septembre 2023 à l' U. F. R. Médecine de Nice

PROFESSEURS CONVENTIONNÉS DE L'UNIVERSITÉ

M.	François	BERTRAND	Médecine Interne
M.	Patrice	BROCKER	Médecine Interne Option Gériatrie
M.	Daniel	CHEVALLIER	Urologie
Mme	Manuella	FOURNIER-MEHOUAS	Médecine Physique et Réadaptation
M.	Patrick	JAMBOU	Coordination prélèvements d'organes
M.	Mathieu	LEBOEUF	gynécologie- obstétrique
Mme	Geneviève	NADEAU	uro-gynécologie
M.	Guillaume	ODIN	Chirurgie maxilo-faciale
M.	Frédéric	PEYRADE	Onco-Hématologie
M.	Bertrand	PICCARD	Psychiatrie
M.	J-François	QUARANTA	Santé Publique

Remerciements

Monsieur le Professeur DARMON David,

Je vous remercie d’avoir accepté de présider ce jury de thèse, je vous suis profondément reconnaissant de votre présence aujourd’hui.

Monsieur le Docteur ROLLAND Fabien,

Je te remercie d’avoir accepté d’etre mon directeur de thèse. Tu as su me guider tout au long de cette aventure intellectuelle. Je te suis particulièrement redevable pour avoir eu l'idée du sujet de cette thèse, ainsi que pour m'avoir fourni les questions de l’application PneumoQuiz, qui ont été essentielles pour mener à bien ce travail de recherche.

Monsieur le Docteur BISROR Armand,

Je vous remercie de m’avoir accompagné avec bienveillance depuis le début de mes études jusqu'à leur aboutissement. Votre soutien indéfectible, tant sur le plan académique que personnel, a été d'une grande importance pour moi tout au long de ce parcours.

À ma mère,

Je te suis profondément reconnaissant pour tous les efforts que tu as déployés tout au long de mes études afin de m'aider du mieux possible. Les mots ne suffiront jamais à exprimer à quel point tu es une source d'inspiration pour moi. Tu m'as appris la valeur du travail, de la persévérance et du courage face aux défis, des qualités qui m'accompagnent chaque jour. Je suis et resterai ton plus grand admirateur. Merci pour tout ce que tu as fait et continues à faire. J'espère de tout cœur continuer à te rendre fier.

À Yassmine,

Merci de m'avoir soutenu dans les moments les plus difficiles, de m'avoir encouragé à persévérer même quand la motivation me manquait, et de m'avoir toujours fait croire que tout était possible. Ta présence à mes côtés jusqu'au bout a été précieuse.

À Caroline,

Je tiens à te remercier du fond du cœur pour ton soutien tout au long de ma thèse. Sans toi, ce travail n'aurait jamais vu le jour. Tes conseils avisés, ta patience, et toutes les heures que tu as consacrées à m'aider m'ont été d'une aide précieuse. Je suis fier de pouvoir te compter parmi mes proches, et je te suis profondément reconnaissant pour tout ce que tu as fait.

À mon petit frère,

Tu me fais sourire à chaque fois que tu dis fièrement que tu as un grand frère médecin et que tu souhaites suivre le même chemin. Quoi que tu décides pour ton avenir, sache que je serai toujours fier de toi et que je serai là pour te soutenir, peu importe la voie que tu choisiras.

À toute ma famille et à mes amis,

Merci pour votre soutien constant, vos encouragements et votre présence tout au long de ce parcours. Vos gestes, grands ou petits, ont été essentiels pour m'aider à arriver jusqu'ici. Je vous suis infiniment reconnaissant.

Liste des abréviations

CNIL: Commission Nationale de l'Informatique et des Libertés

CPP: Comité de Protection des Personnes

GPT : Generative Pre-trained Transformer

IA : Intelligence Artificielle

IA faible : Intelligence Artificielle Faible

IA forte : Intelligence Artificielle Forte

IAg : Intelligence Artificielle Générative

LLM : Large Language Model (Modèle de Langage de Grande Taille)

QCM: Questionnaire à choix multiple

Sommaire

Introduction

L’intelligence artificielle (IA) est difficile à définir en raison des variations des définitions au fil du temps et qui changent selon les experts. Cette difficulté résulte de l'absence de consensus sur la définition du mot « intelligence » et du vaste domaine en constante évolution qu'est l'IA (1).

Historiquement, Alan Turing (1950), considéré comme le père de l’informatique, a estimé qu’une machine est capable d’intelligence si elle parvient à « duper » un être humain qui la questionne, rendant difficile la distinction entre une machine et un humain. Pour valider cela il a créé le Test de Turing (2).

En 1956, Marvin Minsky et John McCarthy, pionniers de l’IA, la définissent comme « la construction de programmes informatiques accomplissant des tâches que les humains exécutent mieux actuellement, car elles nécessitent des processus mentaux de haut niveau tels que l’apprentissage perceptuel, l’organisation de la mémoire et le raisonnement critique. »

Pour cette thèse, nous définirons l’intelligence comme la capacité à traiter l’information et à interagir avec l’environnement pour atteindre un objectif (3). L’intelligence artificielle est l’accomplissement de cet objectif par une machine.

Aujourd’hui, l'IA révolutionne divers domaines, y compris la médecine promettant des progrès considérables dans la qualité des soins. Les agents conversationnels comme ChatGPT, basés sur l'IA, suscitent un intérêt croissant. L'IA en médecine existe depuis plusieurs années. Le premier système d'IA, « MYCIN », permettait de détecter les infections bactériennes avec une précision comparable à celle des cliniciens (4,5). Depuis, des applications en différentes spécialités ont vu le jour, notamment dans le domaine de l’ophtalmologie aux États-Unis pour détecter la rétinopathie diabétique (6).

Toutes ces applications sont considérées comme des IA « faibles » non autonomes, capables d'effectuer des tâches spécifiques nécessitant souvent une intervention humaine (7). En 2022, des IA « fortes », comme les IA génératives (IAg) basées sur l’agent conversationnel ont émergé. Certains les considérant capables de rivaliser voire surpasser l’intelligence humaine.

L’IAg est une technologie émergente qui se base sur des modèles de langage de grande taille (LLM) pour comprendre et générer des textes en langage naturel, ainsi que d'autres types de contenu tels que la musique, les images et les vidéos (8). Le fonctionnement de l’IAg repose sur l'utilisation de plusieurs algorithmes d'apprentissage automatique. Ils sont entraînés de

manière répétitive sur des bases de données énormes avec des techniques d'apprentissage supervisé et par renforcement (9).

L'apprentissage supervisé fait référence à l'entraînement du modèle sur des données étiquetées, dans ce cas du texte humain, pour lui apprendre à générer du texte similaire. L'apprentissage par renforcement implique un processus itératif où le modèle reçoit des signaux de rétroaction positifs ou négatifs en fonction de la qualité de ses productions, l'incitant à s'améliorer continuellement (10,11). Après avoir analysé ces données, l'IAg est capable de repérer des schémas logiques et de les répliquer, générant ainsi du contenu original basé sur les connaissances acquises. Cette technologie est capable de comprendre et de générer du contenu un peu comme le cerveau humain grâce à son réseau neuronal d'algorithmes.

Cependant, il est important de noter que ces IAg ne sont pas capables de « raisonner » comme un humain. Elles fonctionnent en déterminant la probabilité statistique qu'un mot suive un autre, produisant ainsi des phrases qui sont statistiquement les plus probables en fonction des données sur lesquelles elles ont été entraînées.

Bien que les recherches sur l'utilisation de l'IAg en soins primaires soient encore limitées en raison de leur nouveauté, des études existantes suggèrent un potentiel significatif pour améliorer la qualité des soins. Par exemple, ces technologies pourraient répondre aux questions courantes des patients, facilitant la compréhension de l'information médicale (12). Pour les soignants, elles peuvent déjà aider à la rédaction de comptes-rendus divers (13).

La question du possible remplacement des médecins par ces technologies a été soulevée à plusieurs reprises dans les médias (14–16). Cet engouement est dû à la capacité de certaines IA à valider des examens d'études de médecine et à répondre à des questions médicales. Des problèmes ont aussi été soulevés concernant l'utilisation de ces technologies, comme le phénomène « d'hallucination », où l'IA invente de l'information lorsqu'il ne connaît pas la réponse (9).

Ce que l'on ne sait pas aujourd'hui, c'est la fiabilité des réponses fournies en pratique clinique par ces IAg, dont l'évaluation approfondie, notamment en soins primaires, reste à faire.

Quelle est la capacité de ChatGPT à répondre de manière fiable à des questions médicales sur l'asthme ? Une réponse est fiable si elle est exacte et cohérente. L'évaluation de la fiabilité des réponses de cette IAg pourrait déterminer si ces technologies peuvent effectivement améliorer l'efficacité du travail médical en soins primaires, particulièrement en ce qui concerne la prise en charge de l'asthme, un problème de santé publique en France dont les médecins généralistes sont majoritairement la première ligne de prise en charge (17).

1 Matériel et méthode

1.1 Plan Expérimental

Cette étude est une étude observationnelle descriptive quantitative visant à évaluer la fiabilité de ChatGPT, crée par l'entreprise OpenAI, sur des questions médicales relatives à l'asthme.
L'analyse couvre deux versions de ChatGPT (3.5 et 4.0) à deux moments différents (mars 2023 et mars 2024) pour déterminer s'il y a une évolution dans les réponses fournies par ces outils qui sont régulièrement mis à jour.
La version 3.5 de ChatGPT étant gratuite et accessible à tout moment, et la version 4.0 étant payante (20euros par mois) avec une limite du nombre de questions posées par jour.

1.2 Autorisations Réglementaires

Cette étude n'implique pas d'interventions humaines directes ni de collecte de données personnelles sensibles. Ainsi, elle ne nécessite pas d'accord du Comité de Protection des Personnes (CPP) ni de déclaration auprès de la CNIL. Toutefois, elle respecte les principes éthiques de la recherche en médecine.

1.3 Objectif principal

L'objectif principal de cette étude est d'évaluer la fiabilité de ChatGPT 3.5 et 4.0 dans la résolution des questions sur l'asthme.

1.4 Objectif secondaire

L'objectif secondaire est d'évaluer l'évolution des réponses des deux versions de ChatGPT entre mars 2023 et mars 2024.

1.5 Outils d'évaluation

Les questions utilisées dans cette étude sont extraites de l'application de QCM médicaux « PneumoQuiz ® » créé par le Dr Fabien ROLLAND. Ces questions sont spécifiquement conçues pour des professionnels de santé tels que des étudiants en médecine, des internes, et des médecins. Chaque question est référencée, c'est-à-dire que pour chaque réponse, une source est fournie. Chaque question a été par ailleurs relue et corrigée par un expert international : le Pr Pascal CHANEZ.

1.6 Critères d'évaluation

Les réponses fournies par ChatGPT aux questions posées ont été classées en deux catégories : « réponse correcte » ou « réponse incorrecte ».
Pour les questions d'ordre numérique, dont la réponse est en chiffres ou en pourcentage, nous avons décidé qu'une marge d'erreur de dix% serait acceptée. Cette marge a été choisie de façon à permettre une tolérance raisonnable compte tenu des variations possibles dans les réponses générées par les modèles d'IA.

1.7 Échantillonnage et données collectées

Au total, parmi les 587 questions sur l'asthme présentes dans l'application « PneumoQuiz ® », 556 questions ont été posées à ChatGPT versions 3.5 et 4.0 à deux moments différents (mars 2023 et mars 2024) totalisant ainsi 1112 interactions en 2023 et 1112 en 2024.
31 questions ont été exclues car elles étaient en double.

Les questions sont classées dans les groupes suivants :

- « Épidémiologie et facteurs de risque »
- « Physiopathologie »
- « Diagnostic et examens complémentaires »
- « Prise en charge et traitement »
- « Comorbidités, complications et éducation thérapeutique »
- « Allergènes »

1.8 Plan d'analyse statistique

Les données collectées sont enregistrées dans un document Excel.
L'analyse statistique comprend plusieurs étapes. Tout d'abord, une analyse descriptive qui est le calcul des pourcentages de réponses correctes et incorrectes pour chaque version de ChatGPT et pour chaque groupe de questions. Ensuite, il y a une comparaison de la fiabilité des versions 3.5 et 4.0 de ChatGPT en termes de pourcentage de réponses correctes et incorrectes. Pour terminer, il y a une évaluation de l'évolution temporelle des réponses entre mars 2023 et mars 2024 pour chaque version de ChatGPT.
Un test du Chi 2 est utilisé pour comparer les taux de bonnes et mauvaises réponses d'une version à l'autre.

2 Résultats

2.1 Fiabilité de ChatGPT 3.5 et ChatGPT 4.0

Tableau 1. Description de la précision de chaque version de Chat GPT

Catégorie de réponse	Version 3.5 2023	Version 4.0 2023	Version 3.5 2024	Version 4.0 2024
Allergènes	21 (77,8%)	19 (70,4%)	18 (66,7%)	21 (77,8%)
Comorbidités, complications, éducation thérapeutique	50 (54,3%)	55 (59,8%)	47 (51,1%)	54 (58,7%)
Diagnostic et examens complémentaires	53 (72,6%)	58 (79,5%)	54 (74%)	60 (82,2%)
Epidémiologie et facteurs de risque	39 (58,2%)	44 (65,7%)	41 (61,2%)	54 (80,6%)
Physiopathologie	111 (77,6%)	132 (92,3%)	119 (83,2%)	133 (93%)
Prise en charge et traitement	114 (74%)	123 (79,9%)	113 (73,4%)	131 (85,1%)
Ensemble	**388 (69,8%)**	**431 (77,5%)**	**392 (70,5%)**	**453 (81,5%)**

2.1.1 Fiabilité globale

Le tableau 1 présente le nombre et le taux de bonnes réponses par version et par catégorie de questions. Il compare le nombre de bonnes réponses de la version 3.5 (2023) à la version 4.0 (2023) de ChatGPT, puis nous propose l'évolution de ces résultats en 2024. Au total, 556 questions ont été incluses dans cette analyse.

Pour l'ensemble des questions, la version 3.5 de ChatGPT en 2023 procure 388 bonnes réponses, soit un taux de réussite de 69,8 %. Pour la même version en 2024, le taux de bonnes réponses atteint 70,5 %.

Concernant la version 4.0 de ChatGPT en 2023, elle offre 431 réponses satisfaisantes, ce qui représente un taux de réussite de 77,5 %. Ces chiffres passent à 453 bonnes réponses pour la version 4.0 en 2024, soit un taux de réponse positive de 81,5 %.

2.1.2 Fiabilité par catégorie

Pour l'analyse par catégories, il y avait au total 154 questions dans la catégorie « Prise en charge et traitement », 143 dans la catégorie « Physiopathologie », 92 dans la catégorie «Comorbidités, complications et éducation thérapeutique », 73 dans la catégorie « Diagnostic et examens complémentaires », 67 dans la catégorie « Épidémiologie et facteurs de risque », et 27 questions dans la catégorie « Allergènes ».

La version ChatGPT 3.5 a obtenu le meilleur taux de réponses dans la catégorie « Allergènes » en 2023, avec 77,8 % de bonnes réponses, et en 2024 dans la catégorie « Physiopathologie », avec 83,2 % de bonnes réponses.
Les moins bons résultats par catégorie pour cette même version en 2023 et en 2024 ont été obtenus dans le groupe « Comorbidités, complications et éducation thérapeutique », avec 54,3 % de bonnes réponses en 2023 et 51,1 % en 2024.
Pour la version ChatGPT 4.0, les meilleurs résultats ont été obtenus dans la catégorie « Physiopathologie », avec un taux de bonnes réponses de 92,3 % en 2023 et 93 % en 2024.
Concernant cette même version, les résultats les moins favorables sont dans le groupe « Comorbidités, complications et éducation thérapeutique », avec 59,8 % de bonnes réponses en 2023 et 58,7 % en 2024 pour cette même catégorie.

2.2 Evaluation de l'évolution de la fiabilité des deux versions de ChatGPT

2.2.1 Evolution du taux de bonnes réponses d'une version à l'autre

Tableau 2. Comparaison de la fiabilité entre la version ChatGPT 3.5 et 4.0 la même année pour l'ensemble des questions

Version	Bonnes réponses V3.5	Bonnes réponses V4.0	P-valeur	Bonne réponse suivie d'une bonne	Bonne réponse suivie d'une mauvaise réponse	Mauvaise réponse suivie d'une bonne réponse	Mauvaise réponse suivie d'une mauvaise
V 3.5 vs 4.0 en 2023	388 (69,8%)	431 (77,5%)	< 0,01***	358 (64,4%)	30 (5,4%)	73 (13,1%)	95 (17,1%)
V 3.5 vs 4.0 en 2024	392 (70,5%)	453 (81,5%)	< 0,01***	374 (67,3%)	18 (3,2%)	79 (14,2%)	85 (15,3%)

Le tableau 2 présente l'évolution des réponses d'une version à une autre pour la même année.
La version 3.5 de ChatGPT donne 388 bonnes réponses, soit un pourcentage de 69,8 % de bonnes réponses en 2023, contre 431 bonnes réponses, soit un pourcentage de 77,5 % pour la version 4.0 la même année. Cet écart de taux de bonnes réponses de 7 % constitue une différence statistiquement significative au seuil d'un %.
On remarque qu'en 2023, parmi ces 388 bonnes réponses données par la version 3.5, la version 4.0 a répondu également de manière correcte pour 358 d'entre elles. Seules 30 questions qui avaient été bien répondues par la version 3.5 ont été mal répondues avec la version 4.0. Toujours en 2023, la version 4.0 a répondu de manière correcte à 73 questions alors que la version 3.5 avait répondues de manière erronée (soit 13,1 % des questions totales parmi les 556).

En 2024, la version 3.5 donne 392 bonnes réponses (70,5 %) contre 453 bonnes réponses (81,5 %) pour la version 4.0, soit une différence de 11 %. Cette différence est également statistiquement significative au seuil d'un %.

En 2024, l'évolution du taux de bonnes réponses est améliorée. En effet, parmi les 392 bonnes réponses données par la version 3.5, la version 4.0 répond également de manière positive à 374 questions, soit seulement 18 mauvaises réponses. La version 4.0 répond de manière correcte à 79 questions auxquelles la version 3.5 a répondu de manière erronée.

2.2.2 Evolution du taux de bonnes réponses d'une année à l'autre

Tableau 3. Evolution des réponses d'une année à l'autre pour une même version

Version	Bonnes réponses en 2023	Bonnes réponses en 2024	P-valeur	Bonne réponse suivie d'une bonne	Bonne réponse suivie d'une mauvaise réponse	Mauvaise réponse suivie d'une bonne réponse	Mauvaise réponse suivie d'une mauvaise
V3.5 en 2023 vs en 2024	388 (69,8%)	392 (70,5%)	0,844	346 (62,2%)	42 (7,6%)	46 (8,3%)	122 (21,9%)
V4.0 en 2023 vs en 2024	431 (77,5%)	453 (81,5%)	0,119	407 (73,2%)	24 (4,3%)	46 (8,3%)	79 (14,2%)

La version 3.5 de ChatGPT donne en 2023 69,8 % de bonnes réponses contre 70,5 % en 2024. Cet écart de taux de bonnes réponses constitue une différence de zéro virgule sept%, non significative au seuil de cinq % ($p = 0,844$).

Pour la version 4.0, le pourcentage de bonnes réponses en 2023 est de 77,5 % contre 81,5 % en 2024. Cet écart de taux de bonnes réponses constitue une différence de 3.6%, non significative au seuil de cinq % ($p = 0,119$).

On remarque que pour la version 3.5, 42 questions initialement bien répondues en 2023 ont été mal répondues par la même version un an plus tard, et 46 questions initialement mal répondues ont été bien répondues l'année suivante.

On observe cette même tendance avec la version 4.0 pour les questions initialement mal répondues qui ont été bien répondues l'année suivante (46 questions). Cependant, le nombre de questions initialement bien répondues en 2023 qui ont évolué en mauvaise réponse en 2024 est plus faible, soit 24 questions au total.

3 Discussions

3.1 Résultat principal

Cette étude révèle que, sur l'asthme, l'outil d'intelligence artificielle CHATGPT dans ses versions 3.5 et 4.0 donne des réponses fiables. La version 4.0 est d'avantage fiable.
Ces deux versions sont issues d'un même modèle, GPT, et ont été entrainées sur la même base de données. Au moment du recueil des données de notre étude en 2023 et en 2024, les informations de la base de données n'était mise à jour que jusqu'en septembre 2021.
Cependant, les deux versions diffèrent par la taille et le fonctionnement de leur modèle. En effet, la version 4.0 a un modèle plus large que celui de la version 3.5. La version 4.0 bénéficierait d'améliorations dans l'apprentissage par renforcement, qui optimisent continuellement la qualité des réponses en réduisant les erreurs, notamment les phénomènes d'"hallucination" où l'IA invente des informations incorrectes (9). Ces différences pourraient donc être à l'origine de la supériorité en termes de fiabilité de la version 4.0.

3.2 Confrontation des résultats aux données connues de la littérature (validité externe)

Nous allons confronter nos résultats aux données de la littérature. Cependant, aucune étude n'a encore évalué la fiabilité de ChatGPT spécifiquement sur l'asthme ou une autre maladie particulière. Nous nous appuierons donc sur des études qui ont analysé la fiabilité de ChatGPT lors d'examens médicaux généraux et de spécialités, afin de contextualiser nos observations.

3.2.1 Examens généraux

Les résultats de notre étude sont en accord avec ceux de la littérature qui évaluent la fiabilité des différentes versions de ChatGPT lors d'examens médicaux dans divers contextes à travers le monde. Aux États-Unis, la version 3.5 de ChatGPT a validé l'examen médical (USMLE Steps 1,2 et 3) dans deux études distinctes. Dans l'une de ces études, les questions étaient posées en QCM puis en question ouvertes. Pour les QCM, ChatGPT 3.5 a obtenu environ 60%, et pour les questions ouvertes, environ 70%, des résultats similaires à ceux de la seconde étude (18,19). Des résultats comparables ont été observés au Chili, où ChatGPT 3.5 et 4.0 ont validé l'examen d'internat EUNACOM, avec une meilleure performance pour la version 4.0 (20). Ces résultats sont également cohérents avec ceux d'études menées au Japon, en Pologne et en Arabie Saoudite (21–23) .

3.2.2 Examens de spécialité

Dans les études évaluant la fiabilité de ChatGPT sur des spécialités médicales, les résultats sont moins prometteurs. Par exemple, au Japon, la version 3.5 a échoué à un examen d'imagerie médicale, obtenant un score de 40% ($p = 0,013$) (24). De même, deux études polonaises montrent que ChatGPT a échoué aux examens d'internat en néphrologie (45,70%, $P < .0001$) et en médecine interne (49,71%, $P < .0001$).(25,26). L'analyse par catégories de questions révèle que ChatGPT 3.5 a obtenu son meilleur score (71,43%) en allergologie, cohérent avec nos résultats, où cette version performe bien dans ce domaine.

Dans une autre étude comparant ChatGPT 3.5 et 4.0 avec des internes en chirurgie orthopédique aux États-Unis, les deux modèles ont échoué, sous-performant de plus de 40% par rapport aux internes. ChatGPT 3.5 a obtenu 29,4% (IC = 23,23%-36,4%) et ChatGPT 4.0, 47,2% (IC = 40%-54,5%) (27).

Globalement, ces résultats ne concordent pas avec les résultats de notre étude en ce qui concerne la fiabilité de ChatGPT 3.5. Dans ces études d'examens de spécialité ChatGPT 3.5 est peu fiable.

3.3 Discussion des données non significatives et/ou divergentes

L'objectif secondaire de cette étude est d'évaluer l'évolution temporelle des réponses fournies par ChatGPT en 2023 et 2024, afin de déterminer si ces outils d'IA s'améliorent ou non avec le temps. Toutefois, nos résultats ne permettent pas de tirer des conclusions définitives à ce sujet car les différences observées entre les deux années ne sont pas statistiquement significatives.

En conséquence, nous ne pouvons pas affirmer que les réponses de ChatGPT sont cohérentes d'une année sur l'autre ni conclure que ces modèles s'améliorent ou régressent significativement avec le temps.

3.4 Exposition des biais et limites de l'étude (validité interne)

3.4.1 Analyse et interprétation des données

L'analyse et l'interprétation des données ont été réalisées par une seule personne dans cette étude, ce qui soulève la question d'un potentiel biais d'interprétation. De plus, un biais de confirmation a pu survenir, car les réponses fournies par PneumoQuiz ont été systématiquement considérées comme la référence sans vérification approfondie en cas de discordance avec ChatGPT. Cette absence de vérification des sources pourrait avoir influencé l'évaluation de la fiabilité de ChatGPT, surtout si certaines erreurs provenaient des questions elles-mêmes.

3.4.2 Variabilité du nombre de questions par catégorie

Le nombre limité de questions dans certaines catégories, notamment "allergènes" qui compte moins de 30 questions, restreint l'évaluation des connaissances de ChatGPT sur ce sujet.

3.4.3 Sources de la base de données de ChatGPT

Une limite de notre étude réside dans le fait que les deux versions de ChatGPT incluent des informations mises à jour jusqu'en septembre 2021, alors que certaines questions reposent sur des sources publiées après cette date, ce qui pourrait expliquer certaines erreurs. De plus, certaines questions sont basées sur des articles accessibles via des plateformes payantes, soulevant la question de l'accès de ChatGPT à ces sources lors de son entraînement. Enfin, bien qu'OpenAI établisse des partenariats pour obtenir des données exclusives, ces collaborations récentes ne concernaient pas les études académiques disponibles au moment de l'entraînement initial de ChatGPT.

3.4.4 Mise à jour de la base de données

La base de données pour ChatGPT 4.0 a été mise à jour depuis la réalisation du recueil de données de notre étude, les résultats obtenus peuvent donc ne plus être à jour pour cette version. Il serait intéressant d’évaluer la fiabilité de ChatGPT 4.0 depuis sa nouvelle mise à jour.

3.4.5 L’évaluation clinique

Bien que cette étude ait révélé que ChatGPT peut fournir des réponses fiables sur l'asthme dans certaines catégories de questions, elle n'évalue pas la capacité de cet outil à être utilisé comme un assistant quotidien par les médecins généralistes. Cette dimension est pourtant cruciale pour juger de l'utilité clinique de ChatGPT.

3.4.6 Phénomène d'hallucination non évalué

Cette étude n'a pas évalué le phénomène d'hallucination, où ChatGPT pourrait générer des informations incorrectes ou inventées, ce qui est un problème bien documenté dans l'utilisation des modèles de langage (9,28). L'absence de cette évaluation signifie que la fiabilité des réponses observées pourrait être surévaluée, car les cas d'hallucination n'ont pas été systématiquement identifiés et comptabilisés dans cette étude.

3.5 Indications des points forts des résultats

Cette étude se distingue par le fait qu'elle est la première à évaluer de manière approfondie les connaissances de ChatGPT sur l'asthme, en utilisant un large échantillon de questions centrées

sur un même sujet. Contrairement à d'autres recherches qui se sont concentrées sur plusieurs spécialités ou ont utilisé un nombre limité de questions, cette étude fournit une analyse détaillée et exhaustive des capacités de ChatGPT dans un domaine médical spécifique.

3.6 Conclusion avec perspectives du travail

L'objectif de cette thèse était d'évaluer la fiabilité des réponses de ChatGPT 3.5 et 4.0 sur des questions médicales en médecine générale, en particulier sur l'asthme. Ces questions proviennent de la base de données de l'application de QCM médicaux « PneumoQuiz® ».
Nous avons mené une analyse observationnelle descriptive quantitative. Les résultats montrent que ChatGPT 3.5 est fiable pour environ 70 % des questions, tandis que la version 4.0 l'est pour environ 80 %, avec de meilleures performances globales. Ces résultats suggèrent que ChatGPT, notamment 4.0, est un outil potentiellement fiable pour répondre aux questions médicales sur l'asthme pour les professionnels de santé. Toutefois, il est crucial de rester conscient de ses limites et du risque d'erreurs. Par conséquent, il est prématuré de recommander l'utilisation de cette IA en pratique clinique dans son état actuel. Des recherches supplémentaires sont nécessaires pour évaluer son acceptabilité et son utilisation en cabinet de médecine générale.
Une étude future pourrait explorer ces aspects en profondeur, en se concentrant sur l'impact réel de ChatGPT dans la pratique quotidienne. Il serait également pertinent de définir un seuil de performance avant la collecte des données dans les recherches futures, afin d'éviter tout biais d'ajustement post-hoc. Cela garantirait une évaluation plus rigoureuse de la fiabilité de ChatGPT dans un cadre clinique.
De plus, bien que cette étude ait montré certaines limites de ChatGPT, l'Académie Nationale de Médecine, dans sa séance du cinq mars 2024, a recommandé l'apprentissage de l'usage des systèmes d'IA par les professionnels de santé, soulignant qu'il serait contraire à l'éthique de se priver de ces outils dans les pratiques médicales(10).
Cependant, cette recommandation n'a pas précisé quel type d'IA ni les conditions d'utilisation, renforçant la nécessité d'évaluations supplémentaires avant d'intégrer des outils comme ChatGPT de manière généralisée. Enfin, bien que cette étude se concentre sur l'évaluation de ChatGPT en tant qu'outil pour les professionnels de santé, il serait pertinent d'envisager son utilisation comme support pour la population générale, notamment via des chatbots médicaux, afin d'évaluer si l'IA peut fournir des réponses adaptées et compréhensibles par un public non médical.

Références

1. Wang P. On Defining Artificial Intelligence. J Artif Gen Intell. 1 janv 2019;10(2):1-37.

2. TURING AM. I.—COMPUTING MACHINERY AND INTELLIGENCE. Mind. 1 oct 1950;LIX(236):433-60.

3. McCarthy J. WHAT IS ARTIFICIAL INTELLIGENCE? , Stanford, CA 94305, 2007

4. Shortliffe E. H. (1977). Mycin: A Knowledge-Based Computer Program Applied to Infectious Diseases. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 66–69.

5. Mycin. In: Wikipedia. 2023. Disponible sur: https://en.wikipedia.org/w/index.php?title=Mycin&oldid=1166772713

6. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med. 2018;1:39.

7. Flowers, J.C. Strong and Weak AI: Deweyan Considerations. AAAI Spring Symposium: Towards Conscious AI Systems., 2019

8. What Are Large Language Models (LLMs) | IBM [Internet]. 2023 Disponible sur: https://www.ibm.com/topics/large-language-models

9. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 Technical Report . arXiv; 2024 . Disponible sur: http://arxiv.org/abs/2303.08774

10. 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.

11. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al. Training language models to follow instructions with human feedback [Internet]. arXiv; 2022. Disponible sur: http://arxiv.org/abs/2203.02155

12. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol. juill 2023;29(3):721-32.

13. Liu J, Wang C, Liu S. Utility of ChatGPT in Clinical Practice. J Med Internet Res. 28 juin 2023;25:e48568.

14. RMC . Intelligence artificielle: ChatGPT, plus fort que les médecins? Disponible sur: https://rmc.bfmtv.com/actualites/societe/sante/intelligence-artificielle-chat-gpt-plus-fort-que-les-medecins_AV-202305030620.html

15. Hardy F. Site-LeJournalDuMedecin-FR. 2023 ChatGPT va-t-il remplacer le diagnostic des médecins ? Disponible sur: https://www.lejournaldumedecin.com/actualite/chatgpt-va-t-il-remplacer-le-diagnostic-des-medecins/article-normal-69827.html

16. Rothéa C. AlloDocteurs. 2023. ChatGPT, une intelligence artificielle prête à remplacer les médecins ? Disponible sur: https://www.allodocteurs.fr/chatgpt-une-intelligence-artificielle-prete-a-remplacer-les-medecins-34772.html

17. Raherison-Semjen C, Izadifar A, Russier M, Rolland C, Aubert JP, Touboul C, et al. Self-reported asthma prevalence and management in adults in France in 2018: ASTHMAPOP survey. Respir Med Res. nov 2021;80:100864.

18. Kung TH, Cheatham M, Medenilla A, Sillos C, Leon LD, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 9 févr 2023;2(2):e0000198.

19. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 8 févr 2023;9:e45312.

20. Rojas M, Rojas M, Burgess V, Toro-Pérez J, Salehi S. Exploring the Performance of ChatGPT Versions 3.5, 4, and 4 With Vision in the Chilean Medical Licensing Examination: Observational Study. JMIR Med Educ. 29 avr 2024;10:e55048-e55048.

21. Kasai, Jungo & Kasai, Yuhei & Sakaguchi, Keisuke & Yamada, Yutaro & Radev, Dragomir. (2023). Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations. 10.48550/arXiv.2303.18027.

22. Aljindan FK, Al Qurashi AA, Albalawi IAS, Alanazi AMM, Aljuhani HAM, Falah Almutairi F, et al. ChatGPT Conquers the Saudi Medical Licensing Exam: Exploring the Accuracy of Artificial Intelligence in Medical Knowledge Assessment and Implications for Modern Medical Education. Cureus. 15(9):e45043.

23. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. Sci Rep. 22 nov 2023;13:20512.

24. Toyama Y, Harigai A, Abe M, Nagano M, Kawabata M, Seki Y, et al. Performance evaluation of ChatGPT, GPT-4, and Bard on the official board examination of the Japan Radiology Society. Jpn J Radiol. févr 2024;42(2):201-7.

25. Nicikowski J, Szczepański M, Miedziaszczyk M, Kudliński B. The potential of ChatGPT in medicine: an example analysis of nephrology specialty exams in Poland. Clin Kidney J. 22 juin 2024;17(8):sfae193.

26. Suwała S, Szulc P, Dudek A, et al. ChatGPT fails the Polish board certification examination in internal medicine: artificial intelligence still has much to learn. Pol Arch Intern Med. 2023; 133: 16608. doi:10.20452/pamw.16608

27. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT–3.5, ChatGPT-4, and Orthopaedic Resident Performance on Orthopaedic Assessment Examinations. JAAOS - J Am Acad Orthop Surg. 1 déc 2023;31(23):1173.

28. Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. Cureus. févr 2023;15(2):e35179.

Serment d'Hippocrate

Au moment d'être admis(e) à exercer la médecine, je promets et je jure d'être fidèle aux lois de l'honneur et de la probité.

Mon premier souci sera de rétablir, de préserver ou de promouvoir la santé dans tous ses éléments, physiques et mentaux, individuels et sociaux.

Je respecterai toutes les personnes, leur autonomie et leur volonté, sans aucune discrimination selon leur état ou leurs convictions.

J'interviendrai pour les protéger si elles sont affaiblies, vulnérables ou menacées dans leur intégrité ou leur dignité.

Même sous la contrainte, je ne ferai pas usage de mes connaissances contre les lois de l'humanité.

J'informerai les patients des décisions envisagées, de leurs raisons et de leurs conséquences.

Je ne tromperai jamais leur confiance et n'exploiterai pas le pouvoir hérité des circonstances pour forcer les consciences.

Je donnerai mes soins à l'indigent et à quiconque me les demandera. Je ne me laisserai pas influencer par la soif du gain ou la recherche de la gloire.

Admis(e) dans l'intimité des personnes, je tairai les secrets qui me seront confiés. Reçu(e) à l'intérieur des maisons, je respecterai les secrets des foyers et ma conduite ne servira pas à corrompre les mœurs.

Je ferai tout pour soulager les souffrances. Je ne prolongerai pas abusivement les agonies. Je ne provoquerai jamais la mort délibérément.

Je préserverai l'indépendance nécessaire à l'accomplissement de ma mission. Je n'entreprendrai rien qui dépasse mes compétences. Je les entretiendrai et les perfectionnerai pour assurer au mieux les services qui me seront demandés.

J'apporterai mon aide à mes confrères ainsi qu'à leurs familles dans l'adversité.

Que les hommes et mes confrères m'accordent leur estime si je suis fidèle à mes promesses ; que je sois déshonoré(e) et méprisé(e) si j'y manque.

Résumé

Introduction : L'intelligence artificielle (IA) présente un potentiel important pour soutenir divers domaines, y compris la médecine. Cependant, la fiabilité des réponses fournies par ces systèmes dans un contexte clinique nécessite une évaluation approfondie. Cette étude évalue la fiabilité de ChatGPT sur des questions relatives à l'asthme, une maladie respiratoire chronique, principalement prise en charge par les médecins généralistes en France.

Objectif : Évaluer la fiabilité des réponses de ChatGPT 3.5 et 4.0 sur des questions médicales liées à l'asthme et analyser l'évolution de leur qualité entre 2023 et 2024

Matériel et méthode : Étude observationnelle descriptive quantitative utilisant des questions de l'application « PneumoQuizz® » posées à ChatGPT 3.5 et 4.0 en mars 2023 et mars 2024. Les réponses ont été classées en « correctes » ou « incorrectes » et comparées statistiquement par test du Chi 2.

Résultats : La version 3.5 de ChatGPT a obtenu un taux de bonnes réponses de 69,8 % en 2023 et 70,5 % en 2024. La version 4.0 a montré des taux de 77,5 % en 2023 et 81,5 % en 2024. Bien que la version 4.0 soit globalement plus fiable que la 3.5, les différences entre les résultats de 2023 et 2024 ne sont pas statistiquement significatives.

Discussion : Les résultats suggèrent que ChatGPT 4.0 est plus fiable que la version 3.5, mais aucune amélioration significative dans le temps n'a été observée. Des études supplémentaires, notamment dans d'autres domaines médicaux, sont nécessaires avant de pouvoir recommander l'usage de ChatGPT en pratique clinique quotidienne.

Mots-clés : intelligence artificielle, ChatGPT, fiabilité, asthme, médecine générale