



RAPPORT DE STAGE DE FIN D'ÉTUDES

Département : MIASHS UFR 6 Informatique, Mathématique et Statistique

Effectué du **05 mai 2025** au **27 juin 2025**

SUJET DE STAGE

Évaluation de l'interface d'un chatbot et de la simplification des textes médicaux par des modèles de langage à destination des patients.



Réalisée par
HAKIRI Siwar

Enseignant Référent

Nom : DEMANGEOT Marine
Fonction : Responsable L3
marine.demangeot@univ-montp3.fr

Tutrice de stage

Nom : PIRES Rita
Fonction : assistante de recherche technologique
rita.pires@chu-montpellier.fr

Juin 2025

Remerciements

Je tiens à exprimer ma profonde reconnaissance à toute l'équipe ERIOS pour l'accueil chaleureux et la bienveillance dont j'ai bénéficié durant mon stage. Leur soutien constant, leur expertise et leur disponibilité ont été essentiels pour m'accompagner tout au long de cette expérience, qui fut riche en apprentissages et en découvertes.

Je remercie sincèrement ma tutrice de stage, Madame Rita Pires, pour son engagement constant et son accompagnement bienveillant tout au long de cette expérience. Sa disponibilité et ses conseils avisés m'ont aidé à mieux comprendre les différentes étapes du projet. Grâce à son soutien, j'ai pu surmonter les obstacles et développer mes compétences. Sa rigueur et son écoute attentive ont grandement contribué à la qualité de mon travail.

Je souhaite également remercier Madame Louise Robert, dont son expertise et sa rigueur scientifique m'ont permis de développer un regard critique approfondi, notamment sur les dimensions linguistiques et quantitatives de la recherche. Grâce à son encadrement, j'ai pu mieux comprendre les différentes phases préparatoires nécessaires avant d'engager un projet, ce qui a largement enrichi ma méthodologie.

Mes remerciements s'adressent aussi à l'équipe pédagogique de l'UFR 6 de l'Université Paul-Valéry et au corps enseignant de la licence MIASHS, pour leur soutien et la qualité de leur encadrement. Je tiens à remercier particulièrement Madame Marine Demangeot, ma tutrice universitaire, pour sa disponibilité constante, ses conseils avisés et son accompagnement tout au long de cette année, qui ont constitué un atout précieux. Mes remerciements vont aussi à Monsieur Jérôme Pasquet, en charge de l'évaluation de ce rapport, pour son investissement.

Par ailleurs, je suis profondément reconnaissante envers mes parents, dont le soutien inconditionnel, les encouragements et l'amour ont été une source de force tout au long de mon parcours. Leur confiance en mes capacités m'a permis de surmonter les obstacles et de poursuivre mes objectifs avec détermination.

Je souhaite également exprimer ma sincère gratitude envers ma sœur, dont l'aide considérable, les conseils avisés et le soutien moral m'ont permis d'avancer avec confiance. Elle a toujours été un pilier et une source d'encouragement tout au long de mon parcours.

Enfin, je tiens à remercier mes amis, qui, malgré la distance, ont toujours su m'apporter un soutien constant et des encouragements précieux. Leurs présence, même à distance, a été une véritable source de motivation et de réconfort tout au long de cette aventure.

Table des matières

Introduction	4
Chapitre 1 : Présentation d'ERIOS	5
1.1 Genèse du projet ERIOS	5
1.2 Objectifs et missions d'ERIOS	7
1.3 Présentation d'ERIOS assistant	10
Chapitre 2 : Problématique du stage	13
2.1 Contexte spécifique du stage	13
2.2 Enjeux et objectifs des missions confiées	13
2.3 Problématique étudiée :	14
Chapitre 3 : L'IA générative et les modèles de langage dans le secteur de la santé	15
3.1 Définition de l'IA	15
3.2 Applications dans le domaine de la santé	15
3.3 Définition de l'IA générative et LLM :	16
3.3.1 Définition de l'IA générative :	16
3.3.2 Définition d'un LLM (Large Language Models)	17
3.4 Applications de l'IA générative et LLM dans le domaine de santé :	18
3.5 Présentation des LLM utilisés dans les projet ERIOS	20
3.6 Choix du modèle de LLM pour ERIOS Assistant	20
Chapitre 4 : Évaluation de l'interface chatbot et de la simplification des textes médicaux	22
4.1 Méthodologie appliquée à ERIOS assistant	22
4.2 Évaluation de l'interface	24
4.2.1 Évaluation heuristique par des experts :	24
4.2.2 Tests utilisateurs et retours qualitatifs :	28
4.3 Analyse comparative de la simplification des textes médicaux :	35
4.3.1 Tests préliminaires et ajustements :	35
4.3.2 Modèles finaux retenus :	36
4.3.3 Méthodes d'évaluation :	36
4.3.4 Résultats obtenus et interprétations :	37
Chapitre 5 : Conclusion et Perspectives	43
5.1 Limites du travail réalisé	43
5.2 Travaux en cours et évolutions possibles :	43
5.3 Conclusion :	44
Bibliographie	45
Annexes	47

Introduction

Au cours de mon parcours universitaire, j'ai toujours été attirée par les domaines liées à la science des données et plus particulièrement le domaine médical. Cette curiosité m'a orienté à postuler pour un stage afin de finaliser ma licence au sein d'ERIOS (Espace de Recherche et d'Intégration des Outils numériques en Santé), une structure reconnue pour son approche multidisciplinaire alliant recherche fondamentale et appliquée. Les travaux menés au sein de cet établissement mobilisent notamment l'intelligence artificielle, les modèles de langage de grande taille (LLM), accompagné de méthodes qualitatives, dans l'objectif de développer des solutions innovantes au service du secteur médical.

C'est dans ce cadre que s'inscrit le projet auquel j'ai pris part durant mon stage, centré sur l'évaluation d'un chatbot conversationnel destiné à générer des lettres médicales simplifiées à l'intention des patients et des professionnels de santé. Mon implication a porté dans un premier temps sur l'analyse des retours qualitatifs issus des tests utilisateurs et sur l'évaluation des performances des différents modèles de langage mobilisés.

Le déroulement du stage a été structuré en plusieurs phases, avec une répartition claire des tâches au fil des semaines. Chaque semaine, deux sessions de tests utilisateurs étaient organisées, permettant de collecter des retours sur l'ergonomie de l'interface et sur la pertinence des simplifications produites par le chatbot. En parallèle, des évaluations menées par des experts UX ont été organisées sur trois semaines. J'ai donc pu contribué à la collecte et à l'analyse des résultats issus de ces différentes évaluations. Une dernière session avec un expert UX est d'ailleurs prévue la semaine prochaine pour finaliser cette phase.

En ajout à ces tests, une autre partie du travail a porté sur l'analyse poussée des documents médicaux générés. Pour ce faire, nous avons adopté une stratégie de cas par cas, afin d'adapter explicitement les méthodes d'évaluation aux spécificités de chaque texte. Les deux premières semaines du stage ont été consacrées à l'expérimentation des prompts et à la définition des critères d'analyse comparée des LLM. Une fois cette phase préparatoire est achevée, nous avons débuté l'analyse approfondie du premier cas, puis poursuivi progressivement avec les suivants. À ce jour, seulement deux cas restent à analyser pour finaliser l'évaluation comparative.

Ce rapport de stage a pour objectif de présenter le cadre et la problématique du projet, d'expliquer clairement les différentes étapes du travail réalisé durant le stage, et d'analyser les résultats issus des phases de test et d'évaluation. Il cible aussi à mettre en valeur l'importance d'une approche méthodologique mixte, reliant analyses qualitatives et quantitatives, dans l'optique de répondre aux défis de qualité linguistique et d'accessibilité des contenus médicaux produits par intelligence artificielle.

Chapitre 1 : Présentation d'ERIOS

1.1 Genèse du projet ERIOS

Depuis le début des années 2000, le secteur hospitalier connaît une véritable transformation numérique, marquée par l'introduction massive des dossiers patients informatisés (DPI) dans les établissements de santé. L'objectif de cette transition est de rendre la gestion des informations médicales plus moderne afin d'optimiser la qualité, la sécurité, l'accessibilité et la continuité des soins. Au CHU de Montpellier, comme dans de nombreux établissements hospitaliers français, cette évolution s'est traduite par l'implémentation progressive de solutions numériques destinées à substituer les dossiers papiers traditionnels. C'est dans ce contexte que s'inscrit le projet ERIOS (Espace de Recherche et d'Intégration des Outils numériques en Santé) en s'appuyant sur l'expérience de Pr David Morquin, directeur médical du projet ERIOS et professionnel hospitalier au CHU depuis une vingtaine d'année, qui s'est impliqué très tôt dans l'amélioration du DPI dès son introduction.

Cependant, malgré les promesses d'amélioration des procédures cliniques et administratives, l'informatisation du dossier patient n'a pas toujours été à la hauteur des attentes des professionnels, soulignant des restrictions importantes et des dysfonctionnements fréquents. En effet, la transition d'un système entièrement papier vers des solutions numériques, réalisée principalement entre 2000 et 2010, a été pensée sans une réflexion approfondie sur l'ajustement des outils aux besoins spécifiques du travail médical[1]. En conséquence, les logiciels déployés résultent d'une simple reproduction numérique des dossiers papiers, sans réelle adaptation aux exigences et à la complexité des pratiques médicales.

Cette conception inadaptée force souvent les professionnels de santé à s'ajuster à un outil qui ne correspond pas à leurs pratiques, besoins et attentes. Depuis 2012-2014, différentes évaluations, notamment celle menée au CHU de Montpellier, considèrent le DPI comme un outil laborieux à manipuler au quotidien, cela est dû essentiellement aux logiciels généralement élaborés en anglais ainsi qu'à une navigation complexe diffusant les données indispensables à la prise de décision médicale, ce qui entraîne une surcharge mentale[2]. D'autre part, les contraintes organisationnelles telles que l'éloignement des postes informatiques par rapport aux lieux de consultation, ainsi que l'utilisation des équipements informatiques rares et peu performants dans certains services, compliquent la situation. Cette situation oblige parfois les équipes à opter pour des solutions moins sécurisées ou à adopter des pratiques non conformes, ce qui pourrait menacer la protection des données et la qualité des soins. Il est donc d'autant plus préoccupant que ce phénomène contribue à l'augmentation du stress professionnel ainsi que du risque d'épuisement, déjà élevé dans le secteur hospitalier, marqué par une forte charge de travail et des exigences croissantes[3]. Par conséquent, les professionnels de la santé, tels que médecins, infirmiers et autres intervenants, se retrouvent face à une surcharge cognitive importante, due à la diversité des interfaces, la complexité des parcours de navigation et la charge des procédures de saisie. Cela affecte non seulement leur performance, mais aussi leur bien-être au travail.

De plus, la conversion des DPI en outils de contrôle administratif et de traçabilité, plutôt qu'une assistance qui facilite la pratique clinique, entraîne un sentiment de méfiance et d'opposition chez les utilisateurs. Les professionnels considèrent ces systèmes plus comme des dispositifs de surveillance que comme des outils d'aide à leur activité, ce qui affecte leurs implications.

Ces difficultés ont un impact direct sur la qualité des soins cliniques, augmentant le temps consacré à la documentation par rapport à celui dédié aux soins directs, ce qui peut affecter la relation soignant-patient et engendrer un sentiment d'insatisfaction chez les professionnels. À l'échelle mondiale, 40 % des soignants se déclarent en burn-out à cause de l'usage du DPI[4]. En effet, les médecins consacrent plus de 60 % de leur temps de consultation à saisir des informations[5], ce qui les empêche de regarder leurs patients et les oblige à se concentrer sur leurs écrans. Dans ce

cadre, le CHU de Montpellier s'est impliqué dès 2012 dans la numérisation de ses dossiers patients avec le déploiement du logiciel DxCare par l'éditeur Dedalus, un acteur majeur qui devient le premier fournisseur européen de DPI et le troisième au niveau mondial[6], qui s'est progressivement établi comme une solution centrale pour de nombreuses activités hospitalières. Actuellement, le système rassemble près de 16 000 utilisateurs et enregistre jusqu'à 9 900 connexions simultanées, illustrant l'importance de son intégration dans les usages quotidiens. C'est dans ce contexte que s'inscrit le projet ERIOS (Espace de Recherche et d'Intégration des Outils numériques en Santé) en s'appuyant sur les travaux de recherche de Pr David Morquin, qui s'est lancé dans une thèse en sciences de gestion, dans une approche basée sur les sciences humaines et sociales, qui met en avant la différence entre la création des outils numériques de santé et la réalité du terrain, cherchant à proposer des solutions adaptées pour réduire la charge cognitive, prévenir les erreurs, et recentrer les pratiques sur la relation humaine[7]. D'après lui, le travail hospitalier est de nature complexe et imprévu. Il ne peut se baser sur des outils standardisés, il exige au contraire leurs adaptations aux variations et incertitudes quotidiennes afin de faciliter les pratiques pour que les professionnels puissent consacrer plus d'attention à leurs patients. En conclusion de sa thèse, il insiste sur le fait que le dossier patient informatisé doit être co-construit avec les professionnels de santé, en collaboration avec les ingénieurs, afin de répondre aux besoins cliniques tout en intégrant des solutions technique.

Suite à sa thèse, un projet d'innovation technologique et industrielle a été créé par un consortium[8], c'est-à-dire un groupement d'acteurs publics et privés unis autour d'un objectif commun. Cette innovation a orienté la réflexion élaborée dans le cadre du projet ERIOS, co-porté par le CHU de Montpellier, l'Université de Montpellier et l'éditeur Dedalus étant le Leader du projet bien qu'il ait été pensé par le CHU. Donc un consortium à trois acteurs a été créé afin de soutenir le projet.



FIGURE 1 – Présentation des acteurs du consortium ERIOS récupérés auprès du designer social d'ERIOS

Ce projet ambitieux, financé pour un montant de 3,3 millions d'euros sur une période de 3 ans par la Banque Publique d'Investissement BPI depuis avril 2022, vise à élaborer et à expérimenter de nouvelles solutions numériques hospitalières pensées en collaboration avec les professionnels de santé. Avec un budget considérable, la mission principale d'ERIOS est de réaliser 12 expérimentations autour du DPI et des outils numériques en santé, en s'appuyant sur des méthodologies pertinentes telles que celle des essais cliniques[9].

L'objectif est de démontrer l'influence de ces outils sur la qualité des soins, la sécurité des patients et le bien-être des professionnels. Cette approche de participation et d'évaluation est considérée comme une étape clé afin d'aligner le développement des systèmes d'information hospitaliers aux exigences humaines et organisationnelles du travail clinique. Le but de cette approche est de placer l'utilisateur au cœur de l'innovation ; afin de développer des solutions numériques réellement utiles et adaptées aux besoins, tout en réduisant les erreurs ce qui favorise une meilleure communication entre les professionnels.

En résumé, cette évolution met l'accent sur la complexité et l'importance de l'informatisation dans le domaine de la santé, qui ne peut pas se limiter à une simple mutation technique. En effet, elle nécessite une réflexion profonde sur les modifications organisationnelles et les besoins des utilisateurs. Dans ce cadre, le projet ERIOS, grâce à son approche multidisciplinaire qui est centrée sur l'utilisateur, constitue un exemple pertinent et inspirant pour orienter la progression des systèmes d'information hospitaliers, en adaptant les besoins réels du terrain aux défis actuels.

1.2 Objectifs et missions d'ERIOS

Le projet ERIOS (Espace de Recherche et d'Intégration des Outils numériques en Santé) s'inscrit dans une démarche ambitieuse visant à repenser la conception des outils numériques au service des soins de santé modernes. Face à l'essor massif des données médicales et à la complexité croissante des parcours de soins multidisciplinaires, les professionnels de santé sont confrontés à une surcharge informationnelle, à une rigidité des systèmes numériques et à une perte de sens dans leurs pratiques quotidiennes. Ce projet vise ainsi à créer et affiner une méthodologie globale et rigoureuse pour la conception et l'évaluation des outils de santé numérique, afin qu'ils soient mieux adaptés aux besoins réels du terrain. Pour répondre à ces enjeux, ERIOS développe une méthodologie de co-design (voir page 22) centrée à la fois sur les problèmes concrets rencontrés sur le terrain et sur les solutions attendues par les utilisateurs finaux pour aborder de manière exhaustive les défis de la santé. Ils impliquent activement les utilisateurs à chaque étape, ce qui garantit des résultats adaptés à leurs besoins et améliore leurs flux de travail. Cette méthodologie permet d'observer les processus de travail en situation réelle, de mener des ateliers récoltant les besoins, de tester des maquettes et des prototypes et d'étudier les interactions des utilisateurs à travers des scénarios simulés.

Implanté au cœur du CHU de Montpellier, le projet ERIOS réunit une équipe pluridisciplinaire organisée autour de plusieurs volets complémentaires. Il mobilise un laboratoire en intelligence artificielle générative et en ingénierie des modèles de langage (LLM)(voir page 17), une méthodologie rigoureuse d'évaluation et de design, ainsi qu'un espace collaboratif et un laboratoire utilisateur pour favoriser la co-construction avec les usagers. Le projet s'appuie également sur des activités de prototypage et d'intégration technologique, l'implication d'écoles doctorales et de formations universitaires, ainsi qu'une gestion centralisée du centre et un suivi de projet. Cette organisation permet de croiser les expertises en informatique de santé, science des données, design, sciences sociales, linguistique, sciences de l'implémentation, génie logiciel et traitement automatique du langage naturel.

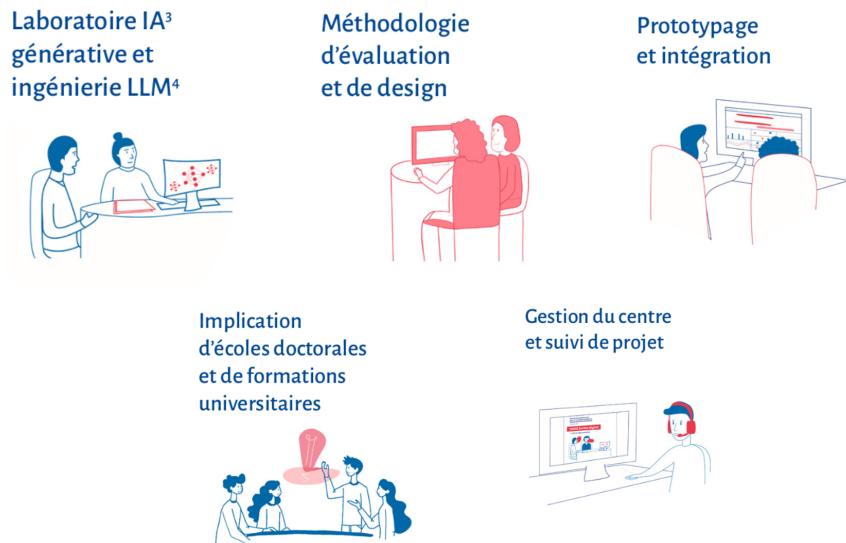


FIGURE 2 – Organisation de l'équipe ERIOS

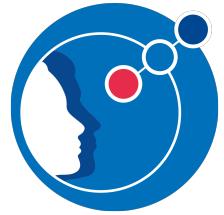
Au cœur du CHU de Montpellier, l'équipe centrale du projet ERIOS se compose de profils aux expertises complémentaires, mobilisés autour d'une ambition commune : intégrer de manière innovante l'intelligence artificielle générative dans les pratiques de santé. La direction médicale et stratégique du projet est assurée par le Professeur M.David, en tant que directeur médical d'ERIOS et responsable de la stratégie IA(voir page 15) et de la gouvernance des données du CHU. Du côté de Dedalus, la phase initiale de prototypage et d'intégration dans le dossier patient informatisé a été menée par V,Quentin et aujourd'hui a évolué vers une spécialisation en UX design assurée par F.Jessica, UX Designer, sous la coordination de F.Loïc, Project Manager. La gestion administrative et le suivi rigoureux du projet, notamment vis à vis des exigences de la BPI, sont assurés par Y.Cécile, responsable de l'équipe « Recherche numérique en santé », et G.Irina, chargée de projet eDOL quant à L'implication étudiante avec l'Université de Montpellier est coordonnée par P.Marin, chargé de coordination pédagogique. Par ailleurs l'interface d'IA générative (voir page 16) est développée et évaluée de manière quantitative par L.Zéno, médecin de santé publique, Y.Kévin, médecin responsable IA générative en santé, et C.Xavier, ingénieur de recherche spécialisé en LLM tandis que l'évaluation qualitative est portée par F.Mylene, designer social, P.Rita et M.Lylia assistante de recherche technologique et R.Louise, docteure en sciences du langage et ingénierie de recherche.



FIGURE 3 – Présentation de l'équipe ERIOS récupéré auprès du designer social

Dans ce cadre, la sélection des cas d'usage fait elle aussi l'objet d'une démarche collaborative : les besoins hospitaliers sont d'abord identifiés avec les professionnels de terrain, puis discutés avec l'éditeur Dedalus, qui valide la pertinence et la faisabilité selon ses critères stratégiques, dans un processus de négociation visant à aligner les objectifs de l'université, du CHU et de l'éditeur.

Ce travail d'alignement interinstitutionnel, symbolisé par le logo du projet, illustre la volonté d'ERIOS de redonner aux soignants un rôle central dans la conception des outils numériques, afin d'améliorer l'efficacité, la sécurité et l'humanité des soins à l'ère du numérique.



Voici un aperçu des expérimentations menées par ERIOS :

1. Projets de tableaux de bord : Ces projets s'inscrivent dans une démarche de co-conception d'outils numériques destinés à accompagner les professionnels de santé dans le suivi et la prise en charge des patients. Chaque tableau de bord a été conçu pour répondre à des besoins spécifiques : suivi légal des mesures d'isolement en psychiatrie, gestion des traitements anti-infectieux, et évaluation de la douleur. Leur point commun est de proposer une visualisation claire et dynamique des données médicales afin de faciliter la compréhension partagée entre les acteurs, améliorer la coordination des soins et soutenir les décisions thérapeutiques au quotidien.



DoloViz

Co-design d'un tableau de bord pour la gestion et le suivi de la douleur des patients et de leurs traitements. L'objectif est d'améliorer la qualité des soins et offrir aux utilisateurs une meilleure compréhension des situations multifactorielles liées à la prise en charge de la douleur, optimisant ainsi les décisions thérapeutiques.



IsoPsy

Co-design d'un tableau de bord pour le suivi légal des patients en isolement ou contention thérapeutique en psychiatrie en application d'un article de loi (dit « l'article 17 »). L'objectif est de permettre à l'ensemble des acteurs impliqués dans ce processus médico-administratif d'améliorer la compréhension partagée de la situation en temps réel d'un ou plusieurs patients ainsi que des tâches à accomplir.



AntibioViz

Co-design d'un tableau de bord pour la prise de décision médicale pour la gestion des traitements anti-infectieux. Dans le processus de prescription thérapeutique, les médecins doivent corrélérer une multitude de données variées tout en tenant compte de leur évolution dans le temps. L'objectif est de faciliter le suivi d'une infection et l'adaptation des traitements anti-infectieux pour un patient donné.

FIGURE 4 – Explications des différents projets à partir des informations recueillies lors de mon stage

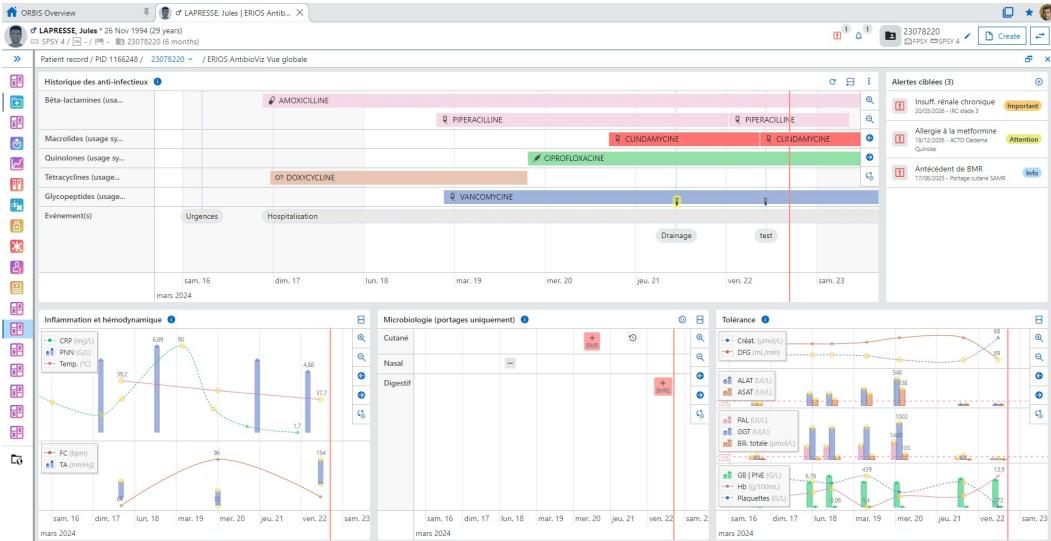


FIGURE 5 – Capture d'écran de l'un des projets de tableaux de bord : DoloViz

2. Projet envisagé : C'est un projet qui s'inscrit dans une réflexion autour de solutions numériques adaptées aux situations de travail en mobilité dans le secteur de la santé, afin de faciliter la consultation et la saisie d'informations en contexte clinique réel.



FIGURE 6 – Explication du projet à partir des informations recueillies lors de mon stage

3. Projets à base d'IA générative et LLM (traitement du langage naturel) : Ces projets visent à explorer le potentiel des modèles de langage dans le domaine de la santé, en s'intéressant à leur usage pour l'analyse des données textuelles, la documentation médicale et l'assistance à la décision. Ils s'inscrivent dans une démarche d'évaluation des apports de l'IA générative en contexte réel, avec une attention portée à la fiabilité des outils, à leur impact sur les pratiques cliniques et à l'amélioration de la qualité des soins.

The image displays four separate project logos arranged in a 2x2 grid. Top-left: Recoconsult, featuring a blue circular icon with a brain-like pattern and the text "Recoconsult". Top-right: LLM4 Quality, featuring a blue circular icon with a network or atom-like pattern and the text "LLM4 Quality". Bottom-left: MedNote, featuring a blue circular icon with a computer monitor and the text "MedNote". Bottom-right: ERIOS Assistant, featuring a blue circular icon with a clipboard and the text "ERIOS Assistant". Each logo is accompanied by a descriptive text box below it.

FIGURE 7 – Explication des projets à partir des informations recueillies lors de mon stage

1.3 Présentation d'ERIOS assistant

ERIOS assistant marque une innovation significative dans l'intégration de l'intelligence artificielle générative dans le secteur de la santé. Il se présente dans l'évolution de transformation des pratiques professionnelles en apportant une aide technologique destinée aux intervenants du secteur médical, en particulier sur la simplification de la rédaction des documents, souvent perçue comme une tâche longue et à faible valeur ajoutée. Pensé et développé en étroite collaboration avec les professionnels de terrain, ERIOS vise à répondre aux besoins identifiés afin de faciliter la production de textes médicaux tout en assurant la qualité, la traçabilité et la confidentialité des données. La documentation médicale joue un rôle important dans le processus de soin, à la fois pour des raisons légales et pour assurer une bonne communication entre les professionnels de santé. Cependant, cette activité est toujours considérée comme un obstacle par les praticiens, qui limite leur temps censé être consacré aux tâches cliniques et relationnelles.

C'est dans ce cadre qu'intervient l'intelligence artificielle générative comme un outil efficace pour rétablir l'équilibre entre les tâches administratives et le cœur de métier, capable de générer des textes cohérents, précis et appropriés au contexte.

The screenshot shows the homepage of the ERIOS Assistant website. On the left, there's a vertical navigation bar with icons and text for 'ERIOS Assistant', 'Explorer les assistants', 'Tableau de suivi', 'Migration des données', and 'Gestion des LLMs'. Below this is a yellow box containing 'Dispositif expérimental du CHU de Montpellier'. Underneath are sections for 'Aujourd'hui' (Lettre de liaison médicale URG PED, Identification d'action de prévention), 'CHU FERREIRA PIRES ANA Administratrice', and a 'Le projet ERIOS' section featuring a photo of a team meeting with a monitor displaying the text 'Equipe ERIOS' and 'Médecins des urgences pédiatriques'.

FIGURE 8 – Capture d'écran de la page d'accueil d'ERIOS assistant

ERIOS assistant étudie plusieurs cas d'usage apportant des réponses à des problématiques précises observées sur le terrain médical.

Génération de textes médicaux professionnels : Le premier cas d'usage s'applique à la création de documents médicaux ayant une haute valeur légale ou organisationnelle tels que les lettres de liaison, les comptes rendus d'hospitalisation, les comptes rendus opératoires ou encore les bilans de suivi. Ces documents sont nécessaires pour assurer l'échange interprofessionnel tout au long du parcours de soins, la continuité et la qualité des prises en charge. Dans de nombreux services hospitaliers, la rédaction de ces documents se base principalement sur des méthodes manuelles, telles que la rédaction d'un compte rendu, ou semi-automatisées, comme l'utilisation de modèles préremplis, demandant un investissement important de temps de la part du personnel médical. Cependant, ERIOS assistant offre la possibilité de générer automatiquement des textes à partir de données médicales structurées ou semi-structurées à partir du Dossier Patient Informatisé (DPI). Les premières expérimentations montrent des améliorations encourageantes en termes de qualité syntaxique et de pertinence médicale des documents générés, mais aussi des limites en terme de gain de temps en fonction des documents générés. Par exemple, dans des services comme la radiologie, où la structure des comptes rendus est fortement standardisée, la génération automatique permet un quasi-automatisme. Toutefois, dans des disciplines comme la psychiatrie, la médecine interne ou la médecine de la douleur, où chaque situation clinique nécessite une étude détaillée et contextualisée, le recours aux professionnels de santé reste toujours indispensable. Dans ces situations, l'IA offre principalement une base textuelle prédéfinie sur laquelle le professionnel peut s'appuyer pour finaliser le contenu.

Génération de textes médicaux simplifiés pour les patients : Un second cas d'usage, particulièrement innovant et socialement important, porte sur la génération de documents médicaux destinés aux patients ou à leurs aidants. Cette fonctionnalité vise à améliorer la compréhension des informations médicales afin de faciliter aux patients la suite de leur suivi médical. Le but d'ERIOS assistant est de reformuler un contenu médical complexe en un langage simple adapté au niveau de compréhension du public ciblé. Les documents générés permettent d'expliquer un diagnostic, décrire les étapes d'un traitement et clarifier les conseils proposés par le professionnel. Cette méthode vise à favoriser l'autonomie du patient en essayant de réduire les

ambiguïtés susceptibles d'entraîner des retours répétés vers le professionnel de santé pour obtenir des explications supplémentaires, ce qui contribue à la diminution de leur charge de travail. Le projet a notamment été expérimenté dans des services d'urgences pédiatriques, où la présence de contrainte de temps et la clarté de l'information a joué un rôle important. Dans ces situations, les parents font face à une épreuve assez difficile, ils sont obligés de prendre des décisions rapidement, sans avoir toujours accès à toutes les informations nécessaires. C'est pour ces raisons que les textes générés sont des lettres d'information adressées aux parents, ainsi qu'aux enfants et aux adolescents, rédigés dans un langage approprié. De plus, l'assistant intelligent développé au sein du projet ERIOS vise à fournir un accès immédiat, sécurisé et fidèle aux informations médicales fournies. Ce cas d'usage devrait être appliqué dans plusieurs services médicaux autres que la pédiatrie où la communication avec le patient est particulièrement essentiel. La simplification des contenus médicaux devient un élément important pour accompagner les patients dans la compréhension de leurs démarches de soin. Par ailleurs, la capacité de l'outil à produire des documents rédigés dans un langage clair et adapté au niveau de compréhension du patient peut être considéré comme un avantage, particulièrement dans des contextes marqués par des diversités linguistiques. Cela pourrait permettre de réduire les inégalités d'accès à l'information liée aux contraintes culturelles et linguistiques. C'est autour de ce cas d'usage que portera mon travail tout au long de mon stage.

Explication et reformulation d'ordonnances : Il s'agit d'un troisième cas d'usage qui s'intéresse à l'amélioration de la compréhension des ordonnances médicales. ERIOS assistant cherche à produire des explications simples et ajustées à partir d'ordonnances complexes, en traduisant le vocabulaire médical en un langage simple. En effet, l'outil vise à offrir aux patients les informations nécessaires pour comprendre, par exemple, les modalités de prise, la durée du traitement, les effets secondaires, ainsi que les mesures de précaution à respecter. Cela afin de faciliter l'admission du traitement et de réduire les éventuelles erreurs des patients. Dans ces situations, une mauvaise compréhension des consignes peut entraîner des erreurs de prise ou des interactions médicamenteuses potentiellement graves. ERIOS assistant permettrait de garantir une démarche thérapeutique bien sécurisée et de renforcer l'indépendance du patient. Enfin, cette fonctionnalité de simplification bénéficierait également aux professionnels impliqués dans la chaîne de soins tels que pharmaciens, infirmiers, aides-soignants ou intervenants à domicile en améliorant la lisibilité des prescriptions et leurs utilisations. L'assistant ne se limiterait donc pas à la relation médecin-patient, mais s'engagerait dans une approche de coordination interprofessionnelle dans le but d'assurer une meilleure communication.

Génération automatique de certificats MDPH : Un quatrième cas d'usage emblématique concerne le pré-remplissage des certificats CERFA (Centre d'Enregistrement et de Révision des Formulaires Administratifs) destinés à la Maison Départementale des Personnes Handicapées (MDPH). Selon l'ingénierie de recherche d'ERIOS ces documents indispensables pour que les personnes en situation de handicap puissent accéder à des aides financières, matérielles ou éducatives, sont souvent perçus comme fastidieux par les médecins : ils sont longs, redondants, et peu valorisés dans la pratique quotidienne. Face à cette contrainte administrative, l'assistant ERIOS se veut proposer une solution efficace en générant automatiquement le contenu de ces formulaires à partir des informations déjà disponibles dans le dossier médical du patient. Cette automatisation permettrait au professionnel de santé de simplement relire et ajuster certaines données pour permettre la réduction du temps consacré à cette tâche et limiter la charge mentale associée.

Chapitre 2 : Problématique du stage

2.1 Contexte spécifique du stage :

Le domaine de la santé fait face à une complexité évolutive des informations médicales destinées aux patients et aux professionnels[10]. Les documents médicaux, quel que soit les comptes rendus d'examen, les protocoles thérapeutiques ou de communications entre soignants et patients, sont généralement rédigés dans un langage technique difficile à comprendre pour la population. Cette contrainte linguistique favorise les inégalités dans l'accès aux soins en terme de compréhension des recommandations médicales. En effet, l'évolution en intelligence artificielle, et plus particulièrement dans le traitement automatique du langage naturel, c'est-à-dire le langage utilisé par l'humain, offre des opportunités importantes pour simplifier ces contenus complexes, « Les modèles de NLP basés sur le deep learning, notamment les architectures transformer comme BERT et GPT-2, offrent un potentiel significatif pour la traduction, la simplification et le résumé de contenus médicaux, favorisant l'eHealth literacy [11] ». Le développement d'un chatbot conversationnel basé sur des modèles de langage (Large Language Models, LLMs), pourrait rendre les informations médicales plus accessibles en proposant des reformulations simplifiées, adaptées aux besoins et au niveau de compréhension des utilisateurs. C'est dans ce contexte que se déroule mon stage au sein d'ERIOS, un espace spécialisé dans la recherche appliquée en santé numérique. Il s'inscrit dans un projet global qui cherche à allier innovation technologique, réduction de la charge administrative des professionnels de santé et facilitation de la compréhension des diagnostics par les patients. Cela en intégrant une démarche d'évaluation qui s'intéresse à la fois à l'interface et à la fois à la qualité des textes générés. Pour répondre à ces objectifs, une interface chatbot en phase de test a été développée. Elle devrait permettre la générations des lettres simplifiés. Les modèles linguistiques utilisés sont entraînés et optimisés pour tenter de générer des lettres à destination des patients et/ou des professionnels de santé à partir des observations médicales, tout en tentant de préserver leur fidélité et leur précision.

2.2 Enjeux et objectifs des missions confiées

Les enjeux du projet sont multiples et interdépendants, à la croisée de la santé publique, des sciences du langage, et des technologies numériques .

1. Enjeux sociaux : La simplification des textes médicaux a pour objectif de réduire les inégalités d'accès à l'information tout en assurant la justesse médicale pour éviter toute confusion susceptible de contrarier la sécurité des soins. L'enjeu est donc important : équilibrer l'accessibilité et la fiabilité de l'information.

2. Enjeux technologiques : La conception de l'interface chatbot repose sur des principes d'ergonomie et d'expérience utilisateur (UX) optimisés pour assurer une interaction naturelle, simple et intuitive. Par ailleurs, une expertise approfondie est nécessaire afin de garantir la sélection des modèles de langage ainsi que leurs configurations, ce qui permet de contrôler la qualité des simplifications automatiques en limitant les biais et les erreurs linguistiques. De plus, l'intégration d'un système capable d'être adapté et prendre en compte les retours des utilisateurs représente un enjeu technique majeur, indispensable pour assurer une évolution continue de l'outil.

3. Enjeux méthodologiques : L'évaluation du projet doit s'appuyer sur des protocoles rigoureux. Il s'agit d'une part d'évaluer l'interface chatbot par des experts UX à l'aide d'une méthode heuristique, puis de collecter des retours qualitatifs issus des tests utilisateurs. D'autre part, la qualité des simplifications doit être interprétée à travers des indicateurs objectifs (indices de lisibilité, fidélité sémantique) et subjectifs (compréhension, satisfaction des utilisateurs). Dans cette approche, il est indispensable de réaliser des tests statistiques dans le but de valider les différences entre les vraies versions et celles simplifiées, ainsi que pour comparer les différents modèles de reformulation.

2.3 Problématique étudiée

C'est dans ce cadre que s'inscrit la problématique principale de mon stage en se basant sur la question suivante :

Dans quelle mesure un chatbot reposant sur des modèles d'intelligence artificielle, peut-il contribuer à la simplification efficace des textes médicaux tout en garantissant une expérience utilisateur satisfaisante et rigoureuse, évaluée à la fois par des méthodes qualitatives et quantitatives prenant en compte la qualité linguistique des contenus générés et l'accessibilité de l'interface ?

Cette problématique relève plusieurs axes :

1. Évaluation de l'interface utilisateur : Il s'agit de déterminer dans quelle mesure l'interface du chatbot est considérée comme accessible et intuitive par divers profils d'usagers. L'évaluation heuristique par des experts vise à déterminer les points forts et les points faibles de l'outil, tandis que les tests utilisateurs apportent une compréhension détaillée des attentes, des pratiques réelles et des difficultés rencontrées. Ces analyses permettent de guider les axes d'amélioration pour optimiser la prise en main de l'outil.

2. Analyse comparative de la simplification automatique : Plusieurs modèles linguistiques (GPT-4o, Mistral Medium3, Gemma3) sont testés et ajustés. L'objectif est de comparer leurs performances en termes de lisibilité, de fidélité sémantique et de satisfaction des utilisateurs via des évaluations d'experts médicaux. Cette approche, qui prend en compte plusieurs critères, est perçue comme indispensable pour identifier le ou les modèles les plus adaptés à la reformulation de textes complexes en langage naturel.

3. Approche méthodologique mixte et statistique : Afin d'aboutir à des conclusions solides, l'évaluation repose sur une approche basée sur des méthodes qualitatives (entretiens, observations) et quantitatives (mesures objectives : score de lisibilité, temps de lecture, écart type, variance. . .). Cette complémentarité est soulignée par l'utilisation de tests statistiques (ANOVA, Wilcoxon, analyses multivariées) afin de vérifier la signification des différences observées entre les différentes versions des textes et les profils d'utilisateurs évalués.

Pour conclure, ce travail vise à tenter d'analyser et d'évaluer l'intégration d'un outil d'assistance médicale, en alliant les dimensions techniques, linguistiques et humaine. Ceci est effectué dans le but d'accompagner le développement des outils numériques répondant aux besoins des patients et des professionnels de santé.

Chapitre 3 : L'IA générative et les modèles de langage dans le secteur de la santé

3.1 Définition de L'IA

L'intelligence artificielle (IA) désigne la capacité d'une machine à reproduire des comportements et des processus cognitifs typiquement humains, tels que le raisonnement, la planification, la prise de décision ou la créativité. Elle permet à des systèmes techniques de percevoir leur environnement , d'interpréter ces données et d'agir de manière adaptée afin d'atteindre des objectifs précis[12]. Ces systèmes d'IA sont conçus pour analyser des informations complexes, résoudre des problèmes variés et ajuster leurs actions en fonction des résultats obtenus, ce qui leur confère une autonomie partielle ou totale selon les cas. En évaluant les effets de leurs propres actions, ils peuvent apprendre et adapter leur comportement, améliorant ainsi leur efficacité avec le temps. L'intelligence artificielle repose sur des algorithmes, des modèles mathématiques et des règles programmées qui simulent les processus mentaux humains. Elle intègre plusieurs fonctions essentielles : la perception des données, le raisonnement logique, l'apprentissage à partir d'expériences passées et l'action dans le monde réel. Grâce à cette combinaison, l'IA automatise des tâches complexes qui nécessitent habituellement une forme d'intelligence humaine, contribuant ainsi à améliorer la performance et la prise de décision dans de nombreux domaines.

3.2 Applications dans le domaine de la santé :

L'application de l'intelligence artificielle dans le domaine médical montre un énorme potentiel. En effet, l'IA permet non seulement d'améliorer la qualité des soins, mais aussi mieux gérer les ressources médicales et d'anticiper les crises sanitaires. Voici les principales applications actuelles de l'IA dans le secteur médical :



FIGURE 9 – Les domaines d'application de l'IA en santé[13]

Pour cette partie nous nous sommes basés sur les principales applications actuelles de l'IA de Midhat Tilawat[14]

Médecine prédictive ; anticiper les maladies avant leur apparition : La médecine prédictive s'appuie sur l'analyse de données médicales massives (big data) pour déterminer les risques de développer certaines pathologies. À l'aide des algorithmes de machine learning, il est possible d'identifier des facteurs de risque qu'on ne peut pas repérer avec l'observation humaine, et de prédire l'apparition de maladies telles que le diabète, les maladies cardiovasculaires ou certains cancers. Par exemple, des modèles d'IA sont en mesure d'analyser les antécédents familiaux, les comportements de vie (alimentation, activité physique), les marqueurs biologiques ou encore les données génétiques pour estimer la probabilité de développer une maladie chronique dans les années à venir. Cela permet de mettre en place des stratégies de prévention ciblées avant l'apparition des premiers symptômes.

Médecine de précision ; des traitements personnalisés pour chaque patient : La médecine de précision, également appelée médecine personnalisée, vise à ajuster les traitements en fonction des caractéristiques individuelles du patient. En effet, l'intelligence artificielle joue un rôle clé en analysant des données massives cliniques et biologiques afin de recommander les traitements les plus efficaces et minimiser les effets indésirables. Par exemple, dans le traitement du cancer, les algorithmes peuvent proposer la thérapie la mieux adaptée à un type spécifique de tumeur en se basant sur le profil génétique du patient.

Aide à la décision ; diagnostic et choix thérapeutique assistés : Les systèmes d'IA peuvent accompagner les professionnels de santé dans la prise de décision, en fournissant des analyses instantanées des données cliniques d'un patient et en proposant des diagnostics ou des stratégies thérapeutiques. Par exemple, un médecin peut être assisté par un système qui, à partir des symptômes, du dossier médical et des résultats d'examens, propose des hypothèses diagnostiques, calcule un score de probabilité pour chaque maladie et suggère les examens complémentaires à réaliser. Des systèmes comme Watson Health d'IBM ont été développés dans cette optique[15]. L'IA ne remplace pas le médecin, mais renforce sa capacité à prendre des décisions en particulier dans les situations compliquées

Robots compagnons ; soutien aux personnes âgées et dépendantes : Les robots d'assistance, dotés d'IA, sont conçus pour accompagner les personnes âgées, isolées ou en perte d'autonomie. Ils assurent des fonctions à la fois pratiques, sociales et émotionnelles. Ces robots ont la capacité de rappeler la prise de médicaments, surveiller les constantes vitales, avertir les services d'urgence en cas de chute et proposer des activités cognitives pour stimuler la mémoire. Par exemple on peut citer Paro, un robot phoque interactif utilisé en gériatrie[16]. Ces technologies permettent de renforcer le lien social, prévenir l'isolement, et alléger la charge des aidants.

Chirurgie assistée par ordinateur ; précision et sécurité : L'intelligence artificielle est intégrée dans les systèmes de robotique chirurgicale, ce qui permet au chirurgien de contrôler à distance des instruments plus petits avec précision, afin de réduire les risques de complication, les douleurs postopératoires et accélérer la récupération du patient. La chirurgie assistée par IA s'impose comme outil essentiel de l'innovation technologique au bloc opératoire.

Prévention ; pharmacovigilance : L'IA contribue fortement à la pharmacovigilance, permet de surveiller en continu les effets secondaires des médicaments en analysant les bases de données de prescriptions, les retours patients ou les publications médicales. Cela améliore la sécurité des traitements et permet un retrait rapide de médicaments problématiques.

3.3 Définition de L'IA générative et LLM :

3.3.1 Définition de l'IA générative :

L'IA générative est une sous-branche de l'intelligence artificielle qui se concentre sur la création de nouveaux contenus qu'il s'agisse de textes, d'images, de sons, de vidéos ou même de données synthétiques à partir des connaissances acquises lors de l'apprentissage[17]. Contraire-

ment à l'IA traditionnelle, principalement axée sur l'analyse, la classification ou la prédiction à partir de données existantes, l'IA générative se distingue par sa capacité à produire des éléments inédits, cohérents et réalistes. Elle repose sur des modèles statistiques avancés, notamment les réseaux de neurones profonds, et s'appuie sur des techniques d'apprentissage automatique telles que les réseaux antagonistes génératifs (GAN), les autoencodeurs variationnels (VAE) ou encore les transformateurs, parmi lesquels les modèles de langage de grande taille (LLM) comme GPT occupent une place centrale. Ces architectures permettent à l'IA de modéliser des distributions complexes et de générer des contenus qui imitent de manière crédible les données d'origine. L'IA générative représente aujourd'hui une avancée majeure dans le champ de l'intelligence artificielle, avec des applications en plein essor dans des domaines variés tels que l'art, la communication, la recherche scientifique, l'éducation, l'industrie, et bien entendu, la santé.

3.3.2 Définition d'un LLM (Large Language Models)

L'un des exemples les plus aboutis d'intelligence artificielle générative dans le domaine du langage est constitué par les grands modèles de langage, ou Large Language Models (LLM). Ces modèles, tels que GPT (Generative Pre-trained Transformer) développé par OpenAI, BERT de Google, LLaMA de Meta, Mistral développé par la start-up française Mistral AI, ou encore Gemma de Google, sont entraînés sur d'immenses corpus textuels issus du Web, de livres, de publications scientifiques ou de bases de données spécialisées[18]. Leur objectif est de modéliser le langage humain de manière à pouvoir comprendre, générer et manipuler des textes avec un haut degré de cohérence. Les LLM reposent sur des architectures de réseaux neuronaux profonds, notamment les transformateurs, qui leur permettent de saisir le contexte global d'un texte, de traiter les dépendances à longue distance entre les mots, et d'anticiper la suite logique d'une phrase. Les textes sont analysés sous forme de tokens, c'est-à-dire des unités élémentaires de langage (mots, morceaux de mots ou caractères), que le modèle encode numériquement pour les traiter. À partir de ces représentations, les LLM peuvent produire des réponses structurées, en s'appuyant sur les probabilités d'occurrence des éléments suivants dans un contexte donné. Grâce à ces capacités, ils peuvent accomplir une grande variété de tâches : rédaction de contenus, résumés automatiques, traduction multilingue, réponses à des questions ouvertes, reformulations, et même génération de code informatique. Un atout essentiel de ces modèles réside dans leur apprentissage non supervisé : ils sont capables d'extraire des régularités linguistiques à partir de vastes volumes de données non annotées, sans intervention humaine directe. Cette approche repose notamment sur des objectifs d'entraînement tels que la prédiction du token suivant (comme dans GPT) ou le masquage de mots à deviner (comme dans BERT). Cette caractéristique leur confère une grande polyvalence et une adaptabilité remarquable dans de nombreux domaines, y compris celui de la santé, où ils commencent à être utilisés pour assister les professionnels dans la rédaction de rapports médicaux, l'analyse de la documentation clinique ou encore le soutien au diagnostic.

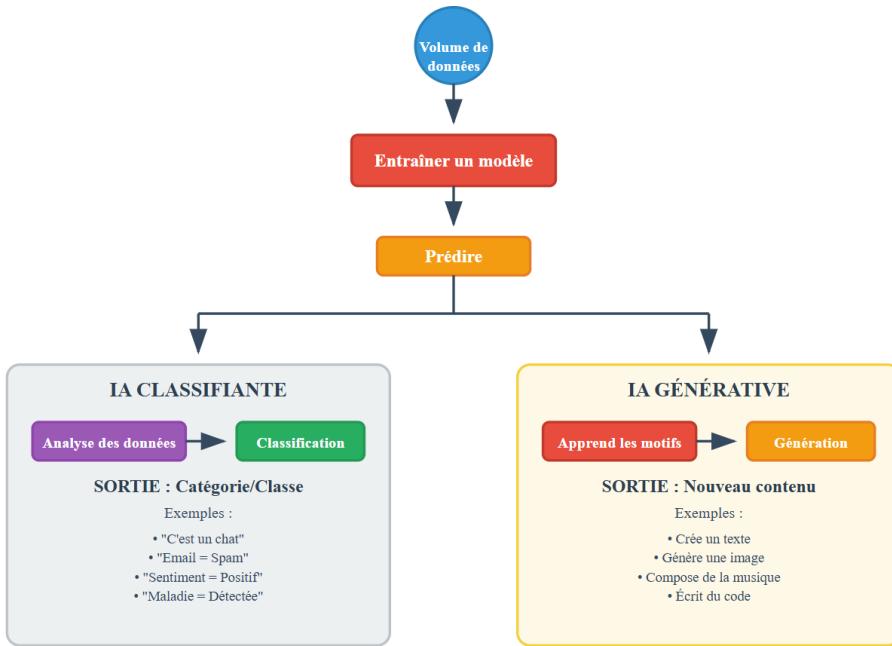


FIGURE 10 – Explication du fonctionnement de L'IA à partir des informations recueillies lors d'un atelier durant mon stage

3.4 Applications de l'IA générative et LLM dans le domaine de santé :

Les progrès récentes en intelligence artificielle, notamment grâce aux IA génératives et aux grands modèles de langage (LLM – Large Language Models), ont transformés le secteur de la santé d'une manière impressionnante . Ces technologies, entraînées sur des quantités importantes de données médicales, scientifiques et cliniques, sont capables de comprendre, générer et analyser le langage naturel, ce qui leur permet d'automatiser de nombreuses tâches médicales, d'accompagnés les professionnels de santé et d'optimiser la prise en charge des patients. De plus, les IA génératives sot en mesure de produire des contenus personnalisés tel que les rapports médicaux et les comptes rendus a destination des patients. Par ailleurs, les LLM spécialisés tels que Med-PaLM [19] ou BioMedLM, sont capable d'interpréter avec précision les termes médicaux et les dossiers patients afin de fournir des réponses fiables. La combinaison de ces outils permet de créer une médecine plus efficace, humaine et innovant tout en accélérant la transformation numérique du parcours de soins et en améliorant l'expérience du patient.

Accélération de la découverte de médicaments : L'intelligence artificielle générative a transformé le processus de découverte de médicaments en proposant des structures moléculaires innovantes, spécifiquement élaborées pour cibler certaines pathologies. En générant virtuellement de nouvelles molécules et en simulant leurs interactions biologiques, l'IA permet d'identifier rapidement des candidats prometteurs. Elle anticipe également les effets secondaires en prédisant les interactions médicamenteuses, ce qui améliore la sécurité des traitements. De plus, l'IA générative joue un rôle croissant dans la recherche biomédicale. En analysant de vastes ensembles de données issus de publications scientifiques ou de dossiers médicaux, elle formule de nouvelles hypothèses, révèle des corrélations entre les variables cliniques et propose des pistes de recherches innovantes. Cette capacité d'analyse accélère le rythme de l'innovation médicale en réduisant les coûts et les délais nécessaires l'homologation des nouveaux médicament.

Transformation de l'imagerie médicale : L'IA générative transforme en profondeur l'imagerie médicale. Elle améliore la résolution des images issues d'examens tels que l'IRM ou le scanner, ce qui facilite le diagnostic de pathologies complexes. Elle est aussi capable de générer des images synthétiques pour entraîner des modèles d'apprentissage tout en respectant la confidentialité des patients. En reconstruction d'images, elle permet une visualisation plus claire et

détaillée des tissus, organes ou anomalies. Ces progrès renforcent la précision diagnostique tout en contribuant à la formation de modèles performant dans un cadre sécurisé.

Reduction de la charge administrative : Un des domaines où l'IA générative démontre déjà son potentiel est la rédaction automatique des documents médicaux. Les professionnels de santé consacrent une part importante de leur temps à la production de comptes rendus, certificats, bilans ou lettres de liaison. L'intégration de modèles de langage de grande taille (LLM), tels que ceux issus des architectures de type GPT (Generative Pre-trained Transformer), permet de générer automatiquement des rapports médicaux structurés à partir de diverses sources : dictées vocales, formulaires, mots-clés ou notes non structurées saisies par les praticiens. Cette automatisation représente un gain de temps considérable pour les professionnels de santé, réduit les erreurs de transcription, standardise la qualité des documents produits et facilite la mise à jour en temps réel des DPI . Au-delà de l'assistance documentaire, l'IA générative s'impose comme un atout stratégique pour la formation continue et la surveillance scientifique au sein des établissements de santé. Elle est capable de produire automatiquement des résumés d'articles scientifiques, de répondre à des questions médicales complexes posées en langage naturel, ou encore de simuler des cas cliniques interactifs. Cette accessibilité instantanée à l'information médical renforce les compétences des professionnels de manière fluide et personnalisée, tout en assurant l'amélioration continue de la qualité des soins.

Amélioration de l'accueil et de l'orientation des patients : L'IA générative, et notamment les modèles de langage de grande taille (LLM), est également au cœur des agents conversationnels, qu'ils soient textuels (chatbots) ou vocaux , capables de dialoguer avec les patients de manière naturelle, fluide et contextuellement pertinente[20]. Ces modelés sont entraînés sur des documents médicaux spécialisés ce qui leur permettent de répondre à des questions fréquentes telles que les horaires de consultation, les démarches administratives, les spécialités disponibles ou encore les modalités de prise en charge. Grâce à leur capacité de compréhension du langage, les LLM sont capables d'adapter les réponses au profil du patient et au contexte de la conversation et les orienter efficacement vers les services ou les professionnels de santé les plus adaptés Par exemple on peut citer Ello, qui utilise la reconnaissance vocale pour traiter jusqu'à 90 % des appels entrants sans intervention humaine, illustrent le potentiel de cette technologie.

Automatiser la gestion des rendez-vous et gestion des flux hospitaliers : L'IA générative, lorsqu'elle est associée à des moteurs de traitement du langage naturel, permet d'automatiser la prise de rendez-vous médicaux tout en tenant compte de contraintes complexes telles que le type de consultation, la disponibilité des médecins ou encore les antécédents médicaux du patient. Des plateformes comme Polaris ou PetalMD[21] intègrent déjà des modules fondés sur des modèles génératifs pour analyser les demandes formulées en langage libre et ajuster les créneaux proposés. Ces outils améliorent la fluidité du parcours de soin en rendant la planification plus intuitive, rapide et personnalisée. En parallèle, l'IA générative et les LLM peuvent être utilisés pour améliorer la gestion des flux hospitaliers. Ils sont en capacité de prédire les pics d'admission aux urgences, de suggérer une répartition optimale des lits ou d'anticiper les besoins en personnel et matériel. Grace a leur capacité à détecter des anomalies dans les processus internes tel que les retards de facturation et des incohérences dans les plannings contribue a l'amélioration de la gestion des établissement de santé et le renforcement de la communication entre les différent services .À plus long terme, ces technologies pourraient évoluer vers des solutions de triage médical automatisé, dans lesquelles un assistant intelligent interagit avec le patient, interprète ses symptômes , pose des questions ciblées et l'oriente vers le niveau de soins le plus adapté (consultation classique, urgence, téléconsultation).

3.5 Présentation des LLM utilisés dans le projet ERIOS :

Dans le cadre du projet ERIOS les modèles de langage de grande taille (LLMs) sont mobilisés pour analyser, structurer et valoriser les données textuelles issues du système de santé. Ces modèles facilitent le traitement automatique de l'information non structurée, telle que les commentaires de patients, les comptes rendus médicaux ou les documents administratifs produits en milieu hospitalier. L'objectif est de réduire la charge cognitive et administrative des soignants, d'améliorer la qualité des soins, et de renforcer le pilotage des établissements à partir de données qualitatives. Bien que les modèles de langage de grande taille (LLMs) offrent des performances impressionnantes dans de nombreuses tâches, ils présentent également des limites majeures qu'il convient de prendre en compte dans les projets en santé, comme ceux du programme ERIOS. L'un des problèmes les plus critiques est celui des hallucinations : il s'agit de réponses générées par le modèle qui sont fausses ou inventées, mais exprimées avec une grande assurance. Dans un contexte médical, cela peut conduire à des erreurs d'interprétation, comme l'attribution incorrecte d'un commentaire patient à une mauvaise catégorie, ou la rédaction de contenu erroné dans un document clinique. Les LLMs sont également sensibles à la formulation des consignes. De légères variations dans l'instruction donnée peuvent conduire à des réponses différentes, parfois contradictoires. Ces caractéristiques rendent indispensable la mise en place de stratégies robustes de contrôle, mais aussi une sélection rigoureuse du modèle utilisé, adaptée au contexte d'usage. C'est pourquoi, dans le cadre d'ERIOS, le choix des LLMs ne repose pas uniquement sur des critères de notoriété ou de performance brute. Une démarche méthodique d'évaluation et de benchmarking est menée pour sélectionner les modèles les plus adaptés aux tâches spécifiques.

ERIOSClassifier : Ce projet vise à exploiter les commentaires libres des patients recueillis via les questionnaires de satisfaction (ISATIS) afin de mieux comprendre leur expérience et d'identifier des pistes d'amélioration pour les établissements de santé. Plus de 27000 commentaires ont été automatiquement classés selon les catégories recommandées par la Haute Autorité de Santé (HAS), telles que la communication, l'accueil, les soins, la gestion de la douleur[22]. Face à la complexité et à la diversité des textes, plusieurs modèles d'apprentissage automatique et grands modèles de langage (LLMs) ont été évalués afin de sélectionner la méthode la plus performante et fiable pour cette tâche de classification automatisée, qui vise à limiter la mobilisation de ressources humaines lourdes pour la relecture manuelle.

ERIOS Assistant : est le projet sur lequel j'ai travaillé durant mon stage, et que je vais présenter dans la section suivante.

3.6 Choix du modèle de LLM pour ERIOS Assistant :

Ce projet a pour objectif de développer un assistant virtuel reposant sur l'utilisation de grands modèles de langage (LLM) pour tester et évaluer différents usages de traitement des données médicales textuelles automatisés par l'IA générative spécifiquement dans le contexte hospitalier. Il s'agit de concevoir des cas d'usage pertinents, de travailler sur l'ingénierie des prompts, et de mener des évaluations en conditions réelles. Le projet vise également à analyser la fiabilité des textes générés et à élaborer des protocoles pour en améliorer la qualité. Pour ce faire, plusieurs modèles de langage sont utilisés notamment :

GPT-4o : modèle multimodal de dernière génération développé par OpenAI. Ce modèle propriétaire, accessible via API cloud, est reconnu pour ses capacités avancées de compréhension et de génération de texte, offrant une grande polyvalence pour diverses applications.

GPT-4 (Microsoft serveur) : version optimisée de GPT-4 proposée par Microsoft, hébergée sur son infrastructure cloud Azure. Ce modèle propriétaire garantit des performances élevées et une intégration sécurisée dans les environnements d'entreprise.

GEMMA3 : modèle développé pour des usages spécifiques, généralement considéré comme propriétaire ou spécialisé. Des informations complémentaires sont nécessaires pour en préciser les caractéristiques exactes.

GPT-4 (VPN) : utilisation sécurisée du modèle propriétaire GPT-4 via un réseau privé virtuel (VPN), permettant d'assurer la confidentialité renforcée des données sensibles traitées.

Mistral Small 3.1 : modèle open source de 24 milliards de paramètres, conçu pour un déploiement local sans dépendance aux services cloud. Il favorise ainsi la souveraineté des données tout en offrant des performances compétitives.

Mistral Medium : modèle open source de taille intermédiaire, conçu pour offrir un excellent compromis entre performance et efficacité. Il peut être déployé localement, sans dépendance aux services cloud, garantissant ainsi la souveraineté des données. Grâce à ses performances robustes, il est adapté à une large gamme d'applications, du prototypage rapide aux déploiements en production.

Deepseek R IIIème génération 70B (LLMÉ 70B) : modèle open source de grande taille, avec 70 milliards de paramètres, reposant sur une architecture Mixture-of-Experts (MoE). Il offre un bon équilibre entre puissance de calcul et efficacité, particulièrement adapté aux traitements volumineux.

Deepseek R1 : variante open source de Deepseek, exploitant l'architecture Mixture-of-Experts, optimisée pour un usage local avec un excellent rapport qualité/prix.

Deepseek R1 QWQ : version modifiée et open source du Deepseek R1, spécialement testée et ajustée pour répondre aux exigences spécifiques du projet.

Mistral Large 2411 : modèle premium développé par Mistral AI. Ce modèle propriétaire offre des performances avancées adaptées aux applications complexes de traitement du langage naturel.

Dans le cadre de mon stage, je travaille sur le projet ERIOS Assistant et on va s'intéresser sur le cas d'usage de la génération de textes médicaux simplifiés à destination des patients afin d'améliorer l'accessibilité de l'information médicale. Ce cas d'usage est expérimenté au sein du service des urgences pédiatriques, un environnement marqué par une forte pression sur les soignants, une surcharge de consultations, et un manque de temps pour fournir des explications claires aux familles. L'objectif est d'aider les professionnels de santé à transmettre rapidement des messages compréhensibles, tout en réduisant les consultations inutiles liées à des problèmes non urgentes .

Pour cela au début j'ai testé avec trois modèles de langage (LLM) : GPT-4o, Gemma 3 et Mistral Large 2411. Cependant, après évaluation, ce dernier a montré certaines limites en termes de précision et de fluidité des textes générés, en particulier dans un contexte médical exigeant. C'est pourquoi l'équipe a décidé de le remplacer par Mistral Medium 3, un modèle plus puissant et mieux adapté aux exigences du projet. Actuellement, je teste avec 3 modèles en parallèle : GPT-4o, reconnu pour la qualité de ses générations mais nécessitant un hébergement cloud ; Gemma 3, rapide et open source, bien adapté à des déploiements maîtrisés, mais parfois moins pertinent sur le vocabulaire médical spécifique ; et Mistral Medium 3, open source, plus récent, permettant également un hébergement local et une gestion autonome des données sensibles. Cette phase de test permettra de déterminer quel modèle répond le mieux aux besoins concrets du service des urgences pédiatriques.

Chapitre 4 : Évaluation de l'interface chatbot et de la simplification des textes médicaux

4.1 Méthodologie appliquée à ERIOS assistant

Le projet s'articule en trois phases principales destinées à concevoir, tester et évaluer un assistant intelligent pour la rédaction de documents médicaux :

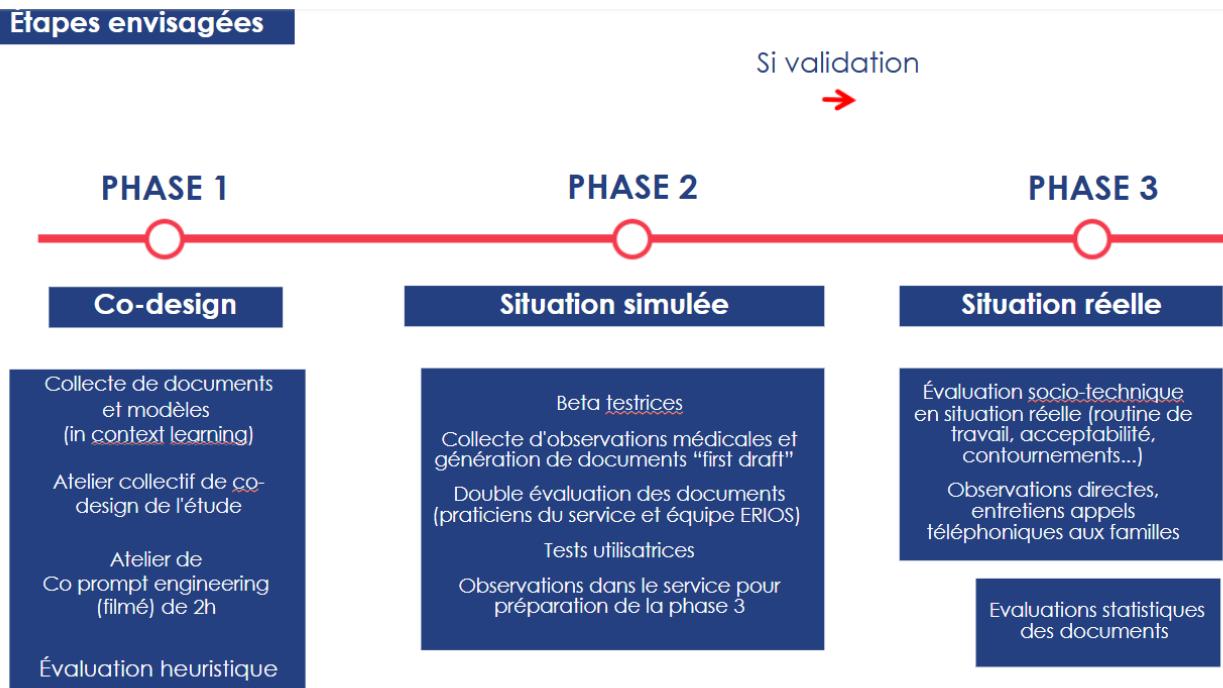


FIGURE 11 – Présentation des différentes phases du projet récupérée auprès de ma tutrice

Phase 1 – Co-design : Cette étape repose sur une co-construction avec les utilisateurs finaux. Elle comprend la collecte de documents cliniques, la création d'un premier prompt, ainsi que l'organisation d'ateliers collaboratifs de co-prompt engineering avec les professionnels de santé.

Phase 2 – Situation simulée : Le dispositif est testé à l'aide de cas fictifs ou anonymisés. Durant cette phase, des professionnels de santé issus de différents services interviennent en tant que bêta-testeurs et utilisent l'outil. Leurs retours sur les documents générés ainsi que ceux de l'équipe ERIOS permettent d'évaluer la qualité du document.

Phase 3 – Situation réelle : Si les résultats des phases précédentes sont satisfaisants, l'assistant est déployé en contexte clinique réel. Cette dernière phase vise à évaluer son intégration dans la pratique quotidienne, son acceptabilité par les utilisateurs et à revérifier la qualité des documents produits.

Avant mon intégration dans le projet, l'équipe avait déjà conduit une préparation approfondie en adoptant une démarche de co-design. Cette méthode repose sur l'analyse profonde du contexte d'usage réel pour créer des outils technologiques plus adaptés aux besoins des utilisateurs. Dans cette logique, un premier prompt a été élaboré, à partir d'une collecte de documents cliniques existants (lettres d'information, courriers médicaux, modèles institutionnels).

Un prompt est une instruction ou un ensemble de consignes, souvent textuelles, fournies à un modèle de langage pour orienter ou encadrer la génération de texte[23]. Il joue un rôle central dans le dialogue entre l'utilisateur et l'intelligence artificielle, en définissant les attentes quant au contenu, au style ou à la structure du texte produit. L'objectif de ce premier prompt était d'identifier les structures textuelles récurrentes, les registres de langage utilisés, ainsi que

les attentes implicites des professionnels de santé en ce qui concerne la communication écrite. Ce premier prompt, structuré au format JSON, visait à traduire ces demandes sous forme de règles claires que l'intelligence artificielle pouvait suivre lors de la génération de textes médicaux. Il constituait une base fonctionnelle de départ, conçue pour établir une communication entre l'outil technologique et les pratiques professionnelles.

Dans une optique de recherche qualitative rigoureuse, tous les ateliers ont été filmés, puis transcrits à l'aide de l'outil Dovetail, une plateforme collaborative permettant de coder, annoter et regrouper les retours exprimés par thématique. À partir de ces transcriptions, les besoins des professionnels ont été tagués puis organisés en quatre grandes catégories :

Contenus indésirables : éléments à éviter, tels que des formulations vagues, redondantes, ou des jugements de valeur non cliniques (Exemple : Il exagère ses douleurs..).

Contenus attendus : informations considérées essentielles pour la compréhension et la communication du message médical.

Besoins spécifiques : attentes relatives à la lisibilité, au ton utilisé, à la clarté des phrases ou au style rédactionnel (Exemple : ton neutre, sans subjectivité, éviter les expressions telles que « heureusement » ou « malheureusement »).

Format du document : structuration du document(titres, sous-sections), longueur, hiérarchie de l'information, mise en page(ordre des sections, niveaux de titres). Des enregistrements d'écran effectuées pendant les ateliers ont permis d'analyser les préférences des professionnels concernant l'ajustement des documents. Ces facteurs ont permis d'améliorer la réflexion sur la forme finale des prompts.



FIGURE 12 – Présentation de l'atelier de co-prompt engineering

À la suite de cette phase d'observation et d'analyse, une nouvelle version du prompt a été élaboré dans un fichier Excel A.1. Ce choix technique a été fait afin de garantir l'accessibilité de l'outil à l'ensemble de l'équipe pluridisciplinaire, quels que soient les niveaux de compétence technique. Afin de faciliter la lecture, la modification et la réutilisation du prompt, l'équipe a décidé de mettre en place un code couleur : le bleu indique les règles communes à tous les cas d'usage (structuration générale, neutralité du ton..), le vert correspond aux règles spécifiques aux lettres d'information , l'orange présente les règles propres à un cas d'usage particulier et le Jaune est réservé aux règles partagées entre tous les courriers médicaux destinés à des professionnels de santé.

4.2 Évaluation de l'interface :

4.2.1 Évaluation heuristique par des experts :

L'évaluation heuristique également appelée audit ergonomique est une méthode d'inspection experte qui consiste à faire analyser une interface par des spécialistes de l'ergonomie ou de l'expérience utilisateur (UX) sans faire appel directement aux utilisateurs finaux. Chaque expert examine l'interface écran par écran, en s'appuyant sur un référentiel reconnu de critères d'utilisabilité, tels que les 10 heuristiques de Nielsen (visibilité du statut du système, correspondance entre le système et le monde réel, contrôle par l'utilisateur, etc.) ou les 8 critères ergonomiques de Bastien et Scapin (guidage, charge de travail, gestion des erreurs, etc.). Cette technique initiée de base par Jakob Nielsen et Rolf Molich (1990), consiste à détecter un maximum de défauts techniques juste avant de mobiliser des utilisateurs réels ; elle complète donc les tests utilisateurs plutôt qu'elle ne les remplace.

Assistants

Découvrez et créez des versions personnalisées de **ERIOS Assistant** qui combinent des instructions, des bases de connaissances supplémentaires et des compétences.

Rechercher dans les assistants

Service concerné: Tous **Test Urgences Pédiatriques** Test Lettre de liaison et CR Ordonnances Document à délivrer

Type de document:

- 1 Lettre de liaison médicale URG PED
L'assistant va générer un compte rendu de le passage aux urgences à partir des données...
Verrouillé GPT4o Version 4
Azure HDS France
- 2 Lettre d'information famille URG PED
L'assistant va générer un document explicatif à destination de la famille susceptible d'être compris...
Verrouillé GPT4o Version 2
Azure HDS France
- 3 Lettre commun médecin/famille URG PED
L'assistant va générer à partir des notes médicales une lettre de liaison commune. Celle-ci contient...
Verrouillé GPT4o Version 4
Azure HDS France
- 4 Lettre d'information parents URG PED
L'assistant génère un texte explicatif à destination des parents de l'enfant suite à sa prise en charge...
Verrouillé GPT4o Version 5
Azure HDS France
- 5 Lettre d'information parents URG PED
L'assistant génère un texte explicatif à destination
- 6 Lettre d'information parents URG PED
L'assistant génère un texte explicatif à destination

FIGURE 13 – Explication de l'interface de ERIOS assistant

Test Urgences Pédiatriques - Lettre d'information parents URG PED

Type de document

L'assistant génère un texte explicatif à destination des parents de l'enfant suite à sa prise en charge aux Urgences

Vous pouvez échanger 3 fois avec l'assistant.

Pour utiliser cette assistant vous devez remplir les informations patient via le formulaire disponible en bas en droite de l'écran.

Écrivez votre message... 0/5000

ERIOS Assistant peut faire des erreurs. Vous devez impérativement relire l'ensemble du texte avant de le réutiliser.

FIGURE 14 – Explication de l'utilisation de l'interface de ERIOS assistant

Dans le cadre de mon stage, nous avons choisi de faire appel à cinq experts UX pour aboutir à la réalisation de notre étude. Le choix du nombre d'experts n'est pas aléatoire : En effet, il repose sur des fondements scientifiques et empiriques solides. D'après les travaux de Virzi (1992) ont démontré qu'un échantillon de cinq évaluateurs permet de détecter moyennement environ 75 à 80% des problèmes techniques liés à l'utilisabilité, tout en optimisant de façon maximale le temps utilisé et les ressources adéquates. Cette «règle des cinq» est de nos jours majoritairement acceptée et utilisée dans de nombreuses démarches UX axées utilisateur, bien qu'elle ne garantit pas une détection intégrale des défauts.

Durant cette évaluation, les experts recueillis précédemment, ont mené leur analyse d'étude indépendamment, en se focalisant exclusivement sur les neuf critères ergonomiques, dont nous avons jugé suffisamment pertinents pour notre étude d'évaluation médicale : la familiarité du système, l'esthétique et le minimalisme, la cohérence et le respect des standards, l'engagement et l'expérience utilisateur (UX), la fiabilité et la transparence, le contrôle et la liberté laissés à l'utilisateur, la gestion des erreurs et leur prévention, ainsi que l'aide, le feedback et l'accompagnement.

De manière indépendante, chaque expert a guidé l'évaluation, en ciblant précieusement les problèmes d'utilisabilités rencontrés, qu'il a par la suite trié en fonction de leur gravité (mineure, majeure, ou critique) et enfin proposé des améliorations pour chaque problème identifié. Les résultats détaillés issus de cette étude individuelle ont ensuite été rassemblés dans le but de souligner les problèmes majeurs identifiés ainsi que les recommandations adéquates proposées.

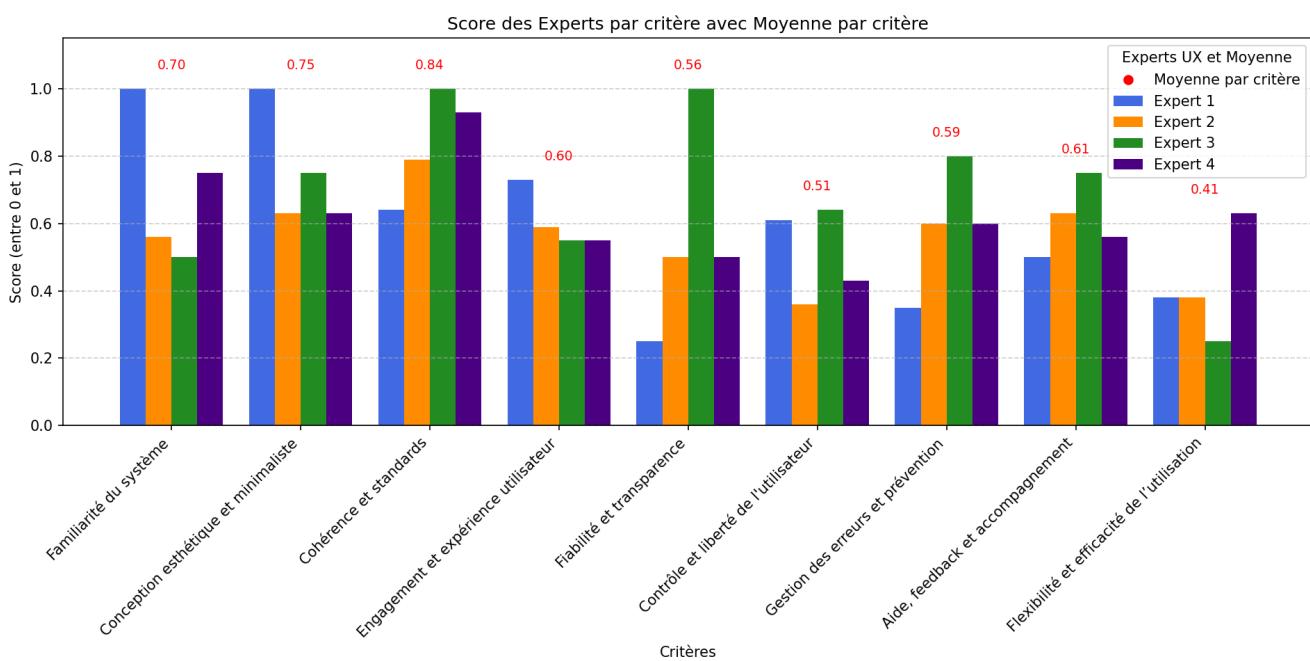


FIGURE 15 – Résultat des Experts UX par critère

Les résultats du graphique et de l'analyse des scores des experts UX révèlent une perception globalement positive de l'interface, mais soulignent également des domaines nécessitant des améliorations notables. Parmi les critères les mieux notés, la cohérence et le respect des standards (score moyen de 0,84) ainsi que la conception esthétique et minimalist (0,75) ressortent comme des points forts. Ces éléments sont particulièrement bien évalués par l'expert 1 et l'expert 4, qui ont délivré des scores proches du maximum, indiquant que l'interface a été visualisée comme claire et harmonieuse, tout en gardant le respect des normes du design numérique. La familiarité du système, avec un score moyen de 0,70, montre que l'interface a été perçue relativement intuitive, essentiellement par l'expert 1 et l'expert 4 qui ont montré une compréhension rapide du fonctionnement global de l'interface. Cependant, certains critères semblent présenter des points de

friction importants. Le plus marquant est la flexibilité et l'efficacité de l'utilisation, qui affiche la moyenne la plus faible (0,41), en particulier pour l'expert 3 (0,25) ainsi que pour les deux experts 1 et 2 (0,38), proposant que l'interface présente une insuffisance en terme de souplesse et peut être perçue comme contraignante. Ce manque de flexibilité est accentué par un score également faible dans le critère du contrôle et de la liberté de l'utilisateur (0,51), ce qui souligne une insuffisance d'options permettant aux utilisateurs de personnaliser leur expérience ou de corriger des erreurs. Le critère de fiabilité et transparence (0,56) montre également une perception hétérogène, avec un fort désaccord entre les experts, comme en témoigne le score parfait de l'expert 3 à l'inverse de la note la plus basse de l'expert 1 (0,25), proposant que certains utilisateurs doutent de la précision des informations fournies par l'interface. Enfin, les critères d'engagement et d'aide/feedback (0,61) ainsi que la gestion des erreurs (0,59) ont présenté des scores moyens, signalant que l'interface pourrait encore être améliorée afin de garantir une expérience utilisateur plus immersive et des fonctionnalités d'aide qui seraient plus efficaces. Pour conclure, bien que l'interface soit perçue positivement dans certains critères, il est important de se concentrer sur l'amélioration de la flexibilité, du contrôle de l'utilisateur et de la gestion des erreurs pour répondre aux attentes des utilisateurs et garantir une expérience plus fluide et adaptable.

Afin d'affiner l'analyse des résultats et de déterminer avec précision les axes prioritaires d'amélioration, une évaluation de la gravité perçue a été ajoutée pour chaque critère et pour chacun des utilisateurs. Contrairement aux scores de satisfaction, cette mesure indique l'impact négatif ressenti en cas de dysfonctionnement. Elle s'appuie sur une échelle de 0 à 4 : 0 : Pas de problème d'utilisabilité, 1 : Problème esthétique uniquement, 2 : Problème mineur d'utilisabilité, 3 : Problème majeur d'utilisabilité, 4 : Catastrophe en termes d'utilisabilité.

Ce croisement entre la performance perçue et la gravité des difficultés rencontrées permet de dépasser une simple lecture quantitative et d'identifier les critères dont les dysfonctionnements ont un réel impact sur l'expérience utilisateur, et même dans les situations où les moyennes paraissent raisonnables. Il s'agit donc d'un outil qui aide à optimiser les décisions dont le but est de guider les améliorations de manière orientée. L'analyse de ces données montre trois critères prioritaires :

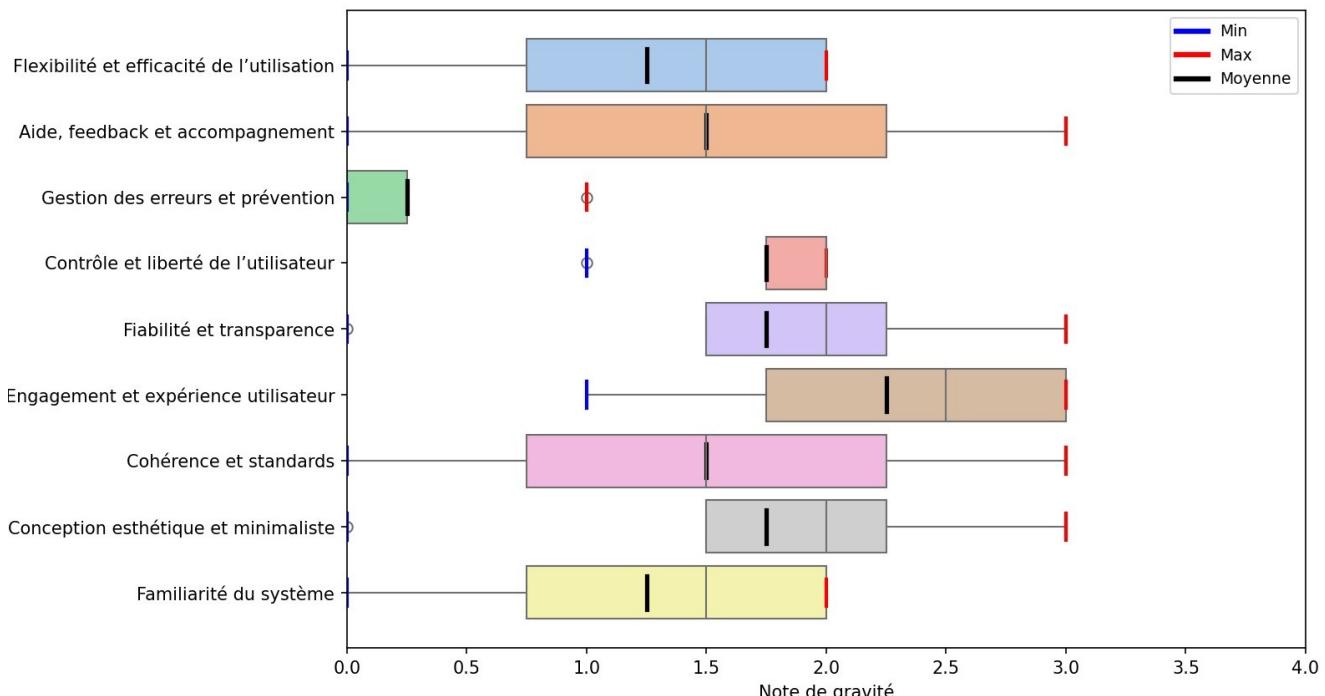


FIGURE 16 – Résultat des notes de gravité par critère

Dans un premier temps, le contrôle et la liberté de l'utilisateur affiche un degré de satisfaction relativement faible (0,51) associée à une gravité moyenne-haute (1,75). Cela reflète un sentiment

de contrainte important chez les utilisateurs, qui éprouvent à revenir en arrière ou à ajuster leurs actions sans difficulté. Au delà de cet aspect, la dispersion des valeurs dans la figure montre une certaine variabilité importante, avec des utilisateurs partageant des sentiments divergents sur ce critère. Ce constat indique qu'il semble exister une hétérogénéité dans l'expérience vécue, certains utilisateurs ressentent cette difficulté comme un obstacle majeur, tandis que d'autres le jugent moins perturbant. Cette dispersion implique que des solutions ciblées pourraient être plus efficaces pour certains profils que pour d'autres.

Dans un second temps, le critère lié à l'engagement et à l'expérience utilisateur combine un score d'efficacité plutôt simple (0,61) avec la gravité la plus élevée de l'ensemble (2,25). Ce résultat laisse penser que l'interface semble avoir des difficultés à capter l'attention ou à proposer une expérience suffisamment fluide, ce qui crée un certain désengagement perceptible. Les valeurs extrêmes observées sur la figure montrent une forte dispersion de la gravité, avec certains utilisateurs signalant des problèmes majeurs dans leur interaction, tandis que d'autres semblent moins affectés. Cette variabilité souligne que l'impact de ce critère pourrait être plus ou moins ressenti en fonction des attentes des experts, mais qu'il reste un axe prioritaire de réflexion pour maintenir l'engagement global.

Enfin, la fiabilité et la transparence obtient un score limité (0,56) et une gravité élevée (1,75), montrent des doutes exprimés par les utilisateurs par rapport à la clarté ou la crédibilité de certaines informations générées par le système. Le graphique montre une faible dispersion, ce qui permet d'affirmer que la majorité des utilisateurs partagent un ressenti similaire vis à vis de la fiabilité du l'outil. Cela renforce l'idée que ce critère nécessite une révision pour améliorer la transparence et la confiance. En revanche, certains critères se distinguent positivement. La gestion des erreurs par exemple présente une gravité très faible (0,25), ce qui démontre que les utilisateurs ne se sentent pas déstabilisés en cas d'erreur. Cette faible gravité se repère dans la faible dispersion des scores, illustrant une cohérence générale dans l'expérience des différents experts. De la même mesure la familiarité du système et la cohérence des standards obtiennent des scores satisfaisants accompagnés de niveaux de gravité faibles à modérés, suggérant une appropriation globale de l'interface assez correcte. La faible dispersion dans ces critères montre une perception homogène de la facilité d'utilisation, confirmant qu'il s'agit de points forts de l'interface.

L'analyse croisée des scores de performance et des niveaux de gravité permet de privilégier clairement les axes d'amélioration. Les critères les plus alarmants à corriger restent l'engagement utilisateur, le contrôle et la liberté d'action, ainsi que la fiabilité perçue du système, car ils présentent une faible efficacité et un fort impact négatif sur l'expérience. Dans une moindre mesure moins préconisée, mais nécessitant une piste d'amélioration rapide, on retrouve la flexibilité d'utilisation et l'aide/feedback, qui soulignent des lacunes fonctionnelles marquantes bien que perçues comme un peu moins graves. À l'inverse, certains éléments comme la gestion des erreurs, la familiarité du système ou encore l'esthétique, peuvent être abordés dans un second temps, car leur impact est jugé moins critique par les experts.

Dans une démarche d'approfondissement des résultats, nous avons décidé de faire une analyse fine des scores attribués à chaque question afin de différencier les difficultés rencontrées selon qu'elle relèvent de l'interface utilisateur ou du chatbot, nous avons examiné les scores question par question, afin d'identifier précisément les questions problématiques et de déterminer si celles-ci sont liées au fonctionnement du chatbot ou à celui de l'interface A.2 A.3 . Cette approche détaillée met en évidence que la majorité des problèmes identifiés concernent l'interface. En effet, tous les critères évalués tel que la familiarité du système, le contrôle et la liberté de l'utilisateur, la fiabilité, la flexibilité et l'efficacité de l'utilisation et enfin le critère qui porte sur l'aide et accompagnement présentent des scores négatifs significatifs sur certaines questions liées à l'interface. Cela traduit une mauvaise compréhension de certaines logiques d'usage, une navigation peu fluide, un manque de possibilités d'action ou de personnalisation, ainsi qu'une assistance parfois absente ou mal formulée. Ces éléments limitent l'efficacité de l'interaction et génèrent des confu-

sions chez les utilisateurs. À l'inverse, le chatbot semble globalement bien perçu sur la plupart des critères, à l'exception de l'engagement et expérience utilisateur, qui constitue un cas particulier. Sur ce point, les résultats indiquent une faiblesse partagée entre l'interface et le chatbot. Certaines questions montrent un manque d'implication ressenti lors de l'échange, proposant une interaction perçue comme trop passive ou monotone. Néanmoins, d'autres réponses positives montrent que la qualité du ton et la réactivité sont appréciées. En somme, cette analyse montre clairement que les axes d'amélioration prioritaires concernent l'interface utilisateur, qui montre des difficultés sur l'ensemble des dimensions évaluées. Le chatbot, pour sa part, présente un fonctionnement plus satisfaisant, à l'exception de certains aspects liés à l'engagement des utilisateurs.

4.2.2 Tests utilisateurs et retours qualitatifs :

Un test utilisateur, ou test d'utilisabilité, est une méthode empirique d'évaluation qui consiste à observer et interroger des personnes représentatives de la cible pendant qu'elles réalisent des tâches scénarisées sur un produit, qu'il s'agisse d'une maquette, d'un prototype ou d'une version en production. L'objectif principal est d'identifier les points de friction rencontrés par les utilisateurs, de mesurer leur efficacité, leur efficience ainsi que leur satisfaction, puis de formuler des solutions d'amélioration adaptées.

Dans le cadre de ce projet, des tests utilisateurs ont été effectués pour enrichir les évaluations heuristiques existantes. Ces tests, impliquant dix professionnels du service des urgences pédiatriques, ont offert un aperçu direct de l'expérience utilisateur et ont permis d'évaluer dans quelle mesure l'interface répond effectivement aux exigences du milieu médical. Pour cela, deux outils ont été privilégiés : le questionnaire SUS (System Usability Scale), un outil standardisé permettant de mesurer rapidement la satisfaction perçue par les utilisateurs, et le test CLEAR, permet d'évaluer la clarté ,la lisibilité et la facilité de l'utilisation d'un contenu ou d'un système. Ces tests utilisateurs, menés auprès de dix professionnels du service des urgences pédiatriques, ont permis de recueillir des données précises sur l'expérience réelle des utilisateurs et d'évaluer son adéquation aux besoins réels.

Dans un premier temps, un questionnaire préliminaire a été administré afin d'évaluer le profil des participants ainsi que leur familiarité avec l'intelligence artificielle générative. Par la suite, cinq scénarios ont été conçus à partir de données patients fictives. Pour chacun de ces scénarios, l'assistant de recherche technologique d'ERIOS, responsable de la conduite des tests, invitait les utilisateurs à accomplir des tâches spécifiques telles que la sélection d'un document, la saisie d'observations, la modification du contenu généré et l'exportation du document. Ces interactions étaient soigneusement observées et évaluées selon plusieurs critères prédéfinis, dans le but de mesurer l'adaptabilité des utilisateurs à ce nouvel outil technologique à travers différents scénarios d'utilisation.

À l'issue de ces scénarios, les utilisateurs ont complété le questionnaire SUS qui comprend dix questions alternant entre formulations positives (questions impaires) et négatives (questions paires). Chaque question est notée sur une échelle de 1 (« Pas du tout d'accord ») à 5 (« Tout à fait d'accord »). Les résultats obtenus sont présentés ci-dessous.

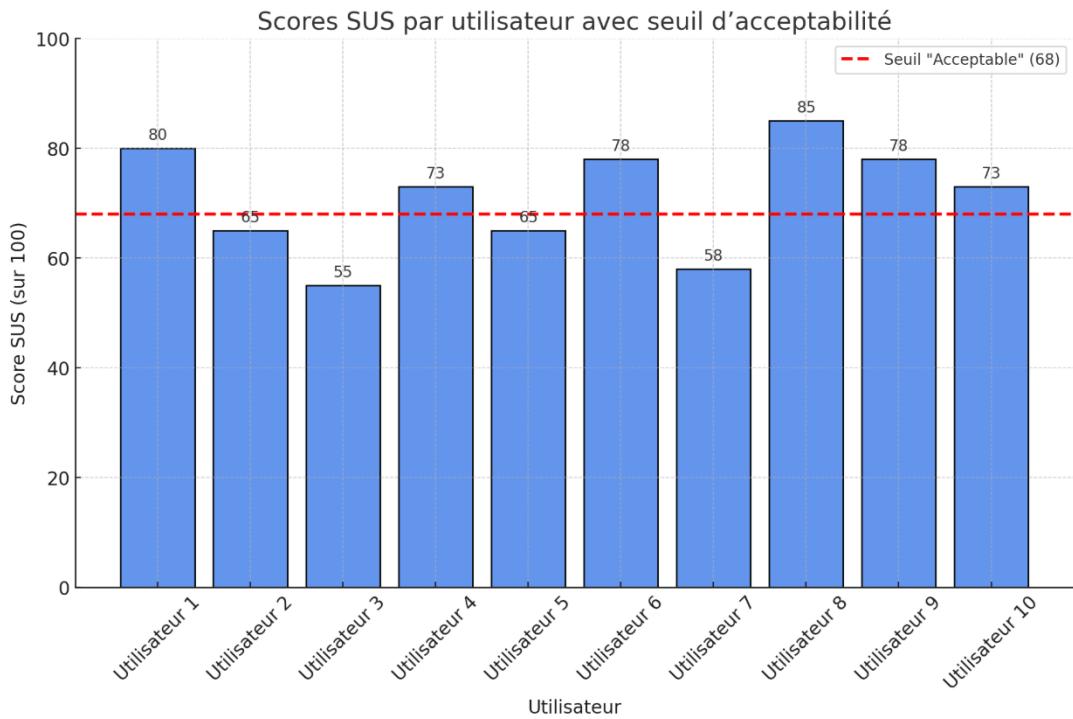


FIGURE 17 – Résultat des scores SUS par utilisateur

L'évaluation de l'utilisabilité à travers ce questionnaire révèle un score moyen de 70,75 sur 100, ce qui le situe au-dessus du seuil de 68, considéré comme la moyenne standard internationale pour ce type de questionnaire. Cette valeur moyenne indique que l'outil évalué présente une utilisabilité globalement satisfaisante. En effet, la majorité des utilisateurs a attribué des scores compris entre 70 et 80, correspondant à une appréciation positive et à une facilité d'usage considérée convenable, sans difficulté.

On peut voir que les deux utilisateurs (1 et 8) ont attribué un score supérieur à 80, ce qui traduit une très bonne utilisabilité selon les critères d'interprétation du SUS, avec une expérience utilisateur jugée comme fluide, agréable et intuitive. Ces scores élevés montrent que le système peut offrir une expérience optimale dans certains cas, probablement lorsque le profil de l'utilisateur correspond bien aux exigences ou au design de l'interface. À l'inverse, deux utilisateurs (3 et 7) ont attribué des scores respectifs de 55 et 58, donc inférieurs au seuil de 60, ce qui montre une perception plus critique du système, voire des difficultés d'utilisation. Bien que ces scores ne soient pas en-dessous du seuil critique de 50, ils restent alarmants et traduisent une expérience utilisateur moins satisfaisante, qui mériterait d'être vérifié plus en détail. De plus, on peut noter que l'écart-type est de 9,55 ce qui montre une variabilité modérée dans les réponses, signalant que les utilisateurs ne partagent pas tous une perception homogène du système.

Afin d'affiner l'analyse des résultats et de déterminer avec précision les retours des utilisateurs, une évaluation de la moyenne des réponses pour chaque question SUS a été ajoutée.

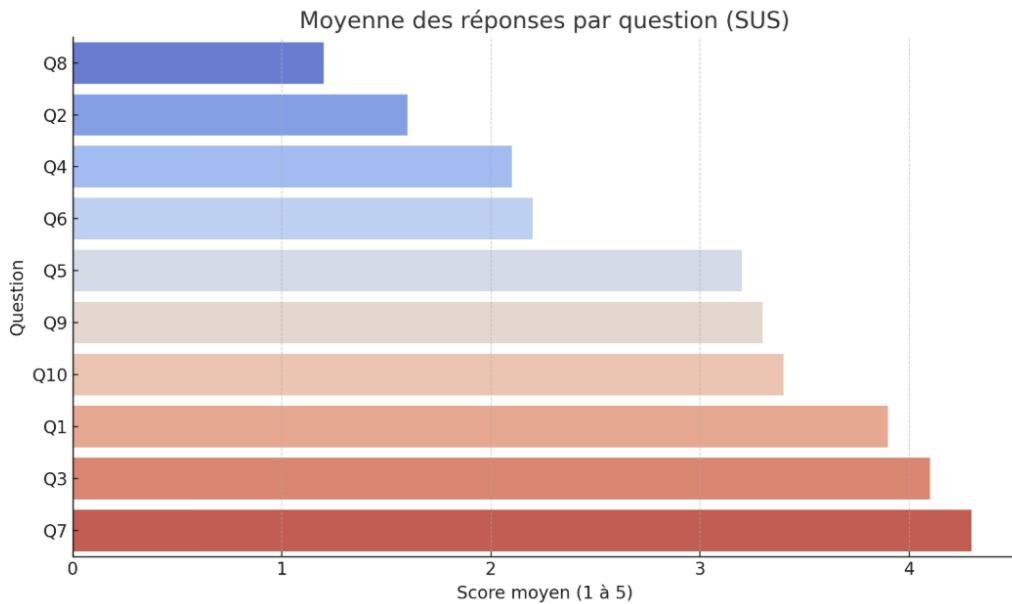


FIGURE 18 – Moyenne des réponses par question SUS

Les résultats issus de cette étude affirme une perception globalement positive de l'utilisabilité d'ERIOS Assistant, avec toutefois quelques nuances. La question 8, «J'ai trouvé le système très difficile à utiliser», affiche le score moyen le plus faible environ 1,3. Ce résultat indique que les utilisateurs sont en désaccord avec cette affirmation, ce qui est extrêmement favorable : ils ne trouvent pas le système difficile à utiliser. De même pour la question 2, «Je trouve ce système inutilement complexe» (1,6), qui montre également un désaccord fort, signifiant que les utilisateurs ne perçoivent pas de complexité inutile. La question 4, «Je pense que j'aurais besoin de l'aide d'un technicien pour pouvoir utiliser ce système» (2,1), indique un désaccord modéré : bien que la majorité semble se sentir autonome, une minorité exprime un besoin d'assistance, ce qui peut signaler des variabilités dans la prise en main selon les profils. Quant à la question 6, «Ce système présente trop d'incohérences» (2,2), indique que les utilisateurs ne perçoivent pas de manière marquée un manque de cohérence, bien que certains aient pu relever des éléments non homogènes. À l'opposé, certaines questions montrent des moyennes élevées, traduisant une expérience utilisateur satisfaisante. Pareil pour la question 7 «Je pense que la plupart des gens apprendraient très rapidement à utiliser ce système» (4,3), illustre une forte confiance dans la facilité d'apprentissage et la question 3, «Ce système est facile à utiliser» (4,1), confirme cette tendance : les utilisateurs s'accordent largement à dire que l'interface est simple et accessible. La question 1, «Je pense que j'utiliserai fréquemment ce système» (3,9), indique un intérêt pour un usage quotidien . Enfin, la question 10, «Je me suis senti très confiant en utilisant ce système» (3,5), montre une confiance relativement solide, bien que perfectible, dans la maîtrise de l'outil. En résumé, l'analyse des réponses montre que les utilisateurs rejettent les affirmations négatives, ce qui indique une bonne perception de la facilité d'utilisation, de la simplicité et de l'autonomie. En parallèle, ils valident les affirmations positives, ce qui confirme une expérience fluide, rassurante et globalement bien accueillie. Toutefois, quelques indices suggèrent des pistes d'amélioration, notamment autour de la cohérence visuelle et du soutien destinés à certains profils qui se montrent moins à l'aise avec l'outil. Ces éléments nous a permis de faire appel à une analyse croisée avec d'autres variables afin de garantir des améliorations ciblées pour répondre aux attentes des utilisateurs les plus critiques et garantir une utilisabilité optimale pour tous les profils A.4.

Dans un deuxième temps, nous avons soumis les utilisateurs au test CLEAR, qui se présente par un questionnaire composé de 5 questions, chacune évaluée par une échelle graduée de 1 à 5. Le niveau le plus bas (1) signifie «pas du tout d'accord», au contraire du plus élevé (5) qui signifie

«être complètement d'accord». Étant donné qu'ERIOS Assistant est un chatbot sans module RAG, il a été décidé d'adapter ce questionnaire en ne conservant que 4 questions. Une fois les résultats obtenus avec ce test, nous avons standardisé les données obtenus par l'intermédiaire des méthodes statistiques, dans le but de comparer les résultats obtenus avec les scores de référence du questionnaire entier. Cette approche m'a permis de garantir la possibilité de comparer les résultats obtenus du test avec les scores de bases références du questionnaire entier. Voici les résultats obtenus à l'issue de l'analyse des données recueillies lors du test CLEAR.

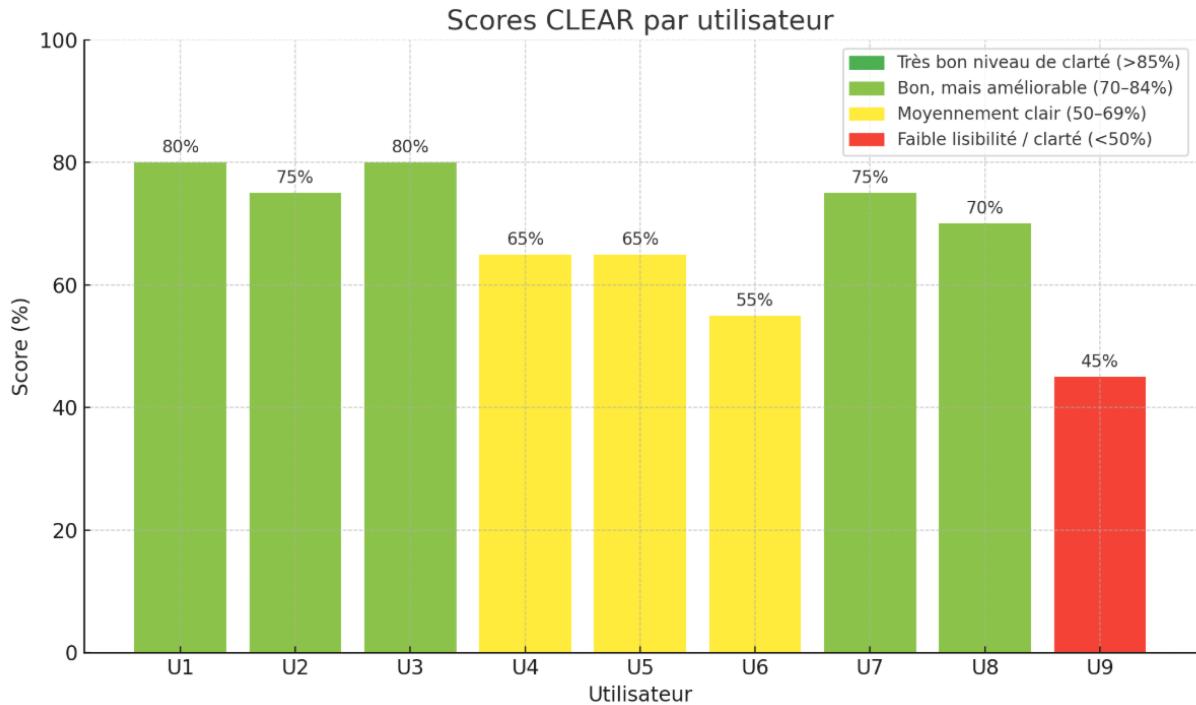


FIGURE 19 – Résultat des scores CLEAR par utilisateur

L'évaluation de la qualité des réponses fournies par le chatbot ERIOS Assistant, à travers le questionnaire CLEAR, montre un score moyen de 67,2 %, légèrement en dessous du seuil de 70 %, considéré comme indicateur d'un bon niveau de clarté. Cette valeur moyenne positionne l'outil dans la catégorie intermédiaire "moyennement clair", selon les critères d'interprétation du test CLEAR. On peut voir que la majorité des utilisateurs a attribué des scores compris entre 65 % et 75 %, montrant une satisfaction modérée : les réponses sont jugées globalement acceptables. Ce positionnement souligne une expérience d'usage ni pleinement fluide ni véritablement problématique, mais qui tendrait à être optimisée. Deux utilisateurs ont attribué un score de 80 %, atteignant ainsi le seuil de très bonne clarté. Ceci traduit une perception positive des contenus générés par le chatbot, jugés bien structurés et orientés vers l'essentiel. Ces résultats témoignent de la capacité du système à fournir des réponses de qualité élevée pour certains profils d'utilisateurs. À l'inverse, un utilisateur a attribué un score de 45 %, situant son expérience en dessous du seuil critique de 50 %, c'est-à-dire faible lisibilité. Ce résultat signale un niveau d'insatisfaction important, ce qui est expliquée par des réponses jugées ambiguës, ou mal formulées. Ce cas particulier mérite une attention spécifique. Pareil pour un autre utilisateur qui affiche un score de 55 %, indiquant une perception plutôt critique du système. Bien que ce score reste dans la zone dite "moyennement claire", il traduit que l'outil pourrait bénéficier d'un travail d'amélioration ciblée, notamment sur la clarté et la précision des formulations. Enfin, l'écart-type observé entre les scores est relativement modéré (environ 11 %), ce qui montre qu'il y'a une variabilité dans les réponses mais non extrême. Cela signifie que les utilisateurs ne perçoivent pas tous la qualité du contenu de manière uniforme. En somme, cette analyse quantitative révèle une performance

correcte mais encore insuffisante du chatbot en matière de clarté et de qualité des réponses. Ces résultats soulignent la nécessité d'un travail de raffinement sur les dimensions de formulation et de structuration, afin d'assurer un retour plus homogène pour tous les profils d'utilisateurs.

Afin d'affiner l'analyse des résultats et d'évaluer avec précision les retours des utilisateurs, une analyse globale sur l'ensemble des réponses au questionnaire CLEAR a été réalisée.

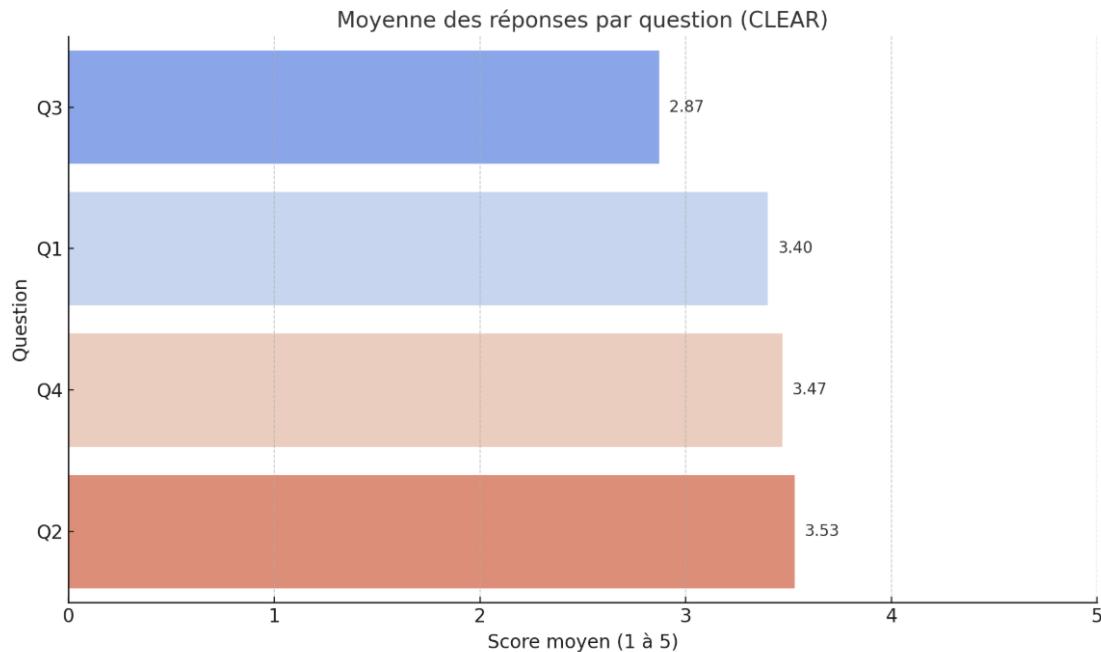


FIGURE 20 – Résultat de la moyenne des réponses par question CLEAR

Les résultats montrent une perception globalement favorable mais contrastée selon les dimensions évaluées. La question relative à la fiabilité du contenu (Q2) obtient la meilleure moyenne environ 3,53, indiquant que les utilisateurs estiment majoritairement recevoir des réponses conformes à leurs attentes. La question 4, portant sur la concentration du contenu sur l'essentiel, sans éléments superflus ni hors sujet, suit de très près avec une moyenne élevée de 3,47, suggérant une perception globalement positive sur la capacité de l'outil à aller à l'essentiel. Pareil pour la question 1, relative à la quantité d'informations obtient également une moyenne correcte environ 3,40, bien qu'il existe une variabilité plus marquée. En revanche, la question portant sur la clarté du contenu tel que(Q3) montre le scores moyens le plus faibles environ 2,87, ce qui indique des limites perçues en termes de lisibilité et de compréhension par rapport aux patients. Certains utilisateurs expriment une insatisfaction due à la présence d'ambiguïtés dans la reformulation.Ces résultats appellent à être affinés par une analyse croisée selon les profils des utilisateurs, afin d'identifier les conditions d'usage ou les caractéristiques individuelles susceptibles d'influencer la perception de qualité.

L'analyse croisée des réponses au questionnaire CLEAR selon les profils des utilisateurs tel que les tranches d'âge et l'ancienneté au CHU A.5 A.6 a permis de mettre en évidence des variations significatives dans la perception de la qualité des réponses fournies par le chatbot ERIOS Assistant. Ceci affirme l'importance d'ajustements ciblés. Les utilisateurs âgés de 45 à 54 ans et ceux disposant de plus de 20 ans d'ancienneté se montrent comme les plus satisfaits, évaluant très positivement la pertinence, la clarté du contenu. À l'inverse, ceux qui sont en poste depuis moins d'un an, ainsi que les utilisateurs de plus de 55 ans apparaissent en difficulté avec des scores très faible en terme d'exactitude et de clarté, ce qui suggère un besoin d'accompagnement renforcé ou une complexité dans la reformulations. Les autres profils, comme les professionnels

âgés de 35–44 ans ou ayant entre 2 et 10 ans d'ancienneté, adoptent des avis plus nuancés : ils reconnaissent certaines qualités de l'outil tel que la pertinence ou la lisibilité tout en exprimant des limites en termes de précision des informations. Cette hétérogénéité dans les évaluations invite à adapter l'interface et les contenus du chatbot pour répondre au besoin des différents profils ainsi d'améliorer l'expérience globale tout en préservant la qualité perçue par les plus expérimentés.

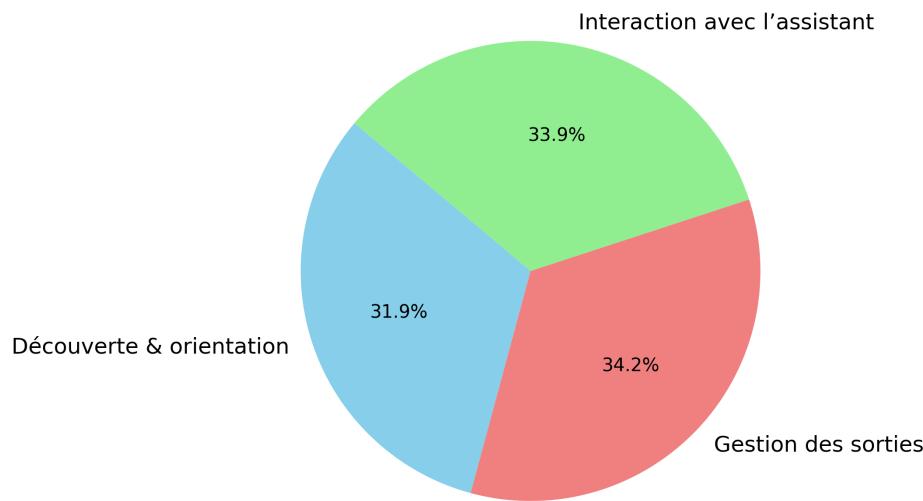


FIGURE 21 – Répartition des trois catégories évaluées

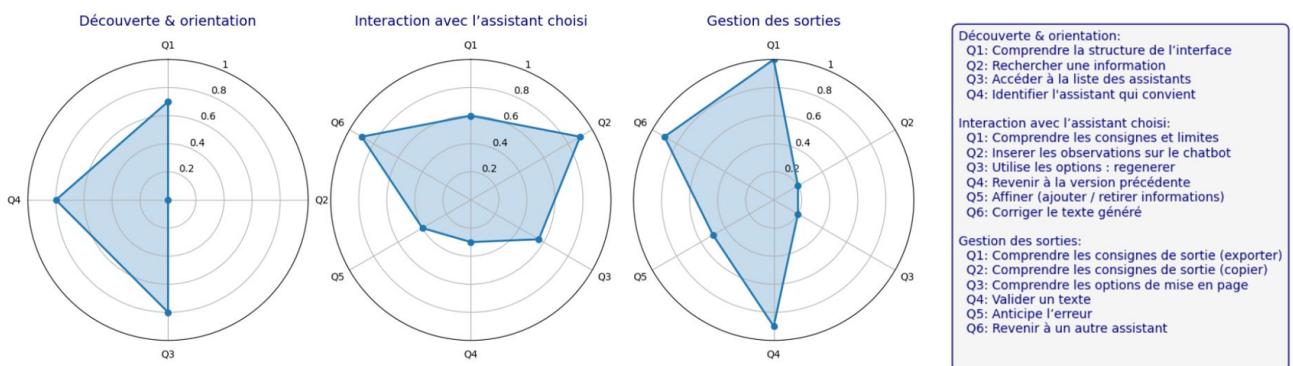


FIGURE 22 – Résultat des questions par catégories

L'évaluation portait sur la capacité de 10 utilisateurs à accomplir un ensemble de 16 tâches, réparties en trois catégories : Découverte et orientation, Interaction avec l'assistant et Gestion des sorties. Chaque tâche a été notée de manière binaire (1 = réussite, 0 = échec), et la moyenne obtenue pour chaque tâche constitue un indicateur clair de son accessibilité, sa compréhension et son exécution effective. L'analyse visait à identifier les points forts de l'interface, mais surtout à révéler les zones de friction nécessitant une amélioration. Les résultats montrent que les performances des utilisateurs varient sensiblement selon les catégories.

La première catégorie, « Découverte et orientation », avec 31,9 % des tâches, affiche une moyenne globale de 0,575, traduisant une appropriation partielle de la structure initiale de l'interface. Certaines tâches, comme « Accéder à la liste des assistants » et « Identifier l'assistant qui convient », obtiennent un score élevé (0,8), indiquant que ces fonctionnalités sont bien intégrées et facilement identifiables. En revanche, la tâche « Rechercher une information » enregistre un taux

d'échec total (0,0), ce qui révèle une faille majeure dans la conception de cette fonctionnalité. Cette difficulté peut être liée à un manque de visibilité de l'outil de recherche, à un libellé ambigu ou à une absence de guidance explicite dans l'interface.

La deuxième catégorie, « Interaction avec l'assistant », qui représente 33,9% des tâches, présente une moyenne légèrement supérieure, autour de 0,664. Les utilisateurs semblent à l'aise avec les actions élémentaires telles que « Insérer des observations » ou « Corriger le texte généré », qui atteignent toutes deux un score de 0,9. Ces bons résultats suggèrent une prise en main intuitive des fonctions centrales de l'assistant. En revanche, d'autres fonctionnalités apparaissent plus difficiles d'accès. C'est le cas de « Revenir à la version précédente » (0,3) ou « Affiner les informations » (0,4), qui affichent des taux de réussite faibles. Ces résultats laissent penser que certaines options avancées souffrent d'un manque de lisibilité ou d'un positionnement peu ergonomique, rendant leur usage peu évident, notamment pour des utilisateurs non experts.

Enfin, la troisième catégorie, « Gestion des sorties », couvrant 34,2 % des tâches, enregistre une moyenne de 0,628, avec des résultats très contrastés selon les tâches. Les actions de validation ou de navigation, telles que « Valider un texte » ou « Revenir à un autre assistant », sont bien maîtrisées (0,9), de même que l'exportation du contenu (1,0), ce qui montre que les fonctions finales les plus visibles répondent aux attentes des utilisateurs. À l'inverse, les tâches plus techniques, comme « Comprendre les consignes pour copier » et « Comprendre les options de mise en page », affichent chacune un score de seulement 0,2. Ces faibles performances révèlent un déficit de compréhension important, probablement causé par un manque d'explications contextuelles, un vocabulaire peu accessible ou une présentation peu claire des options disponibles.

Dans l'ensemble, cette évaluation met en évidence une interface globalement efficace sur ses fonctions principales. Cependant, les fonctionnalités plus avancées comme la recherche d'information ou la gestion détaillée des options restent difficilement accessibles, ce qui met en évidence la nécessité d'améliorer la conception de l'interface afin d'optimiser l'expérience utilisateur pour les différents profils.

Après avoir analysé les comportements des utilisateurs, nous avons identifié les biais liés à leur interaction avec le contenu généré, en tenant compte de leur profil technologique, notamment leur expérience et fréquence d'usage des outils d'IA générative, y compris en contexte médical. Cela nous a permis d'évaluer l'impact de ces profils sur l'apparition des biais. L'analyse s'est ainsi concentrée sur plusieurs types de biais récurrents les voici :

BIAIS	Description rapide	Exemples / Conséquences	Barrières / Actions possibles
COVERAWYSIATI	/ On se fie uniquement à ce qu'on voit sans chercher ce qui manque	Valider un texte fluide mais incomplet	Vérifier s'il manque des informations
Effet de Halo	Qualité du style influence confiance sur le contenu	Texte clair → impression de pertinence globale	Se concentrer sur le fond, pas juste sur la forme
Surconfiance / Expertise	Médecin comprend, mais suppose que le parent comprendra aussi	Texte complexe validé malgré inadéquation pour le parent	Penser au niveau du destinataire
Safety-I	Éviter que les choses tournent mal : détecter et corriger erreurs	Médecin détecte un problème et corrige le texte	Correction proactive, vigilance
Safety-II	Faire en sorte que tout se passe bien : analyser les ajustements et réussites	Médecin adapte spontanément le texte pour l'améliorer	Anticipation, adaptation, amélioration continue

Interruptions	Test de la capacité à revenir sur une tâche après coupure	Baisse possible de la vigilance, validation rapide	Gestion des interruptions, reprise du contrôle
----------------------	---	--	--

L'analyse des biais cognitifs pour chaque hypothèse montrent des résultats contrastés avec des scores moyens globalement faibles, variant entre 0,3 et 0,6 A.7. Le biais de sur-confiance, avec une moyenne de 0,4, montre qu'une proportion non négligeable d'utilisateurs accepte encore des réponses erronées sans les remettre en question. Ce chiffre reste préoccupant dans un contexte où les conséquences d'une mauvaise interprétation peuvent être graves, notamment en santé. Toutefois, les scores obtenus pour l'effet de halo (0,3) et le biais WSYATU (0,3) sont plus encourageants. Contrairement à ce que ces biais suggèrent, les utilisateurs ne se sont pas laissés tromper par la simple apparence d'expertise de l'IA (« effet IA médicale »), ni par la première réponse reçue. Autrement dit, la majorité des participants ont conservé une distance critique, en cherchant à vérifier, compléter ou comprendre davantage les informations fournies, ce qui est un signe positif de discernement. Par ailleurs, les deux comportements liés à la sécurité proactive affichent les scores les plus élevés (0,6 chacun), ce qui traduit une tendance majoritaire à corriger un texte incorrect ou à reformuler un contenu déjà clair pour le rendre encore plus compréhensible. Ces actions témoignent d'un certain niveau de responsabilité dans l'usage de l'outil, et d'une volonté d'assurer la qualité des productions textuelles, même si tout n'est pas encore optimal. L'ensemble des résultats suggère donc une vigilance globale des utilisateurs face aux réponses générées par l'IA. Toutefois, l'existence persistante de comportements de sur-confiance appelle à renforcer les dispositifs de vérification et de sensibilisation, afin de limiter les risques liés à une confiance excessive ou non critique dans les outils d'intelligence artificielle.

Dans le but d'avoir une analyse profonde, nous avons cherché à croiser les résultats des biais cognitifs avec les profils d'utilisation des outils d'IA générative. L'objectif était de mieux comprendre comment le niveau d'exposition et la nature des usages influencent la vigilance critique face aux réponses produites par l'IA. Les résultats montrent que les utilisateurs expérimentés et réguliers, comme l'utilisateur 7, développent une posture critique équilibrée, capable de compenser certains biais (sur-confiance, effet de halo) par une bonne détection et correction des erreurs ainsi qu'une reformulation active des réponses. À l'inverse, les utilisateurs peu familiers ou occasionnels (utilisateurs 9 et 10) manifestent une vigilance moindre, avec une forte propension à accepter les erreurs sans vérification, ce qui augmente les risques d'utilisation non critique. Les profils intermédiaires présentent quant à eux une posture en construction, oscillant entre prudence et automatismes, témoignant d'une prise de conscience progressive des enjeux liés à l'IA. Ces observations soulignent l'importance d'une formation ciblée pour favoriser une utilisation raisonnée et réflexive des outils d'IA, afin de réduire les risques associés aux biais cognitifs dans le contexte médical.

4.3 Analyse comparative de la simplification des textes médicaux :

4.3.1 Tests préliminaires et ajustements :

Dans cette deuxième phase du projet, centrée sur la situation simulée, Dans cette deuxième phase du projet, centrée sur la situation simulée, nous avons continué à élaborer et optimiser un prompt initial pour la génération de documents cliniques à l'aide de modèles de langage. Ce travail d'amélioration du prompt est resté un processus itératif, mené en parallèle de la génération et de l'analyse des documents produits.

nous avons testé le prompt initial avec trois modèles de langage : GPT-4o, Gemma 3, et Mistral Large . Pour chacun de ces modèles, nous avons généré des documents à partir du prompt, puis analysé les résultats afin d'identifier les points faibles, les manques ou les incohérences. Cette phase itérative nous a permis d'ajuster progressivement le prompt, en l'adaptant aux objectifs

de notre projet ainsi qu'aux spécificités de chaque modèle testé. Une fois un prompt suffisamment robuste et cohérent obtenu, nous avons entamé la phase de génération des textes. C'est à ce moment-là que notre partenaire technique, Mistral, a informé notre équipe de la mise à disposition d'un nouveau modèle, Mistral Large, une version améliorée du Mistral Medium. À leur demande, nous avons intégré ce nouveau modèle dans notre dispositif expérimental, afin de tester ses performances dans les mêmes conditions que les autres modèles précédemment évalués. Par ailleurs, nous avons décidé de structurer notre expérimentation autour de cinq types de documents représentatifs des usages cliniques envisagés : Lettre d'information à destination des parents, Lettre d'information à destination des enfants/adolescents, Lettre d'information à destination de la famille, Lettre de liaison commune.

Pour chacun de ces types de documents, nous avons mobilisé une observation médicale complexe issue de situations cliniques réelles anonymisées. Ces observations ont été choisies en raison de leur richesse descriptive, permettant d'explorer la capacité des modèles à gérer un vocabulaire varié, des termes médicaux spécialisés, différents niveaux d'explication, ainsi que des enjeux de clarté et d'adaptation au public cible. À partir de chaque observation clinique, nous avons généré les différents types de lettres en testant un modèle de LLM spécifique, puis affiné le prompt si nécessaire. Ainsi, chaque type de document a donné lieu à une version de prompt distincte, adaptée au format, au destinataire et au contenu clinique sous-jacent. Cette démarche nous a permis d'évaluer de manière plus fine la performance des modèles dans des contextes variés, proches de situations professionnelles réelles.

4.3.2 Modèles finaux retenus :

À l'issue des tests préliminaires et des ajustements, nous avons retenu trois modèles de langage pour la phase finale d'expérimentation : GPT-4o, Gemma 3 et Mistral Large. L'objectif était de comparer leurs performances sur des cas cliniques concrets et représentatifs. Pour cela, nous avons décidé de générer 150 lettres par type de document, réparties équitablement entre les trois modèles (50 lettres par modèle). Chaque type de lettre a été associé à une situation médicale spécifique, choisie pour sa pertinence clinique et sa capacité à faire émerger des différences dans les documents générés par les différents modèles. Ainsi, la lettre d'information destinée aux parents portait sur un cas de gastroentérite d'un enfant de 1 an, pour lequel nous avons utilisé la version 5 du prompt. La lettre pour enfants et adolescents sur une perte de connaissance d'un enfant de 15 ans, également avec la version 5 du prompt, la lettre à destination de la famille sur un cas d'une traumatologie d'un enfant de 10 ans, utilisant la version 2 du prompt et la lettre de liaison commune sur une situation de bronchiolite d'un bébé de 2 mois avec une version 4 du prompt. Pour le dernier cas d'usage, nous avons généré un total de 150 lettres, dont 75 destinées aux parents et 75 aux enfants. Là encore, chaque sous-groupe a été réparti entre les trois modèles, avec 25 lettres générées par modèle pour chaque destinataire. Le cas étudié porte sur un cas d'une crise clastique chez un enfant de 10 ans, traité pour l'instant avec la version 4 du prompt. Le travail sur le prompt est toujours en cours pour ce cas délicats, en raison de leur complexité clinique et linguistique. Cette répartition équilibrée et ces choix de cas cliniques ont été pensés pour garantir une diversité de registres lexicaux, de niveaux de technicité et de formulations, afin de mieux évaluer la capacité des modèles à produire des contenus informatifs, accessibles et adaptés aux différents publics visés.

4.3.3 Méthodes d'évaluation :

Dans le cadre de l'évaluation de la lisibilité des textes générés par les modèles de langage, nous avons initialement envisagé d'utiliser des métriques classiques telles que les scores de Flesch Reading Ease et de Flesch-Kincaid Grade Level, largement répandus dans les études de lisibilité. Ces formules, développées à l'origine pour la langue anglaise, estiment la difficulté de lecture à partir de la longueur moyenne des phrases et des mots (en nombre de syllabes), et sont particuliè-

rement utilisées dans les contextes éducatifs et médicaux. Cependant, au cours de notre revue de littérature, nous avons constaté que ces formules sont peu adaptées à la structure linguistique du français, notamment en raison des différences morphosyntaxiques et phonétiques entre les deux langues. Cette observation nous a conduits à approfondir nos recherches afin d'identifier des formules de lisibilité spécifiquement conçues ou adaptées pour la langue française. C'est ainsi que nous avons retenu trois indicateurs pertinents : le score de Flesch Douma, une adaptation franco-phone du Flesch original ; la formule de Kandel-Moles, développée dans le cadre de l'analyse de textes éducatifs en français ; et l'indice LIX (Lire Index), initialement scandinave mais souvent utilisé pour des textes en français, notamment dans les travaux comparatifs de lisibilité.

Ces trois formules tiennent compte de critères mesurables tels que le nombre moyen de mots par phrase, la proportion de mots longs (définis en français comme ayant plus de neuf lettres) ou encore la complexité syntaxique globale. Après avoir appliqué ces indicateurs à l'ensemble des textes générés, nous avons procédé à une analyse statistique visant à évaluer la stabilité des scores au sein de chaque modèle. Les textes ont été répartis aléatoirement en plusieurs groupes, et un test de normalité de Shapiro-Wilk a d'abord été réalisé pour déterminer si les données suivaient une distribution normale. En fonction des résultats, nous avons choisi soit une ANOVA (analyse de variance) pour les distributions normales, soit un test de Kruskal-Wallis, non paramétrique, lorsque la normalité n'était pas respectée. Ces tests ont permis de vérifier l'existence ou non de différences significatives entre groupes, ce qui constitue un indicateur clé pour mesurer la cohérence et la stabilité des performances linguistiques des modèles étudiés.

4.3.4 Résultats obtenus et interprétations :

Dans ce premier cas d'usage, il s'agit d'évaluer la capacité des LLM à générer une lettre d'information à destination des parents, dans un contexte sensible : un cas de gastroentérite concernant un enfant âgé d'un an. Ce type de communication exige une grande clarté, une structure accessible et un vocabulaire compréhensible pour un public non spécialiste, souvent confronté à des émotions telles que l'inquiétude ou l'anxiété. C'est pourquoi l'analyse de la lisibilité des textes produits constitue un indicateur pertinent pour comparer les performances des modèles utilisés (Mistral, Gemma et ChatGPT).

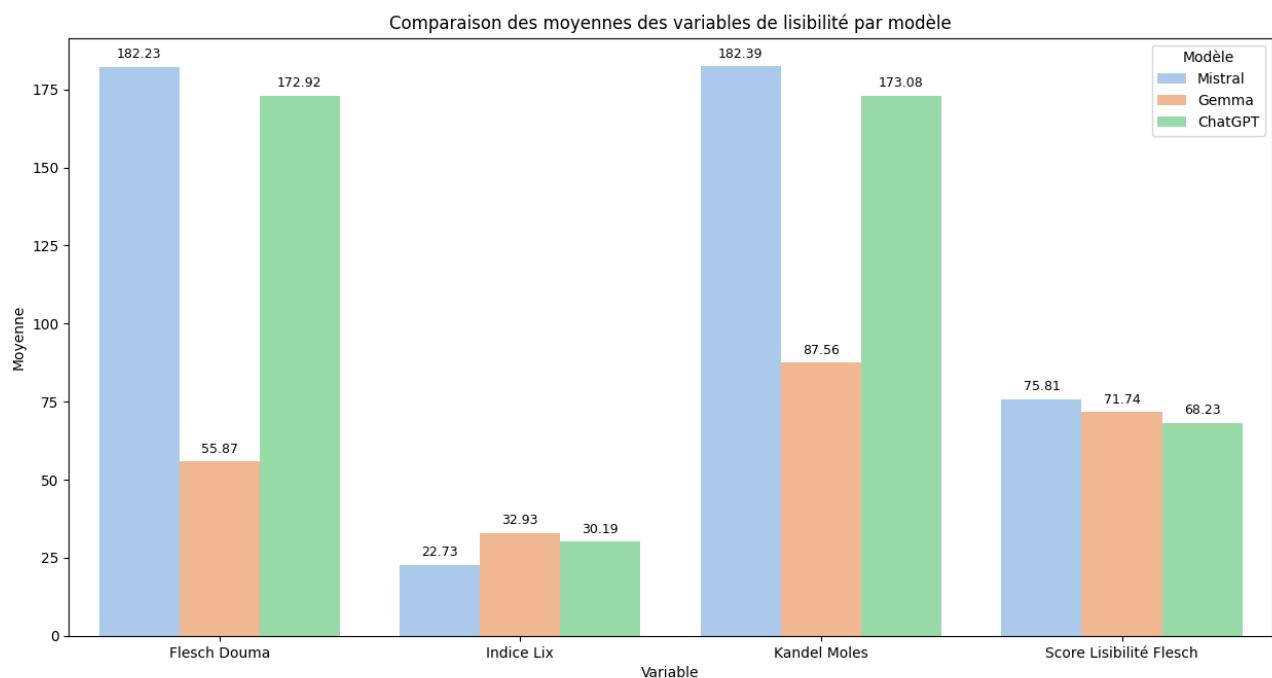


FIGURE 23 – Résultat des différents scores de lisibilité pour les 3 LLM

Les résultats, représentés dans le graphique ci-dessus, montrent des disparités marquante en fonction des indices de lisibilité représentés. Sur l'indice Flesch Douma, qui évalue la complexité d'un texte (plus le score est haut, moins il est lisible), le modèle Gemma se démarque nettement grâce à un score bien inférieur à ceux de Mistral et ChatGPT. Cette écart suggère que les textes produits par Gemma sont plus simples et plus fluide en accessibilité, ce qui constitue un avantage majeur pour la compréhension du message par les parents. Une tendance similaire est observée sur l'indice Kandel et Moles, qui mesurent la complexité lexicale : Gemma montre un score (87,56) bien inférieur par rapport à Mistral et ChatGPT qui ont obtenu respectivement des notes de 182,39 et 173,08, renforçant l'idée que ses textes sont plus faciles à comprendre sur le plan du vocabulaire utilisé. En revanche, le modèle Mistral s'est distingué en obtenant la meilleure performance en terme de score de lisibilité Flesch (75,81), un indicateur qui évalue directement la facilité de lecture (plus il est élevé, plus le texte est facile à lire). Ce score performant indique que, malgré une complexité plus importante perçue selon d'autres indices, le texte généré par Mistral reste globalement fluide et structuré. Enfin, ChatGPT affiche des résultats moyens sur l'ensemble des indicateurs, accompagné d'une lisibilité satisfaisante mais une complexité lexicale légèrement plus marquée que Gemma.

En résumé, cette étude met en évidence l'importance du choix du modèle selon les contraintes du contexte : si l'objectif est de maximiser la compréhension immédiate et intuitive du texte, notamment pour un public non expert, Gemma semble être l'outil le plus adéquat. Cependant, si l'on cherche à maintenir un équilibre entre variété du contenu, clarté syntaxique et accessibilité, Mistral ou ChatGPT se présentent comme des alternatives pertinentes, à condition d'adapter ou simplifier le texte en aval si nécessaire. Cette première comparaison souligne ainsi les enjeux de lisibilité dans la génération de contenus médicaux et l'impact direct qu'un choix de modèle peut avoir sur l'expérience des utilisateurs.

Afin de compléter l'analyse comparative des performances des modèles, des tests intra-modèle ont été effectués afin d'évaluer leur stabilité interne. Cette dernière correspond à l'aptitude d'un modèle à produire des textes homogènes lorsqu'il est soumis à des règles identiques. Pour cela, des classes de textes triées aléatoirement ont été comparés à l'aide de tests statistiques (ANOVA et Kruskal-Wallis), en se basant sur plusieurs indicateurs de structure (nombre de mots, de phrases, mots longs...) et de lisibilité.

Les résultats pour ChatGPT A.8 montrent une bonne stabilité globale satisfaisante, avec l'absence de différences significatives sur la majorité des variables. Toutefois, la longueur moyenne des phrases ainsi que les indices de lisibilité Flesch Douma et Kandel-Moles affichent des variations significatives ($p < 0,05$), ce qui souligne une certaine variabilité dans la complexité linguistique des textes. Le modèle reste correct et cohérent dans l'ensemble, mais peut adopter des formulations plus ou moins difficiles selon les différents cas. En revanche, le modèle Mistral se démarque par un équilibre remarquable. Tous les indicateurs testés affichent des résultats non significatifs, indiquant que le modèle produit des textes hautement homogènes, tant sur le plan structurel que lexical, ce qui représente un avantage notable dans des contextes de génération répétée ou de production automatisée à grande échelle. En ce qui concerne Gemma, les résultats montrent une stabilité moyenne. Aucune différence significative n'est détectée, mais certaines valeurs tels que la diversité lexicale s'approchent du seuil de significativité, laissant présenter une légère sensibilité aux variations internes. Au delà de tout ceci, les textes générés restent globalement cohérents.

Pour conclure, cette analyse a permis de classer les modèles selon leur stabilité : Mistral est le plus stable, suivi par Gemma, puis ChatGPT, qui demeure néanmoins fiable. Ce critère de stabilité soutient les résultats sur la lisibilité, mettant en lumière l'optique professionnels exigeant une constance stylistique et linguistique.

Après avoir réalisé plusieurs analyses, nous avons cherché à identifier un phénomène particulier et à regarder le comportement des LLM vis-à-vis de celui-ci. Nous avons donc décidé de nous concentrer sur le traitement de l'expression « temps de recoloration capillaire », afin

d'observer si elle était reformulée, mentionnée sans explication, mentionnée avec une explication, ou simplement absente des lettres générées. L'analyse démontre des approches contrastées entre les modèles face à cette expression A.9. Mistral ne la cite pas dans toutes les lettres générées, ce qui reflète une stratégie d'évitement des expressions techniques, probablement au profit d'une simplification maximale. Gemma, quant à lui, affiche une tendance à l'explication. En effet, sur 50 lettres l'expression est expliquée dans 10 cas et reformulée dans 5, sans jamais apparaître sans clarification. Cette constance démontre une volonté claire de rendre le contenu compréhensible pour un public non experts. ChatGPT adopte une position plus délicate. L'expression est reformulée dans 6 lettres, expliquée dans 1, citée sans explication dans 1, et absente dans 42 cas. Ce comportement exprime une logique moins homogène, où le modèle alterne entre simplification et reformulation.

Ces écarts mettent en lumière des choix implicites dans la gestion de la technicité : effacement chez Mistral, vulgarisation systématique chez Gemma, et flexibilité contextuelle chez ChatGPT. Le choix du modèle a donc un impact direct sur la clarté des contenus, notamment dans les communications sensibles destinées à un public non expert.

Au regard de l'ensemble des résultats, le choix du LLM dans ce cas d'usage s'avère déterminant. Gemma se distingue par une bonne vulgarisation. Mistral, plus stable, perd parfois sa précision et chatGPT reste intermédiaire, avec une clarté inégale. En définitive, le choix du modèle doit être guidé par le public cible et les exigences de lisibilité, en particulier dans des contextes sensibles où la compréhension du message est primordiale.

Dans ce second cas d'usage, l'objectif principal fut similaire : il avait pour but d'évaluer l'aptitude des modèles à générer une lettre d'information fluide et cohérente, mais cette fois-ci livrée à un public moins âgé, composé essentiellement d'enfants et d'adolescents dans le contexte d'un épisode de perte de connaissance. L'approche d'analyse reste inchangée : comparaison des scores de lisibilité pour mesurer l'accessibilité des textes A.10, et identification de phénomènes linguistiques spécifiques afin d'observer comment les LLM gèrent certains termes ou notions sensibles. L'objectif est de mieux comprendre comment chaque modèle ajuste son discours à destination d'un public sensible ou à une situation probablement stressante.

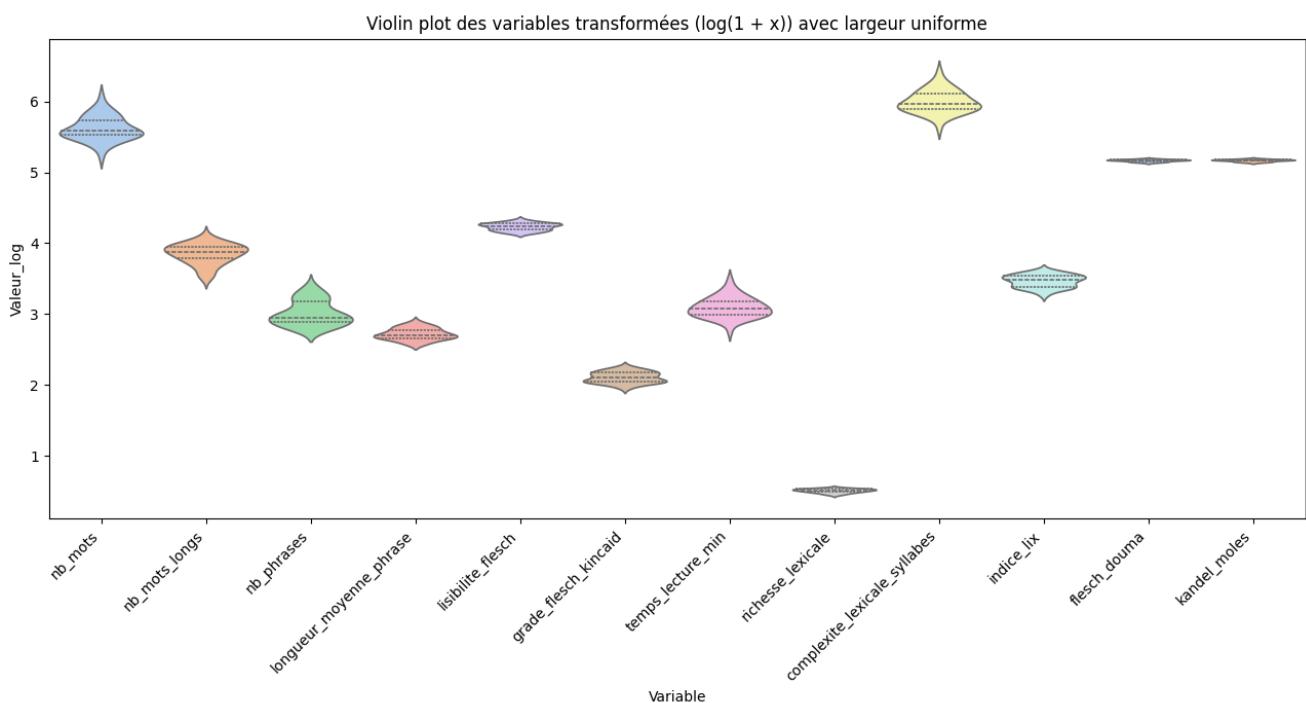


FIGURE 24 – Dispersion des différentes variables du modèle Mistral

L'analyse des trois diagrammes en violon met en évidence des différences claires dans la stabilité des écrits produits par les modèles GPT, Mistral et Gemma A.11 A.12. Le modèle GPT présente des distributions très concentrées autour des médianes pour la majorité des variables, telles que le nombre de mots, le nombre de phrases, le temps de lecture estimé ou encore les indices de lisibilité (Flesch, Kincaid, Lix, Douma). Cette faible dispersion souligne une certaine constance dans les textes produits, impliquant un comportement coordonné et homogène du modèle, et également sur des informateurs complexes comme la richesse du texte ou la complexité lexicale. À l'inverse, Mistral montre des distributions plus dispersées, traduisant une variabilité plus marquée dans les textes produits. Cela est particulièrement visible sur les variables corrélées à la structure textuelle (nombres de mots, la longueur moyenne des phrases), mais également sur d'autres indicateurs de lisibilité, ce qui peut refléter une sensibilité plus forte aux variations d'instructions ou de contexte dans les prompts. Le modèle Gemma quant à lui se situe au milieu de ces 2 extrêmes : ses distributions sont dans la majorité regroupées, notamment pour les variables liées à la complexité lexicale et à certains indices de lisibilité, mais un peu plus dispersées sur les mesures structurales comme la longueur moyenne des phrases ou le temps de lecture. Cette position intermédiaire suggère que Gemma cherche un équilibre entre stabilité et adaptabilité. En résumé, GPT se distingue par une solidité et une régularité parfaite pour des tâches ayant recours à une production textuelle normalisée, Mistral offre une plus grande plasticité textuelle potentiellement utile dans des contextes créatifs ou exploratoires, tandis que Gemma propose une solution médiane adaptée à des usages mixtes.

Afin d'approfondir notre analyse des documents générés dans le cadre de ce cas d'usage, nous avons décidé d'étudier la densité et la diversité des termes médicaux présents dans les textes. Cette étape vise à évaluer la richesse lexicale et la précision du langage médical utilisé, en identifiant les types de vocabulaire mobilisés et leur fréquence. Pour ce faire, nous avons procédé à une classification manuelle des termes médicaux en cinq catégories principales :Examens médicaux, Symptômes, Diagnostics, Maladies, Médicaments. L'analyse vise à comparer les trois modèles de langage étudiés en identifiant, pour chaque document, la variété des termes utilisés par catégorie, ainsi que les expressions spécifiques à chaque LLM. Ceci contribue à localiser les différences lexicales entre les 3 modèles et de noter leur capacité à générer un contenu médical détaillé, hétérogène et adapté au contexte clinique.

L'analyse des lettres produites par les trois modèles Mistral, GPT, et Gemma révèle des différences notables dans la fréquence des termes médicaux, traduisant des variations dans la richesse descriptive, la précision clinique et les centres d'attention A.13. Le modèle Mistral fait un usage très soutenu de termes techniques liés aux examens médicaux comme EEG (77 occurrences), scanner (50), et ECG (38), et met l'accent sur des expressions symptomatiques précises comme faiblesse (68), douleur dans la poitrine (78), ou fatigue (64). Il emploie également davantage de diagnostics neurologiques spécifiques comme hémiplégie gauche (39) et séquelle (29), suggérant un niveau de détail et de pathologisation élevé. À l'inverse, GPT présente une distribution plus modérée de ces termes, avec une mention plus fréquente de "perdu connaissance" (39) et une occurrence plus faible de termes techniques. Ce modèle se distingue aussi par un emploi limité des diagnostics complexes, comme en témoigne l'absence de mention d'hémiplégie ou de pathologies intestinales graves. De son côté, Gemma produit un discours dense en termes d'examens médicaux (notamment ECG = 76, EEG = 93, scanner = 59) mais beaucoup plus pauvre en termes de diagnostics détaillés. On note un nombre élevé d'occurrences de symptômes vagues comme faiblesse (71) et perdu connaissance (41), sans pour autant développer des diagnostics neurologiques poussés. De plus, Gemma surutilise certains examens comme la radiographie (44), mais reste imprécis sur les interprétations. Enfin, les trois modèles diffèrent aussi quant à leur traitement des maladies spécifiques : Mistral mentionne l'invagination intestinale (24) et le syndrome du grêle court (20), ce qui n'apparaît presque pas chez GPT, tandis que Gemma mentionne davantage torsion de l'intestin (10) et reste partiel sur d'autres pathologies.

Dans ce troisième cas d'usage, qui porte sur une lettre destinée à la famille d'un patient pris en charge pour un cas de traumatologie, nous avons décidé d'analyser deux segments phonétiques afin d'observer le comportement des trois modèles de langage (LLM). Nous nous sommes notamment intéressés à l'explication du terme « accès intraveineux ». Nous avons également identifié que l'expression « temps de recoloration » apparaissait dans l'observation médicale associée à ce cas. Comme ce même terme figurait déjà dans le premier cas d'usage et constituait l'un des phénomènes que nous avions analysés, nous avons décidé de vérifier s'il était reformulé de manière similaire dans les deux cas.

Pour commencer, nous nous sommes intéressés au traitement du terme « accès intraveineux » dans les lettres générées par les trois modèles, afin de déterminer s'il était reformulé, expliqué ou simplement cité sans précision

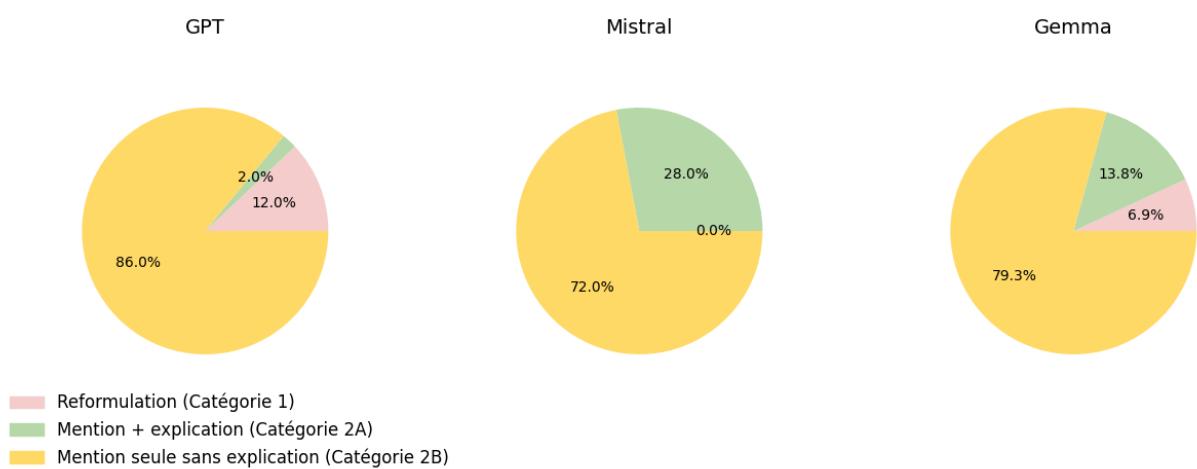


FIGURE 25 – Comparaison des 3 modèles au phénomène cherché

Les résultats montrent des différences notables entre les modèles. Le modèle GPT mentionne le terme dans 86 % des cas sans fournir d'explication, ce qui traduit une tendance à reproduire le vocabulaire médical sans le contextualiser. Seuls 12 % des cas relèvent d'une reformulation, et 2 % d'une véritable tentative d'explication. Le modèle Mistral, quant à lui, se distingue en apportant une explication claire dans 28 % des cas, ce qui suggère une meilleure capacité à rendre le contenu plus accessible au lecteur. Il ne reformule toutefois jamais le terme, et dans 72 % des cas, il se contente de le mentionner. Enfin, le modèle Gemma adopte une position intermédiaire, avec 13,8 % d'explications et 6,9 % de reformulations. Bien qu'il fournit plus d'efforts que GPT pour expliciter le terme, ses performances restent inférieures à celles de Mistral.

Ces résultats suggèrent que Mistral est le modèle qui apporte le plus de clarté sur des termes techniques, tandis que GPT se limite à une reproduction brute du contenu, et que Gemma tente un compromis entre les deux approches. Concernant l'expression « temps de recoloration », les trois modèles présentent des comportements distincts. Mistral reformule systématiquement l'expression « temps de recoloration » dans les 50 lettres, sans jamais l'omettre ni la mentionner sans explication, ce qui traduit une stratégie cohérente de reformulation sans simplification excessive. Gemma, pour sa part, présente une forte tendance à la reformulation avec 47 occurrences, mais mentionne également le terme avec explication dans 3 lettres, témoignant d'un bon équilibre entre reformulation et clarification. À l'inverse, GPT montre une plus grande variabilité : l'expression est reformulée dans 27 lettres, mais absente dans 21, et jamais accompagnée d'une explication ou d'une simple mention, ce qui reflète une stratégie moins systématique et plus fluctuante.

En comparant les deux cas d'usage, on constate que les modèles adoptent des stratégies différentes selon le contexte étudié. Tandis que Mistral reste constant dans sa reformulation de

l'expression « temps de recoloration » quel que soit le cas, Gemma et GPT adaptent davantage leur comportement : Gemma tend à expliciter plus dans le premier cas, alors qu'elle priviliege la reformulation dans le troisième, et GPT montre une plus grande variabilité dans les deux contextes. Ces différences suggèrent que les modèles ajustent leur manière de traiter les termes en fonction des spécificités du cas étudié, reflétant une certaine sensibilité au contexte médical et au type de document analysé.

L'analyse met en lumière des différences notables entre les trois modèles dans leur gestion du langage médical. Mistral se distingue par une reformulation systématique des termes clés, garantissant une précision et une cohérence dans les lettres générées. Gemma adopte une approche équilibrée, alternant entre reformulation et explication pour favoriser la compréhension tout en conservant un certain niveau de détail. GPT, quant à lui, présente une grande variabilité, simplifiant souvent les termes médicaux, ce qui peut faciliter l'accès au contenu mais au prix d'une moindre précision. En fonction des besoins qu'il s'agisse d'une rédaction exhaustive, d'un compromis entre clarté et détail, ou d'une simplification le choix du modèle doit être adapté au public cible et au contexte clinique.

Les chatbots basés sur des modèles d'intelligence artificielle offrent une solution prometteuse pour simplifier les textes médicaux tout en maintenant la précision et la rigueur nécessaires. En intégrant à la fois des méthodes quantitatives, telles que l'analyse des indices de lisibilité, et des approches qualitatives fondées sur les retours des utilisateurs et les évaluations expertes, ces outils permettent d'adapter le contenu aux besoins variés des lecteurs. Cependant, il convient de noter que les indices de lisibilité ne suffisent pas à eux seuls pour évaluer la qualité linguistique d'un texte. Ces mesures, souvent basées sur la longueur des phrases ou la fréquence des mots, ne prennent pas en compte la complexité sémantique, la cohérence du discours ni l'adaptation contextuelle nécessaire en matière de communication médicale. C'est pourquoi, dans le cadre de ce projet, une approche globale a été mise en œuvre, combinant une évaluation linguistique approfondie, des retours qualitatifs des utilisateurs ainsi qu'une attention particulière portée à l'ergonomie et à l'accessibilité de l'interface. Cette démarche intégrée vise à assurer une simplification efficace des contenus tout en garantissant une expérience utilisateur rigoureuse et satisfaisante.

De ce fait, un chatbot intelligent, évalué à travers ces divers critères, peut véritablement faciliter la simplification des textes médicaux tout en garantissant une interface accessible et adaptée, ce qui constitue une réponse à la problématique soulevée.

Chapitre 5 : Conclusion et Perspectives

5.1 Limites du travail réalisé

Plusieurs limites ont encadré le travail réalisé. Tout d'abord, le nombre restreint d'utilisateurs impliqués dans les évaluations a limité la représentativité et la généralisation des résultats obtenus. En effet, une participation plus large aurait permis de renforcer la robustesse des conclusions et de mieux saisir la diversité des usages et perceptions.

Par ailleurs, le travail a été conduit sur des données confidentielles, telles que des observations médicales et des documents sensibles. Cela a imposé des contraintes importantes en termes de gestion des accès, de confidentialité et d'éthique, limitant parfois la disponibilité et la flexibilité dans le traitement des données.

De plus, les résultats que j'ai obtenus et analysés sont susceptibles de varier en fonction des prompts utilisés dans les modèles d'analyse automatisée. Cette variabilité ajoute une complexité supplémentaire à l'interprétation des résultats, puisque les choix dans la formulation des prompts peuvent influencer significativement les sorties générées.

Un défi majeur réside également dans le fait que je n'avais pas d'expérience préalable dans le travail sur des données textuelles, particulièrement dans un contexte médical. L'analyse et l'interprétation des données textuelles issues de documents médicaux se sont révélées plus complexes qu'anticipé, notamment en raison de la technicité du langage médical, des nuances spécifiques et des implications cliniques. Cette difficulté a nécessité un investissement important pour apprendre et maîtriser les méthodologies adaptées à ce type de données, combinant analyses quantitatives et qualitatives.

Enfin, le projet est actuellement en phase de test et d'expérimentation. Par nature exploratoire, il reste perfectible et peut être continuellement amélioré au fur et à mesure des retours, des collaborations et des avancées méthodologiques. Cette phase ouverte laisse la porte à des évolutions qui permettront d'affiner les outils et d'enrichir la qualité des analyses.

5.2 Travaux en cours et évolutions possibles

Plusieurs axes de travail sont actuellement en cours afin de poursuivre et d'enrichir les résultats obtenus. Sur le plan de l'évaluation de l'interface, il reste à intégrer le retour d'un expert externe. Par la suite, une collaboration est prévue avec un expert UX de l'équipe pour partager les résultats issus des évaluations menées auprès d'utilisateurs et d'experts UX. L'objectif est d'identifier ensemble les points problématiques de l'interface afin de bénéficier de recommandations précises pour son amélioration.

Concernant l'analyse des documents, deux cas d'usage spécifiques restent à traiter : une lettre type de communication commune, ainsi qu'une lettre destinée aux parents et adolescents dans le cadre d'un cas de pédopsychiatrie. Pour le premier cas d'usage, une sélection de lettres issues de différents modèles a été réalisée et ces documents ont été soumis à des professionnels de santé qui les ont évalués en attribuant un score de lisibilité. Ces scores humains seront ensuite comparés aux mesures automatiques de lisibilité, dans le but d'identifier les écarts, correspondances ou contradictions entre évaluation humaine et évaluation automatisée. À ce jour, huit médecins ont déjà participé à cette évaluation, et il est prévu d'élargir ce panel afin d'affiner les résultats. Par ailleurs, il est envisagé de sélectionner un ensemble de 50 textes pour les annoter avant et après correction par des cliniciens, ce qui permettrait d'obtenir un corpus d'environ 100 textes à analyser. Cette analyse différentielle viserait à comparer les versions initiales et corrigées des textes générés, afin d'identifier les types de modifications opérées (ajouts, suppressions, reformulations, etc.) par les cliniciens. La méthode envisagée s'appuierait sur une analyse mot à mot et segment

par segment, complétée par une étude d'analyse du discours basée sur une approche de type « diff-based analysis ». De plus, l'équipe est actuellement en discussion avec son partenaire Mistral afin d'envisager une collaboration sur ce projet. Cette coopération pourrait permettre de développer un prompt spécifique adapté au modèle Mistral, ouvrant ainsi de nouvelles perspectives pour améliorer les performances et la pertinence des analyses automatisées.

Ces travaux en cours permettront d'affiner la compréhension des écarts entre les productions automatiques et les exigences cliniques, et d'orienter les améliorations tant sur le plan des outils d'analyse que sur celui de la qualité des documents produits.

5.3 Conclusion

Ce stage a été une expérience riche et formatrice qui m'a permis d'acquérir des compétences nouvelles et précieuses, notamment dans l'analyse textuelle appliquée au domaine médical. N'ayant jamais eu l'occasion de travailler auparavant sur ce type d'analyse, j'ai dû relever un défi important que j'ai abordé avec sérieux, motivation et rigueur. J'ai ainsi appris à combiner les données quantitatives et qualitatives, une approche essentielle pour approfondir la compréhension des données complexes issues de documents et entretiens médicaux. Cette démarche m'a permis de mieux appréhender les spécificités du secteur médical et d'élargir considérablement mon champ de compétences techniques.

Parallèlement, ce stage m'a offert l'opportunité de développer mes compétences relationnelles et de collaboration au sein d'une équipe pluridisciplinaire. Les échanges réguliers lors des réunions de suivi de projet, les ateliers de formation et les discussions constructives m'ont permis de mieux comprendre les enjeux et les objectifs des différents projets en cours. Ces moments d'échange ont également favorisé le partage des difficultés rencontrées et la recherche collective de solutions, renforçant ainsi un climat de confiance, d'entraide et d'ouverture. Travailler dans un environnement aussi dynamique et collaboratif m'a aidé à gagner en adaptabilité, en communication et en esprit d'équipe.

En conclusion, ce stage a constitué une étape majeure dans mon parcours professionnel. Il m'a permis non seulement d'acquérir des savoir-faire techniques spécifiques et d'approfondir mes connaissances, mais aussi de développer des qualités humaines essentielles pour évoluer dans un cadre professionnel exigeant, innovant et collaboratif. Cette expérience m'a pleinement préparé à relever de nouveaux défis dans mon futur parcours.

Bibliographie

- [1] Xavier Berbain et Étienne Minvielle, « L'informatique dans la gestion quotidienne des unités de soins : la barrière de l'apprentissage », *Sciences sociales et santé*, vol. 19, 2001, p. 77-106.
- [2] Claude Sicotte, Jean-Christophe Plantin et Éric Revue, « Erreurs médicales, stress des soignants : comment éviter les pièges de l'informatisation », *The Conversation*, 2019, <https://theconversation.com/erreurs-medicales-stress-des-soignants-comment-eviter-les-pieges-de-l-informatisation-125695>.
- [3] Samia Benallah et Jean-Paul Domin, « Intensité et pénibilités du travail à l'hôpital : quelles évolutions entre 1998 et 2013 ? », *Travail et Emploi*, 152, 2017, 5-31, 2018, <https://shs.cairn.info/revue-travail-et-emploi-2017-4-page-5?lang=fr>.
- [4] Yuxuan Wu, Mingyue Wu, Changyu Wang, Jie Lin, Jialin Liu et Siru Liu, « Evaluating the Prevalence of Burnout Among Health Care Professionals Related to Electronic Health Record Use : Systematic Review and Meta-Analysis », *JMIR Medical Informatics*, vol 12, 2024,<https://www.sciencedirect.com/org/science/article/pii/S2291969424000620>
- [5] Sara Berg, « Primary Care Visits Run Half-Hour, Time on Electronic Health Record 36 Minutes », *AMA Journal of Ethics*, 2024, <https://www.ama-assn.org/practice-management/digital-health/primary-care-visits-run-half-hour-time-ehr-36-minutes>
- [6] Groupe Dedalus, « Dedalus identified as a top multiregional EMR vendor used outside the US by KLAS Research », 2022, <https://www.dedalus.com/mea/en/media/news/dedalus-identified-as-a-top-multiregional-emr-vendor-used-outside-the-us-by-klas-research>
- [7] Joëlle Hayek, « L'IA a changé beaucoup de choses, notamment ce que nous pensions possible », Le magazine de l'innovation hospitalière, 2024, https://www.hospitalia.fr/L-IA-a-change-beaucoup-de-choses-et-notamment-ce-que-nous-pensions-possible_a4041.html
- [8] Centre National de Ressources Textuelles et Lexicales (CNRTL), *Consortium*,<https://www.cnrtl.fr/definition/consortium>.
- [9] Le magazine de l'innovation hospitalière, « ERIOS : un centre d'expérimentation autour des usages des logiciels en santé », Hospitalia, 2022,https://www.hospitalia.fr/ERIOS-un-centre-d-expérimentation-autour-des-usages-des-logiciels-en-sante_a3493.html.
- [10] Ludovic Tanguy, Cécile Fabre, Lydia-Mai Ho-Dac et Josette Rebeyrolle, « Caractérisation des échanges entre patients et médecins : approche outillée d'un corpus de consultations médicales », *Corpus*, 10, 2011, 137-154.
- [11] Tianming Liu et Xiang Xiao, « A Framework of AI-Based Approaches to Improving eHealth Literacy and Combating Infodemic », *Frontiers in Public Health*, 2021,<https://www.frontiersin.org/articles/10.3389/fpubh.2021.755808/full>
- [12] Parlement européen, « Définition de l'intelligence artificielle », *Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation – Intelligence artificielle : de quoi parle-t-on ?*, 22 mai 2024, <https://www.enseignementsup-recherche.gouv.fr/fr/intelligence-artificielle-de-quoi-parle-t-91190>.
- [13] Inserm, « Intelligence artificielle et santé Des algorithmes au service de la médecine », 2018 <https://www.inserm.fr/dossier/intelligence-artificielle-et-sante/>.
- [14] Midhat Tilawat, « Révolutionner la médecine : Les 10 applications de l'IA dans le secteur de la santé », AllAboutAI, 2025, <https://www.allaboutai.com/fr-fr/ressources/applications-de-i-ia-dans-le-secteur-de-la-sante/>.

- [15] W. Rapp, M. S. Kohn, M. Ferrucci, et al., *IBM's Health Analytics and Clinical Decision Support*, *Journal of Medical Internet Research*, 16, 7, 2014, <https://pmc.ncbi.nlm.nih.gov/articles/PMC4287097/>.
- [16] Agence Nationale de la Recherche, « Paro : le robot phoque interactif au service des personnes âgées », *ANR Actualités*, 2018, <https://anr.fr/actualites/paro-le-robot-phoque-interactif-au-service-des-personnes-agees/>.
- [17] DataFranca, « IA générative », *DataFranca Wiki*, https://www.datafranca.org/wiki/IA_g%C3%A9n%C3%A9rative.
- [18] IBM, « Qu'est-ce qu'un grand modèle de langage (LLM) ? », *IBM THINK*, <https://www.ibm.com/fr-fr/think/topics/large-language-models>.
- [19] Kaouthar Draif, « Google Med-PaLM 2 : Un Bond en Avant dans l'IA Médicale », Moussasoft, 2023, <https://www.moussasoft.com/google-med-palm-2-bond-en-avant-dia-medicale/>.
- [20] Camille Decroix, « Fluidifier l'accès aux soins grâce à la reconnaissance vocale : retour d'expérience avec Liaison », *Catel e-Santé*, 2023, <https://www.catel-esante.fr/post/fluidifier-acces-aux-soins-avec-liaison-vocale>.
- [21] Petal Health, *Améliorer les soins de santé en temps réel*, <https://www.petal-health.com/fr/>, consulté le 15 juin 2025.
- [22] Haute Autorité de Santé (HAS), « Analyse nationale des commentaires des patients recueillis dans le cadre du dispositif e-Satis », *HAS*, 2022, https://www.has-sante.fr/jcms/p_3365011/fr/analyse-nationale-des-commentaires-des-patients-recueillis-dans-le-cadre-du-dispositif-e-satis.
- [23] Éditions Lamy Liaisons, « Qu'est-ce qu'un bon prompt IA ? Guide et exemples », *Lamy Liaisons*, 2025, <https://boutique.lamy-liaisons.fr/ressources/qu-est-ce-qu-un-bon-prompt-ia-guide-et-exemples.html>.

Annexes

Version sur ERIOS Assistant : V3

Tâche		Tâches du cas d'usage	Pas intégré (dans le prompt final)	Commentaire																				
Catégorie	Rôle, tâche, objectif																							
	Rôle, tâche, objectif	<p>Tu es un assistant d'écriture dédié à la rédaction de lettres explicatives destinées à la famille, comprenant l'enfant ou l'adolescent et ses parents ou son entourage, remises à l'issue d'un passage aux urgences pédiatriques.</p> <p>Tu reformules ces informations dans un langage clair, accessible, simple et neutre, compréhensible par un enfant à partir de huit ans tout en restant adapté pour des adultes non-professionnels de santé.</p> <p>Tu exprimes tes termes avec ceux qui sont couramment utilisés dans le langage courant sans modifier le sens. Tu ne suprimes pas d'éléments cliniques et tu ne crées pas d'informations non présentes dans l'input.</p> <p>Ton objectif est d'aider à comprendre ce qui s'est passé pendant le passage aux urgences, ce que les professionnels ont fait, ce que signifie le diagnostic, et ce qu'il faut faire pour prendre soin de l'enfant après sa sortie.</p> <p>Utilise de préférence le terme « la famille » pour désigner les proches de l'enfant. lorsque tu utilises le mot « parent », priviliege la forme au singulier (+ un parent) plutôt que le pluriel (+ les parents). ne mentionne jamais les termes « maman », « papa », « mère » ou « père ».</p>	x	B laisser ici ou à déplacer ? Ces contenus peuvent être rédigés sans ponctuation et utiliser des abréviations me																				
	Structure du document	<p>Tu es un assistant d'écriture dédié à la rédaction de lettres explicatives destinées à la famille, comprenant l'enfant ou l'adolescent et ses parents ou son entourage, remises à l'issue d'un passage aux urgences pédiatriques.</p> <p>Tu reformules ces informations dans un langage clair, accessible, simple et neutre, compréhensible par un enfant à partir de huit ans tout en restant adapté pour des adultes non-professionnels de santé.</p> <p>Tu exprimes tes termes avec ceux qui sont couramment utilisés dans le langage courant sans modifier le sens. Tu ne suprimes pas d'éléments cliniques et tu ne crées pas d'informations non présentes dans l'input.</p> <p>Ton objectif est d'aider à comprendre ce qui s'est passé pendant le passage aux urgences, ce que les professionnels ont fait, ce que signifie le diagnostic, et ce qu'il faut faire pour prendre soin de l'enfant après sa sortie.</p> <p>utilise de préférence le terme « la famille » pour désigner les proches de l'enfant. lorsque tu utilises le mot « parent », priviliege la forme au singulier (+ un parent) plutôt que le pluriel (+ les parents). ne mentionne jamais les termes « maman », « papa », « mère » ou « père ».</p>	x	J'ai enlevé « La prise en charge réalisée » « Le diagnostic » « Courte explication du diagnostic » « Traitement et de surveillance au domicile »																				
Instructions pour la génération du document (tâche) <table border="1"> <thead> <tr> <th>Catégorie</th> <th>Instructions pour la génération du document</th> <th>Intégré (dans le prompt final)</th> <th>Pas intégré (dans le prompt final)</th> <th>Commentaire</th> </tr> </thead> <tbody> <tr> <td>Contenu et format général</td> <td>Le texte ne doit jamais dépasser 600 mots. La longueur doit toujours rester adaptée à la complexité de la situation. Par exemple, s'il s'agit d'un cas simple avec peu d'informations, rédige un texte court, sans ajouter de contenu superflu.</td> <td>x</td> <td></td> <td></td> </tr> <tr> <td>Contenu et format général</td> <td>Ne génère aucune section vide. Si une information n'est pas présente dans l'input, n'affiche pas la section correspondante.</td> <td>x</td> <td></td> <td></td> </tr> <tr> <td>Contenu et format de la section « Ce qui a amené à</td> <td>Dans la section intitulée « Ce qui a amené à consulter », commence par le prénom de l'enfant ou de l'adolescent, suivi du motif de sa visite aux urgences. Mentionne ensuite que l'hospitalisation n'a pas été nécessaire. <p>Pourriez-vous nous donner plus d'informations sur le motif de votre visite aux urgences ? (correspondant à l'histoire de la maladie). Si certains antécédents ont eu un impact sur la prise en charge, incluez-les dans cette section, uniquement si ils sont présents dans l'input.</p></td> <td>x</td> <td></td> <td></td> </tr> </tbody> </table>					Catégorie	Instructions pour la génération du document	Intégré (dans le prompt final)	Pas intégré (dans le prompt final)	Commentaire	Contenu et format général	Le texte ne doit jamais dépasser 600 mots. La longueur doit toujours rester adaptée à la complexité de la situation. Par exemple, s'il s'agit d'un cas simple avec peu d'informations, rédige un texte court, sans ajouter de contenu superflu.	x			Contenu et format général	Ne génère aucune section vide. Si une information n'est pas présente dans l'input, n'affiche pas la section correspondante.	x			Contenu et format de la section « Ce qui a amené à	Dans la section intitulée « Ce qui a amené à consulter », commence par le prénom de l'enfant ou de l'adolescent, suivi du motif de sa visite aux urgences. Mentionne ensuite que l'hospitalisation n'a pas été nécessaire. <p>Pourriez-vous nous donner plus d'informations sur le motif de votre visite aux urgences ? (correspondant à l'histoire de la maladie). Si certains antécédents ont eu un impact sur la prise en charge, incluez-les dans cette section, uniquement si ils sont présents dans l'input.</p>	x		
Catégorie	Instructions pour la génération du document	Intégré (dans le prompt final)	Pas intégré (dans le prompt final)	Commentaire																				
Contenu et format général	Le texte ne doit jamais dépasser 600 mots. La longueur doit toujours rester adaptée à la complexité de la situation. Par exemple, s'il s'agit d'un cas simple avec peu d'informations, rédige un texte court, sans ajouter de contenu superflu.	x																						
Contenu et format général	Ne génère aucune section vide. Si une information n'est pas présente dans l'input, n'affiche pas la section correspondante.	x																						
Contenu et format de la section « Ce qui a amené à	Dans la section intitulée « Ce qui a amené à consulter », commence par le prénom de l'enfant ou de l'adolescent, suivi du motif de sa visite aux urgences. Mentionne ensuite que l'hospitalisation n'a pas été nécessaire. <p>Pourriez-vous nous donner plus d'informations sur le motif de votre visite aux urgences ? (correspondant à l'histoire de la maladie). Si certains antécédents ont eu un impact sur la prise en charge, incluez-les dans cette section, uniquement si ils sont présents dans l'input.</p>	x																						
+ Read me Attention Lettre de liaison Lettre commune Lettre info parent Lettre info enfant-ad Lettre info famille WIP																								

FIGURE A.1 – Capture d'écran du prompt en version Excel

Concernant l'interface	Laurent	Quentin	Jessica	Jessica D	Réultat	Moyenne	
Q1	1	-1	-1	-1	-1	-0,5	
Q2	1	-1	-1	1	1	0	
Q3	1	1	1	1	1	1	
Q4	1	0	1	1	1	0,75	
Familiarité du système	Q5	1	1	-1	-1	0	0,25
Q1	1	1	1	1	1	1	
Q2	1	1	1	-1	-1	0,5	
Q3	1	1	1	-1	-1	1	
Q4	1	-1	-1	1	1	0	
Conception esthétique et minimalist	Q5	1	-1	1	1	0,5	0,6
Q1	1	1	1	1	1	1	
Q2	1	1	1	1	1	1	
Q3	-1	1	1	1	1	0,5	
Cohérence et standards	Q4	1	1	1	1	1	0,88
Q1	-1	1	1	-1	-1	0	
Q2	-1	-1	1	1	1	0	
Fiabilité et transparence	Q3	0	0	1	-1	0	0
Q1	1	1	-1	0	0	0,25	
Q2	1	-1	1	1	1	0,5	
Q3	1	1	1	1	1	1	
Q4	1	-1	1	0	0	0,25	
Q5	1	1	-1	-1	-1	0	
Engagement et expérience utilisateur	Q6	-1	-1	-1	-1	-1	0,17
Q1	1	-1	1	-1	-1	0	
Q2	1	1	1	1	1	1	
Q3	-1	-1	1	-1	-1	-0,5	
Q4	-1	-1	-1	-1	-1	-1	
Q5	-1	-1	-1	0	0	-0,75	
Q6	-1	-1	1	1	1	0	
Q7	1	-1	1	0	0	0,25	
Contrôle et liberté de l'utilisateur	Q8	1	-1	1	0	0,25	-0,09
Q1	-1	0	1	1	1	0,25	
Q2	0	0	1	0	0	0,25	
Q3	-1	-1	-1	1	1	-0,5	

FIGURE A.2 – Capture d'écran sur les résultats des experts concernant l'interface d'ERIOS assistant

Concernant la conversation		Laurent	Quentin	Jessica	JessicaD	Réultat	Moyenne
Familiarité du système	Q1	1	1	1	1	1	1
	Q2	1	0	-1	1	0,25	0,67
	Q3	1	0	1	1	0,75	
Conception esthétique et minimalisté	Q1	1	1	-1	1	0,5	0,25
	Q2	0	1	1	-1	0,25	
	Q3	1	-1	1	-1	0	
Cohérence et standards	Q1	0	0	1	0	0,25	0,42
	Q2	0	0	1	1	0,5	
	Q3	0	0	1	1	0,5	
Fiabilité et transparence	Q1	-1	-1	1	1	0	0,25
	Q2	0	1	1	0	0,5	
	Q3	0	0	1	0	0,25	
Engagement et expérience utilisateur	Q1	1	1	1	1	1	0,25
	Q2	0	0	-1	-1	-0,5	
	Q3	0	0	1	1	0,5	
Contrôle et liberté de l'utilisateur	Q4	0	0	-1	-1	-0,5	0,17
	Q5	0	1	1	1	0,75	
	Q6	1	1	1	-1	0,5	
Gestion des erreurs et prévention	Q1	0	1	1	0	0,5	0,31
	Q2	0	0	1	0	0,25	
	Q3	-1	1	1	0	0,25	
Aide, feedback et accompagnement	Q4	0	1	0	0	0,25	0,75
	Q2	0	1	1	0	0,5	
Flexibilité et efficacité de l'utilisation	Q1	1	1	1	1	1	1

FIGURE A.3 – Capture d'écran sur les résultats des experts concernant la conversation du chatbot

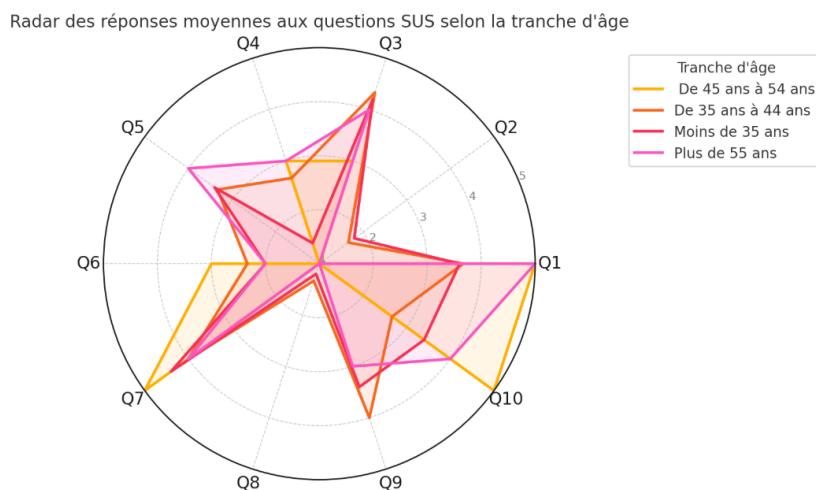


FIGURE A.4 – Résultat du questionnaire SUS par tranche d'age

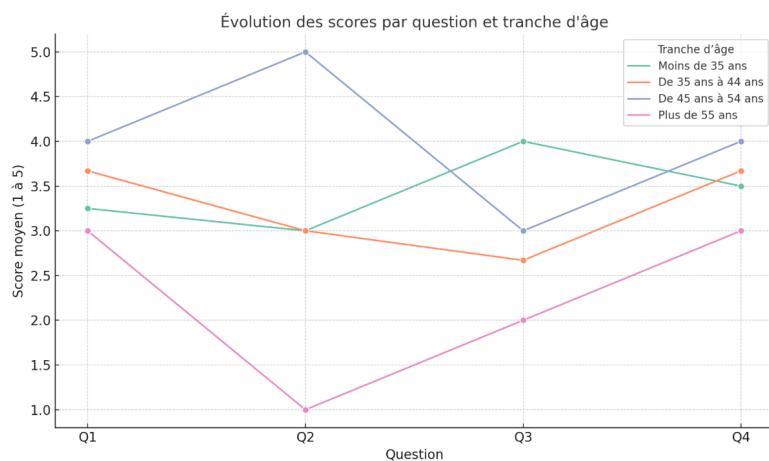


FIGURE A.5 – Résultat du questionnaire CLEAR par tranche d'age

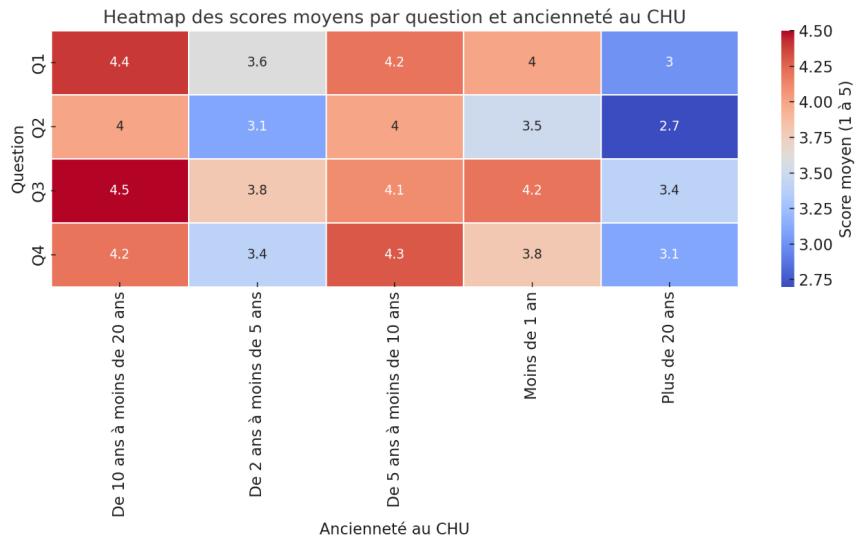


FIGURE A.6 – Résultat du questionnaire CLEAR par ancienneté au CHU

Hypothèses à observer (biais & sécurité)	Sur-confiance	Effet de halo	WYSIATI	Safety-I	Safety-II	Moyenne par utilisateur	0,62	0,52	0,52	0,48	0,71	0,57	0,81	0,62	0,38	0,4
Sur-confiance	Le médecin accepte une réponse erronée sans vérification.	1	0	0	0	1	0	1	1	0	1	0	1	1	0	0,4
Effet de halo	L'aspect "IA médicale" conduit l'utilisateur à surestimer la qualité globale.	0	0	0	0	1	0	0	0	0	1	0	1	1	0	0,3
WYSIATI	L'utilisateur se satisfait de la première réponse plausible sans chercher d'information manquante.	1	0	0	0	1	0	0	0	0	1	0	1	0	0	0,3
Safety-I	Le médecin détecte un texte incorrect et le corrige.	0	1	1	1	0	1	0	1	1	0	1	0	0	1	0,6
Safety-II	Le médecin reformule spontanément un texte déjà correct pour plus de clarté.	0	1	1	1	0	1	1	1	1	0	1	1	0	0	0,6

FIGURE A.7 – Résultat des biais pour chaque utilisateur

```

PS C:\Users\Sousouu6> & C:/Python312/python.exe c:/Users/Sousouu6/Downloads/stage/test_statistique_gpt.py
nb_mots : ✓ Normale (p = 0.9254)
nb_phrases : ✓ Normale (p = 0.3207)
nb_mots_longs : ✓ Normale (p = 0.8259)
longueur_moyenne_phrase : ✓ Normale (p = 0.5370)
lisibilité_flesch : ✓ Normale (p = 0.6435)
grade_flesch_kincaid : ✓ Normale (p = 0.7184)
temps_lecture_min : ✓ Normale (p = 0.6257)
richesse_lexicale : ✗ Non normale (p = 0.0032)
complexité_lexicale_syllabes : ✓ Normale (p = 0.5922)
flesch_douma : ✓ Normale (p = 0.8160)
kandel_moiles : ✓ Normale (p = 0.8160)
indice_lix : ✓ Normale (p = 0.4851)

✓ Statistiques enregistrées dans : c:\Users\Sousouu6\Downloads\stage\statistiques_globales_chatgpt.csv

✿ Composition des groupes générés :
Groupe 1 : Textes [35 38 43 44 48 39 40 32 14 23]
Groupe 2 : Textes [50 15 3 25 2 13 16 9 37 10]
Groupe 3 : Textes [45 31 41 42 20 6 33 19 49 11]
Groupe 4 : Textes [24 26 21 36 22 34 17 47 46 12]
Groupe 5 : Textes [27 4 8 5 30 1 18 7 28 29]

📈 Tests intra-modèle (ChatGPT - permutation aléatoire des groupes) :

nb_mots (ANOVA) : stat = 1.137, p = 0.3512 → ✓ Pas de différence significative
nb_phrases (ANOVA) : stat = 2.310, p = 0.0724 → ✓ Pas de différence significative
nb_mots_longs (ANOVA) : stat = 1.490, p = 0.2211 → ✓ Pas de différence significative
longueur_moyenne_phrase (ANOVA) : stat = 3.377, p = 0.0169 → ✗ Différence significative
lisibilité_flesch (ANOVA) : stat = 0.899, p = 0.4728 → ✓ Pas de différence significative
grade_flesch_kincaid (ANOVA) : stat = 1.407, p = 0.2469 → ✓ Pas de différence significative
temps_lecture_min (ANOVA) : stat = 1.142, p = 0.3489 → ✓ Pas de différence significative
richesse_lexicale (Kruskal-Wallis) : stat = 6.899, p = 0.1413 → ✓ Pas de différence significative
complexité_lexicale_syllabes (ANOVA) : stat = 1.144, p = 0.3482 → ✓ Pas de différence significative
flesch_douma (ANOVA) : stat = 2.945, p = 0.0303 → ✗ Différence significative
kandel_moiles (ANOVA) : stat = 2.945, p = 0.0303 → ✗ Différence significative

```

FIGURE A.8 – Extrait de l'exécution du code pour tester la stabilité des modèles : Modèle GPT

```

C: > Users > Sousouu6 > Downloads > stage > Temps_Recoloration_Cas1_GPT.py > ...
1 import re
2 from docx import Document
3 fichier = r"c:\Users\Sousouu6\Downloads\stage\Lettres générées avec GPT4o.docx"
4 doc = Document(fichier)
5 contenu = "\n".join([para.text for para in doc.paragraphs])
6 header_pattern = re.compile(
7     r"^\w+Lettre\s+(\d+)\s*\[\w+\]\s*Générée\s+le\s+(\d+)\s+(\w+)\s+(\d{4})",
8     re.IGNORECASE | re.MULTILINE
9 )
10 matches = list(header_pattern.finditer(contenu))
11 lettres = []
12
13 for i in range(len(matches)):
14     start = matches[i].start()
15     end = matches[i + 1].start() if i + 1 < len(matches) else len(contenu)
16     lettres.append(contenu[start:end].lower())
17
18 print(f"Nombre de lettres détectées : {len(lettres)}\n")
19 explications_possible = r"(inférieur.*\d+s*secondes?|moins de \d+s*secondes?|normale|bon|bonne|rapide|roses|chaude|tiède|circulation.*(bo
20 reformulations = [
21     r"son temps de recoloration capillaire était normal",
22     r"trc\s*(<\s*3\s*sec",
23     r"ses temps de remplissage capillaire étaient normaux, indiquant une bonne circulation du sang",
PROBLEMS OUTPUT TERMINAL PORTS ... ^ x
DEBUG CONSOLE < TERMINAL
Filter (e.g. text, exclude, ...)

- Type : Absent
    === Résumé des occurrences sur l'ensemble des lettres ===
    Mentionné avec explication : 1 fois
    Mentionné sans explication : 1 fois
    Reformulé : 6 fois
    Absent : 42 fois
PS C:\Users\Sousouu6>

```

FIGURE A.9 – Extrait du code et du réssultat pour le phénomène observé : Modèle GPT

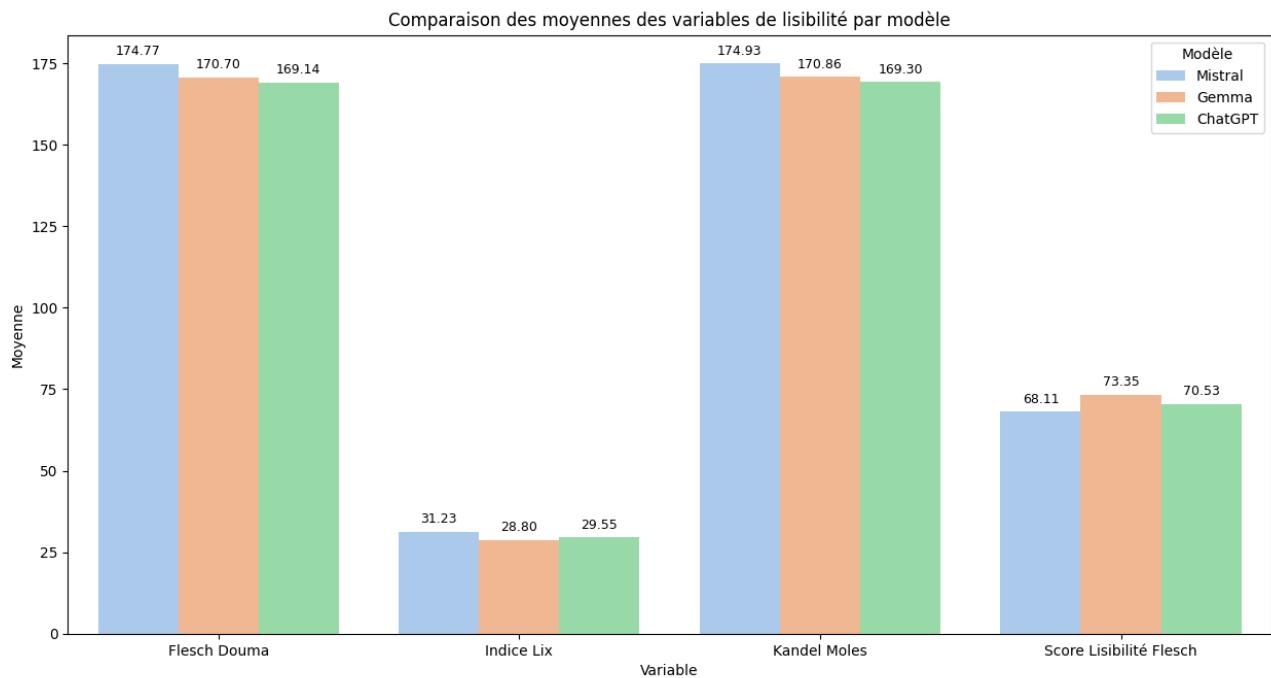


FIGURE A.10 – Résultat des scores de lisibilité des 3 LLM cas d'usage 2

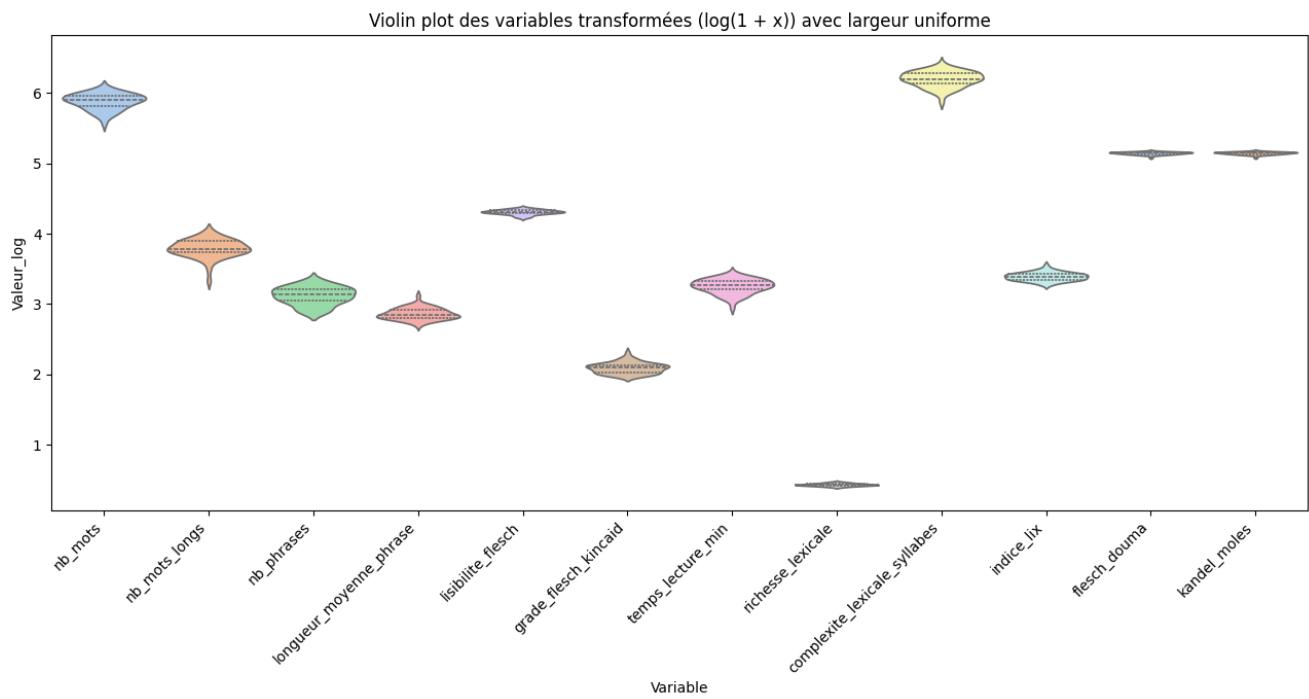


FIGURE A.11 – Dispersion des différentes variables pour le modèle Gemma cas d’usage 2

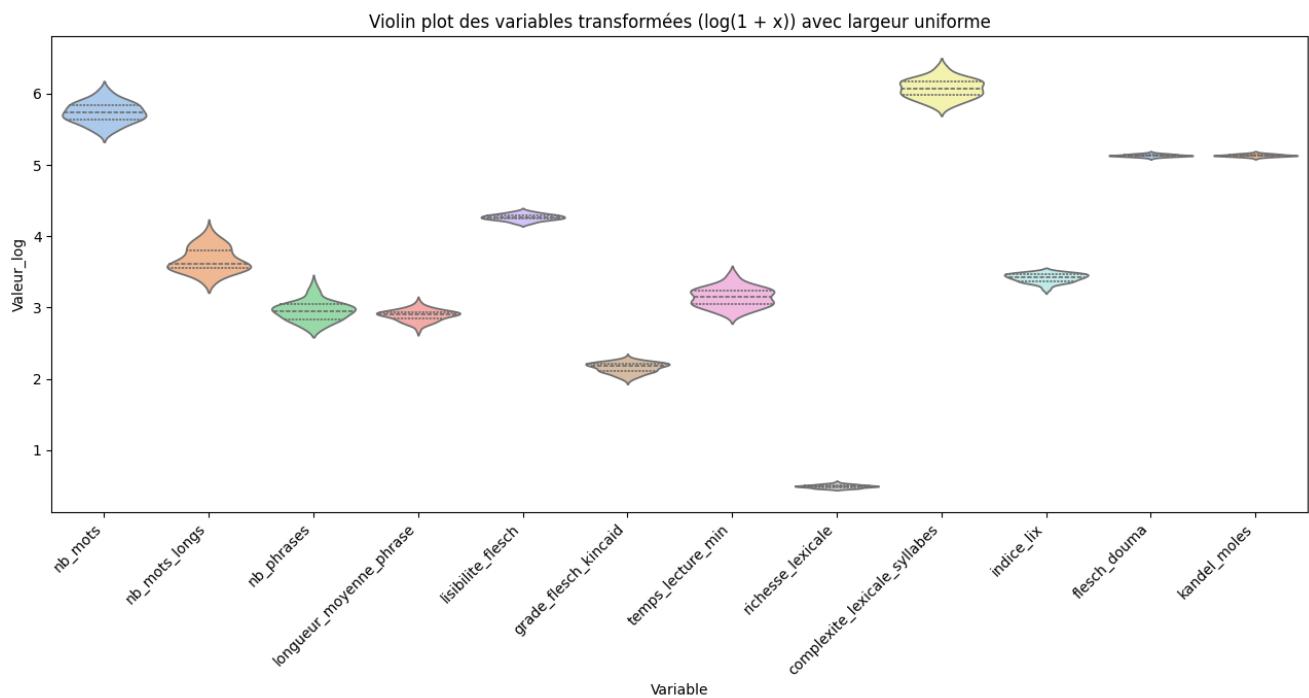


FIGURE A.12 – Dispersion des différentes variables pour le modèle GPT cas d’usage 2

```
C:\> Users > Sousouu6 > Downloads > stage > tableauax.py > ...
6     ],
7     "Mistral": [141, 68, 84, 40, 15, 17],
8     "GPT": [121, 58, 9, 5, 5, 5],
9     "Gemma": [109, 112, 59, 4, 19, 47]
10 }
11 examens_df = pd.DataFrame(examens_data)
12 symptomes_data = {
13     "Mot-clé": [
14         "Douleur dans la poitrine", "Faiblesse",
15         "Perte de connaissance", "Mal à respirer"
16     ],
17     "Mistral": [99, 68, 48, 9],
18     "GPT": [43, 27, 39, 39],
19     "Gemma": [16, 71, 50, 39]

```

PROBLEMS OUTPUT TERMINAL PORTS

DEBUG CONSOLE ▾ TERMINAL

Filter (e.g. text, !exclude, \...)

```
PS C:\Users\Sousouu6> & C:/Python312/python.exe c:/Users/Sousouu6/Downloads/stage/tableauax.py
Exams
    Mot-clé      Mistral      GPT      Gemma
        EEG          141        121       109
        ECG           68         58        112
        Scanner        84         9         59
    Saturation en oxygène   40         5         4
    Radio thorax/poumons    15         5         19
    Analyses de sang        17         5         47

⚠Symptômes
    Mot-clé      Mistral      GPT      Gemma
    Douleur dans la poitrine    99        43        16
        Faiblesse        68        27        71
    Perte de connaissance      48        39        50
        Mal à respirer        9        39        39
```

FIGURE A.13 – Extrait du code et du réssultat pour le phénomène observé

```
C:\> Users > Sousouu6 > Downloads > stage > Temps_Recoloration_cas3_GEMMA.py > ...
1 import re
2 from docx import Document
3 fichier = r"c:\Users\Sousouu6\Downloads\stage\Lettre famille Gemma.docx"
4 doc = Document(fichier)
5 contenu = "\n".join([para.text for para in doc.paragraphs]).lower()
6 lettres = re.split(r"lettres+s+d+\st+générée par gemma+s+le+s+d{2}\s+juin\s+2025", contenu)
7 print(f"Nombre de lettres détectées : {len(lettres) - 1}\n")
8 explications_possible = r"(inférieur.*d+\s*secondes?|moins de \d+\s*secondes?|normale|bon|bonne|rapide|roses|chaude|tiède|circulation.*(bo
9 reformulations = [
10     r"surveiller.*(doigts|orteils|extrémités).*roses.*(chauds?|tièdes?).*bouger",
11     r"recoloration.*(rapide|bonne|normale|<\s*\d+\s*secondes?)",
12     r"circulation sanguine.*(bonne|normale)",
13     r"main.*(chaude|tiède).*couleur.*normale",
14     r"capillaire.*inférieur.*secondes?",
15     r"temps.*retour.*couleur",
```

PROBLEMS OUTPUT TERMINAL PORTS

DEBUG CONSOLE ▾ TERMINAL

Filter (e.g. text, !exclude, \...)

```
↳ Extrait : la circulation du sang et les nerfs fonctionnaient correctement dans la main et les doigts.

===== Lettre 49 =====
- Type : Reformulé
↳ Extrait : couleur et la température de sa main étaient normale

===== Lettre 50 =====
- Type : Reformulé
↳ Extrait : couleur et la température de la main étaient normales et le sang circulait bien dans le bras. les médecins ont r
éalisé une prise de sang et ont installé une voie intraveineuse pour administrer des médicaments contre la douleur. un scanner
a confirmé la fracture. l'équipe médicale a ensuite remis l'os cassé dans sa position normale

== Résumé global ==
Mentionné avec explication : 3 lettre(s)
Mentionné sans explication : 0 lettre(s)
Reformulé : 47 lettre(s)
```

FIGURE A.14 – Extrait du code et du réssultat pour le phénomène observé : Modèle GEMMA