L'intelligence artificielle en santé au travail : évaluation de son potentiel pour renforcer la précision et la clarté des préconisations médicales*

AUTEUR :
C. Broutin, Lille

EN RÉSUMÉ

Dans une étude menée dans les Hauts-de-France, 78 % des préconisations rédigées par les médecins du travail présentaient au moins un défaut de qualité susceptible d'en altérer la compréhension et, *in fine*, de nuire au maintien en emploi. Face à ce constat, une expérimentation a été conduite afin d'évaluer la capacité du modèle de raisonnement o1 d'OpenAI, l'entreprise à l'origine de ChatGPT, à détecter ces défauts, selon cinq critères, en comparaison avec un consensus multidisciplinaire. Après une optimisation itérative du prompt, l'analyse de 385 préconisations tirées au sort dans une base de données a montré 74,6 % de concordances, 10,5 % de discordances justifiées, 13,2 % de discordances par excès et 1,5 % de discordances par défaut, sans aucune hallucination. Ces résultats suggèrent que l'utilisation d'o1 pourrait constituer un outil prometteur d'aide à la relecture.

MOTS CLÉS
Intelligence artificielle – IA / Surveillance médicale / Suivi médical

La qualité de la formulation et la précision des préconisations des médecins du travail favorisent le maintien en emploi des salariés, évitent les risques d'interprétation erronée ainsi que les conséquences juridiques associées. Cet article rend compte d'une étude sur la capacité d'une intelligence artificielle à évaluer les préconisations des médecins du travail selon plusieurs critères. Cette étude fait suite à la thèse du Dr N'Guessan (« *Analyse critique de la rédaction actuelle des préconisations chez les médecins du travail en Hauts-de-France* ») [1] qui avait pour objectif d'examiner et de catégoriser les préconisations formulées par les médecins du travail à l'intention des employeurs, en s'appuyant sur cinq critères de mauvaise qualité :

- imprécisions, difficultés de compréhension et d'application pour l'employeur (critère 1) :

** Cet article résulte d'un mémoire soutenu en octobre 2025 (L'intelligence artificielle en santé au travail : un levier pour accroître la précision et la clarté des préconisations médicales ? Mémoire pour le diplôme d'état de spécialité en médecine du travail. Lille : Université de Lille, Faculté de Médecine Henri Warembourg ; 2025. Lien vers le mémoire complet : https://www.iasantetravail.com/ia-préconisations). Les modèles d'intelligence artificielle présentés correspondent à l'état des connaissances au moment du recueil et de l'analyse des données ainsi que de la rédaction, réalisés entre la fin de l'année 2024 et le début de l'année 2025.*

L'intelligence artificielle en santé au travail : évaluation de son potentiel pour renforcer la précision et la clarté des préconisations médicales

préconisations comportant des imprécisions sur la temporalité, avec une fin imprécise dans le temps ou avec une mention *« renouvelable »* sans périodicité indiquée ;

- induction d'un doute sur la force d'obligation de l'avis médical (critère 2) : préconisations utilisant le conditionnel ou des termes introduisant un choix, une marge de manœuvre ou un doute quant à leur mise en place par l'employeur ;
- informations ne relevant pas de l'annexe 4[1] et posant des difficultés juridiques (critère 3) : avis comportant des informations relevant d'échanges avec l'employeur ou avec le salarié ou des commentaires sans formulation de préconisations à proprement parler ;
- changements de postes ou inaptitudes médicales déguisées (critère 4) : avis préconisant un autre poste pour le salarié ou comportant des aménagements manifestement trop importants pour permettre le maintien à son poste antérieur ;
- rupture du secret médical ou atteinte à la vie privée du salarié (critère 5) : préconisations mentionnant l'invalidité, le bénéfice de la reconnaissance de la qualité de travailleur handicapé (RQTH), une sollicitation de Cap emploi, une maladie ou toute information sur la santé du salarié.

Douze services de prévention et de santé au travail interentreprises (SPSTI), constituant avec l'Institut de santé au travail du Nord de la France (ISTNF) le groupe de travail *« Prévention de la désinsertion professionnelle »*, ont fourni les préconisations rédigées durant le mois de septembre 2023. Celles-ci étaient d'abord cotées en double aveugle selon les critères ci-dessus, par deux évaluateurs (une stagiaire en master 2 de droit social et un interne en médecine du travail). Ensuite, les analyses respectives étaient confrontées avec l'expertise de trois professionnels de la médecine du travail afin d'identifier et comprendre les écarts d'interprétation entre les deux évaluateurs initiaux, permettant ainsi une harmonisation de la lecture et de l'application des critères. Enfin, une nouvelle phase de cotation en double aveugle, fondée sur cette harmonisation, permettait d'atteindre le consensus multidisciplinaire (CM).

Cette étude avait permis d'analyser 4 217 préconisations formulées en septembre 2023 dans les Hauts-de-France. Parmi elles, 3 776 ont fait l'objet d'un CM : seules 22 % ne répondaient à aucun critère de mauvaise qualité. L'analyse de la répartition des critères, portant sur les 2 947 préconisations restantes (78 %), a montré que :

- 53 % présentaient des imprécisions ou des difficultés de compréhension et d'application pour l'employeur (critère 1) ;
- 28 % comportaient des informations ne relevant pas de l'annexe 4 (critère 3) ;
- 9,7 % laissaient un doute quant à leur mise en œuvre (critère 2) ;
- 6,2 % impliquaient une rupture du secret médical (critère 5) ;
- 1,7 % faisaient état d'un changement de poste ou d'aménagements jugés excessifs (critère 4).

Malgré ces taux élevés de préconisations de mauvaise qualité, les avis des médecins du travail sont en réalité rarement contestés juridiquement. Une étude visant à recenser le nombre de contestations des avis depuis la réforme de 2016 a, en effet, identifié 207 arrêts rendus par les cours d'appel, dont seuls 22 % concernaient des mesures d'aménagement. Ce chiffre, quasiment négligeable par rapport au nombre total d'avis (636 917 préconisations formulées en 2022), peut s'expliquer notamment par le délai de contestation limité à 15 jours par l'employeur ou le salarié ainsi que par la durée moyenne des litiges, estimée à 13 mois [2]. Au-delà des éventuelles répercussions juridiques, une rédaction imprécise peut fragiliser la communication et entamer le lien de confiance avec l'employeur. Or cette confiance est essentielle pour favoriser une collaboration efficace, améliorer les conditions de travail, prévenir les risques professionnels et assurer le maintien en santé et en emploi des salariés. Des recommandations mal formulées ou percues comme inapplicables risquent d'entraîner une incompréhension, un défaut d'application ou encore une impression de manque de rigueur. Cela peut non seulement nuire à la crédibilité du médecin du travail mais aussi inciter l'employeur à ignorer ou minimiser les préconisations, voire à limiter sa collaboration avec le SPST.

Améliorer la qualité rédactionnelle des préconisations présente donc plusieurs bénéfices majeurs : assurer une bonne compréhension et une meilleure application des mesures de prévention ; faciliter les relations entre l'employeur et le salarié ainsi qu'entre l'employeur et le médecin du travail ; éviter une perte de temps dans les échanges entre les différentes parties ; réduire le risque de contentieux prud'homal ou ordinal ; préserver et valoriser l'image du médecin du travail. La rédaction des préconisations doit ainsi reposer sur la clarté, la précision et la pertinence, afin de renforcer la transparence, la confiance et la portée de la démarche de prévention [1]. Le travail présenté ici étudie la capacité d'une intelligence artificielle à

1. Annexe 4 : document établi par le médecin du travail proposant des mesures d'aménagement, d'adaptation ou de transformation du poste, ou du temps de travail, en application de l'article L. 4624-3 du Code du travail.

évaluer, selon les critères énoncés ci-dessus, les préconisations des médecins du travail.

LES GRANDS MODÈLES DE LANGAGE

Les grands modèles de langage (*large language models*, LLMs) sont des systèmes d'intelligence artificielle (IA) conçus pour comprendre, interpréter et générer du texte en langage naturel. Ils reposent sur des réseaux de neurones, une architecture inspirée du cerveau humain, composée de couches interconnectées de *« neurones »* artificiels qui traitent et transmettent l'information. Contrairement aux programmes traditionnels qui suivent des instructions explicites, les LLMs apprennent à partir d'énormes quantités de données textuelles, identifiant des motifs, des structures grammaticales et des contextes sémantiques pour prédire et générer des séquences de mots cohérentes [3].

Un aspect clé des LLMs est leur caractère non déterministe, ce qui signifie que, pour une même entrée, le modèle peut produire des réponses différentes à chaque exécution [4, 5]. Cette variabilité est due aux mécanismes de probabilité intégrés dans leur fonctionnement, où chaque mot généré est sélectionné en fonction de sa probabilité conditionnelle, permettant ainsi une diversité de réponses et une créativité accrue dans la génération de contenu [5].

Les LLMs peuvent trouver des applications variées dans de nombreux domaines, même si leurs performances restent variables selon les contextes d'usage, les données disponibles et les attentes des utilisateurs. Dans l'assistance virtuelle, ils peuvent fournir des réponses automatisées et interactives aux utilisateurs, susceptibles de contribuer à l'amélioration de l'expérience client [6]. En traduction automatique, ils peuvent convertir du texte d'une langue à une autre avec un niveau de précision souvent élevé, mais inégal selon les langues, les domaines et la complexité des contenus, ce qui peut néanmoins faciliter la communication multilingue. Pour la création de contenu, les LLMs peuvent générer des articles, des rapports et des scripts, offrant ainsi un appui aux créateurs dans la production de matériel original, sans garantir pour autant une adéquation parfaite aux attentes initiales. Dans l'analyse de données, ils peuvent aider à interpréter et résumer de vastes ensembles de données textuelles, bien que la pertinence des informations extraites dépende fortement de la qualité des entrées et du cadrage de la tâche. Enfin, dans le domaine de l'éducation, les LLMs peuvent assister l'apprentissage et l'enseignement en fournissant des explications et des réponses personnalisées aux étudiants [7], avec toutefois des résultats qui peuvent différer selon les besoins pédagogiques et le niveau de fiabilité attendu.

Cependant, malgré leurs capacités avancées, les LLMs présentent certaines limites. Ils peuvent générer des réponses factuellement incorrectes ou biaisées, voire des hallucinations (encadré 1), reflétant les données sur lesquelles ils ont été entraînés [8]. De plus, leur compréhension du contexte est limitée à l'information présente dans les données d'entraînement, ce qui peut engendrer des interprétations erronées. Il est donc essentiel d'utiliser ces modèles avec discernement et de les compléter par une supervision humaine, en particulier dans des domaines sensibles comme la médecine ou le droit.

L'utilisation d'un LLM nécessite la formulation d'un prompt. Il s'agit d'une instruction ou une question formulée par l'utilisateur pour orienter le modèle dans la génération de sa réponse. La qualité et la précision du prompt influencent directement la pertinence des réponses produites. Parmi les bonnes pratiques de conception de prompts figurent la définition claire du rôle du modèle, le détail des attentes, l'utilisation

↓ **Encadré 1**

> LES HALLUCINATIONS

Les hallucinations dans les LLMs désignent des situations où le modèle affirme des informations qui, bien que syntaxiquement et sémantiquement correctes, sont factuellement fausses [8]. Sur les *benchmarks* SimpleQA et PersonQA, conçus pour évaluer la fiabilité factuelle des modèles de langage, SimpleQA mesure la précision sur des connaissances générales simples et la capacité du modèle à éviter les hallucinations, tandis que PersonQA évalue l'exactitude biographique, la désambiguïsation entre individus portant des noms similaires et la robustesse face aux erreurs d'attribution ou aux informations inventées. Sur ces deux tests, o1 utilisé pour cette étude hallucine dans respectivement 44 % et 16 % des cas, ce qui est inacceptable dans un contexte médical [9].

d'exemples illustratifs (« *few-shot prompting* »), l'optimisation de la longueur du prompt et l'intégration de la chaîne de raisonnement [10 à 13]. La performance de raisonnement des LLMs tend à diminuer à mesure que la longueur du prompt augmente [13].

Les LLMs ont également démontré une efficacité notable dans le domaine médical. Des études indiquent que certains systèmes égalent, voire surpassent, les médecins humains dans des compétences essentielles telles que le raisonnement clinique et les connaissances médicales. Par exemple, une étude a révélé que GPT-4 d'OpenAI, sans ajustement spécifique, dépasse le score de passage de l'examen de licence médicale des États-Unis (*United States medical licensing examination*, USMLE) de plus de 20 points, surpassant à la fois les modèles généralistes antérieurs et ceux spécifiquement entraînés sur des connaissances médicales [14]. Il convient toutefois de préciser que ces résultats portent sur des tests standardisés et ne signifient pas que les LLMs surpassent les médecins dans la pratique clinique globale.

De plus, GPT-4 a démontré une capacité à générer des réponses perçues comme empathiques lors d'interactions avec des patients. Une revue systématique l'a évalué sur des questions liées aux compétences relationnelles dans l'USMLE, où GPT-4 a répondu correctement à 90 % des questions, suggérant une aptitude à simuler certaines facettes de l'empathie humaine [15]. Une autre étude a évalué l'impact de l'assistance de GPT-4 sur le raisonnement des médecins en matière de gestion clinique. Les résultats ont montré que les médecins ainsi assistés obtenaient des scores significativement plus élevés que ceux utilisant uniquement les ressources conventionnelles. Aucune différence significative n'a été observée entre les performances des médecins assistés par GPT-4 et celles de GPT-4 seul, ce qui suggère que l'assistance par un LLM peut améliorer le raisonnement des médecins dans des tâches de gestion clinique complexes [16].

LE MODÈLE O1 : UNE AVANCÉE EN RAISONNEMENT COMPLEXE

Le modèle o1 d'OpenAI, l'entreprise américaine à l'origine de ChatGPT et des différents modèles de la série GPT, se distingue par ses performances avancées en matière de raisonnement complexe. Comme ses prédécesseurs, o1 est basé sur l'architecture *Transformer*, une structure innovante qui utilise des mécanismes d'attention pour gérer efficacement les dépendances, c'est-à-dire les relations entre des mots ou concepts qui peuvent être éloignés dans une phrase ou dans le texte mais dont le lien est essentiel à la compréhension du sens global. Cette architecture permet au modèle de traiter et de comprendre des contextes étendus, améliorant ainsi la qualité et la pertinence des réponses générées [9]. Introduite par Vaswani et al. en 2017, l'architecture *Transformer* a révolutionné le traitement du langage naturel en se passant des réseaux récurrents, s'appuyant uniquement sur des mécanismes d'attention pour traiter les séquences de données [5].

L'un des aspects clés du modèle o1 est son apprentissage par renforcement à grande échelle [17]. Cette méthode d'entraînement permet au modèle d'apprendre à prendre des décisions en recevant des récompenses pour les actions correctes et des pénalités pour les erreurs, optimisant ainsi ses performances en affinant ses capacités de prise de décision et en améliorant sa chaîne de raisonnement. La chaîne de raisonnement (*chain of thought*, COT) fait référence au processus interne par lequel le modèle décompose et analyse les informations pour générer une réponse cohérente et pertinente [18]. En améliorant cette chaîne de raisonnement, o1 est capable de structurer ses réponses de manière logique et de résoudre des problèmes complexes plus efficacement. L'approche de chaîne de raisonnement renforce également son alignement éthique en intégrant des politiques de sécurité directement dans le processus de raisonnement, minimisant ainsi les risques de comportements indésirables [9]. Les performances du modèle o1 dans des *benchmarks* rigoureux et son architecture avancée indiquent qu'il possède les capacités nécessaires pour égaler, voire dépasser, les systèmes actuels tels que GPT-4 dans le domaine médical.

En tirant parti de ces capacités avancées, l'objectif de ce travail est d'évaluer dans quelle mesure le modèle o1 peut identifier, dans les préconisations médicales formulées par les médecins du travail, les erreurs potentielles répondant aux cinq critères de mauvaise qualité précédemment cités (et le critère 0 d'absence d'erreur), avant leur transmission à l'employeur.

L'exploitation du potentiel de reformulation ou de génération intégrale d'une préconisation n'a pas été retenue et testée, le rôle de l'IA

étant ici considéré comme un outil d'accompagnement et d'assistance aux professionnels dans leurs tâches, sans substitution à leur expertise.

MATÉRIELS ET MÉTHODES

POPULATION D'ÉTUDE ET RÉFÉRENCE

Afin d'évaluer les capacités d'analyse du modèle o1, ses résultats ont été confrontés à ceux de préconisations ayant déjà fait l'objet d'une évaluation dans le cadre de la thèse du Dr N'Guessan [1]. L'analyse comparative a porté spécifiquement sur les préconisations ayant donné lieu à un consensus multidisciplinaire (CM) dans ce travail antérieur.

CATÉGORISATION DES DISCORDANCES

Chaque préconisation a été étudiée en fonction des cinq critères de mauvaise qualité précédemment cités. En cas de discordance entre le modèle o1 et le CM, celle-ci a été classée dans l'une des catégories suivantes par l'auteur avec l'aide d'un professeur en santé au travail :

- pas de discordance : les réponses du modèle o1 et du CM sont concordantes ;
- discordance justifiée : o1 donne un avis différent du CM, mais qui a été validé comme correct après réévaluation. Il peut s'agir, par exemple, d'une rupture du secret médical non détectée initialement par le CM et identifiée par le modèle, relevant ainsi d'une erreur du jugement initial ;
- discordance par excès : o1 détecte une erreur en raison d'un jugement trop sévère, par exemple lorsqu'il considère l'expression *« RV médical »* comme relevant du critère 1 au motif que l'abréviation *« RV »* serait jugée imprécise ;
- discordance par défaut : o1 ne repère pas une erreur qui aurait dû être détectée et qui a été identifiée par le CM ;
- hallucination : o1 invente une réponse de toute pièce, sans rapport avec la question posée.

OPTIMISATION DU PROMPT

Afin d'atténuer ces hallucinations et d'obtenir des réponses plus précises, des prompts clairs et bien structurés ont été fournis au modèle. De plus, la performance de raisonnement des LLMs tendant à diminuer à mesure que la longueur du prompt augmente, des instructions concises mais exhaustives ont été privilégiées.

Le prompt a été testé et affiné itérativement sur une centaine de préconisations tirées au hasard et équitablement réparties dans la base de données, afin d'éviter le surentraînement. En effet, la base de données n'était pas constituée de manière aléatoire, ce qui entraînait la présence d'un style d'écriture homogène et d'erreurs récurrentes provenant d'un même auteur sur plusieurs préconisations successives. Autrement dit, il y avait un risque de surajuster le prompt aux biais de formulation de quelques médecins seulement. Sans une sélection véritablement aléatoire couvrant l'ensemble de la base, il devenait impossible de savoir si les performances obtenues auraient été généralisables à d'autres utilisateurs.

Cinq itérations successives ont été nécessaires avant d'obtenir un prompt permettant des taux de discordance par défaut et d'hallucinations proches de 0 %. Les résultats ont été confrontés aux évaluations du CM à chaque itération. Le prompt final attribue au modèle le rôle d'un médecin du travail, lui demande d'analyser chaque préconisation de manière exhaustive selon les cinq critères prédéfinis, en procédant étape par étape (intégration de la chaîne de raisonnement) et en se concentrant uniquement sur les informations fournies, sans ajout d'informations extérieures. Ce prompt validé a ensuite été retesté sur 50 nouvelles préconisations tirées au hasard de la base de données afin de confirmer son efficacité.

CALCUL DE L'EFFECTIF ET ANALYSE STATISTIQUE

Le test de McNemar a été utilisé afin de comparer les proportions d'erreurs identifiées par le CM et celles identifiées par le modèle o1. Les intervalles de confiance des proportions ont été calculés selon la méthode de Wilson. Afin d'obtenir une précision d'environ ± 5 % (IC 95 %), l'effectif minimal garantissant une puissance suffisante pour détecter d'éventuelles différences significatives a été calculé selon la formule : $n = z^2 \times p\,(1 - p) / e^2$, avec $z = 1{,}96$ (intervalle de confiance à 95 %), $p = 0{,}5$ (variance maximale) et $e = 0{,}05$ (marge d'erreur), soit $n = 385$ préconisations.

Le prompt validé a ensuite été appliqué à ces 385 nouvelles préconisations, extraites aléatoirement de la base de données. Les résultats obtenus ont été soumis à une analyse comparative avec ceux issus de la thèse du Dr N'Guessan. Le niveau d'accord entre les deux approches a été évalué à l'aide du coefficient kappa de Cohen, qui mesure la concordance entre deux observateurs au-delà du simple effet du hasard. Son interprétation

a suivi les seuils proposés par Landis et Koch : $\kappa < 0,20$ (accord faible), $\kappa = 0,21–0,40$ (accord faible à modéré), $\kappa = 0,41–0,60$ (accord modéré), $\kappa = 0,61–0,80$ (accord substantiel) et $\kappa > 0,80$ (accord presque parfait). Une analyse descriptive par critère a également été réalisée.

Aucune des préconisations entrées dans le modèle ne comportait d'informations permettant d'identifier le salarié, l'entreprise, l'employeur, le médecin du travail ou le SPSTI émetteur de la préconisation.

RÉSULTATS

ANALYSE COMPARATIVE GLOBALE ENTRE LE CM ET O1

Parmi les 385 préconisations étudiées, des écarts notables sont observés entre le CM et le modèle o1. Globalement, le modèle détecte plus fréquemment les critères d'imprécision (critère 1), de doute sur l'obligation (critère 2), d'informations hors annexe 4 (critère 3) et de changement de poste (critère 4) que le CM. Les différences sont statistiquement significatives et traduisent une tendance du modèle à multiplier la détection d'erreurs. À l'inverse, l'absence d'erreurs (critère 0) est plus fréquemment observée dans les évaluations du CM que dans celles du modèle, confirmant la propension de ce dernier à surestimer les anomalies. Le critère 5, correspondant à la rupture du secret médical, se distingue par des résultats plus proches entre le CM et o1, bien qu'il reste aussi plus élevé avec le modèle (tableau I).

L'analyse de concordance confirme ces constats : l'accord mesuré par le kappa de Cohen demeure faible pour les critères 0 à 3, atteint un niveau modéré pour le critère 4 et devient presque parfait pour le critère 5 ***(tableau I)***.

ANALYSE DESCRIPTIVE PAR CRITÈRE

Les taux de concordance entre le CM et le modèle o1 varient sensiblement selon les critères. Pour le critère 1 (imprécision), la concordance est de 61,3 %, mais accompagnée d'un taux non négligeable de discordances par excès, traduisant une tendance du modèle à surévaluer les erreurs. Le critère 2 (doute sur obligation) présente un taux de concordance de 45,5 %, avec des discordances par excès particulièrement marquées, confirmant cette même tendance. À l'inverse, les critères 3 à 5 se distinguent par des taux de concordance plus élevés : 77,1 % pour le critère 3, plus de 90 % pour le critère 4, et 96 % pour le critère 5. Les discordances y restent marginales, sans aucune discordance par défaut pour les critères 4 et 5. Il est à noter également l'absence complète d'hallucinations détectées (tableau II).

↓ Tableau I

> COMPARAISON DU TAUX DE DÉTECTION DES DIFFÉRENTS TYPES D'ERREURS ENTRE LE CONSENSUS MULTIDISCIPLINAIRE (CM) ET LE MODÈLE O1 : POURCENTAGES, INTERVALLES DE CONFIANCE (MÉTHODE DE WILSON), SIGNIFICATIVITÉ (TEST DE MCNEMAR) ET NIVEAUX D'ACCORD (KAPPA DE COHEN, SEUILS DE LANDIS ET KOCH).

	CM		o1		Significativité	Accord CM / o1	
Critère	n (%)	IC 95%	n (%)	IC95%	valeur-p	Kappa de Cohen	Interprétation (Landis & Koch)
0 (Absence d'erreur)	83 (21,56 %)	(17,74 - 25,94)	18 (4,68 %)	(2,98 - 7,27)	$5,42\times10^{-20}$	0,13	Accord faible (0,00 – 0,20)
1 (Imprécision)	236 (61,30 %)	(56,34 - 66,03)	355 (92,21 %)	(89,09 - 94,49)	$3,01\times10^{-36}$	0,10	Accord faible (0,00 – 0,20)
2 (Doute sur obligation)	68 (17,66 %)	(14,18 - 21,79)	273 (70,91 %)	(66,18 - 75,22)	$3,89\times10^{-62}$	0,14	Accord faible (0,00 – 0,20)
3 (Infos hors annexe 4)	32 (8,31 %)	(5,95 - 11,50)	95 (24,68 %)	(20,63 - 29,22)	$2,17\times10^{-19}$	0,13	Accord faible (0,00 – 0,20)
4 (Changement de poste)	12 (3,12 %)	(1,79 - 5,37)	35 (9,09 %)	(6,61 - 12,38)	$2,38\times10^{-07}$	0,48	Accord modéré (0,41 – 0,60)
5 (Rupture secret médical)	50 (12,99 %)	(9,99 - 16,71)	63 (16,36 %)	(13,00 - 20,39)	$2,44\times10^{-04}$	0,82	Accord presque parfait (0,81 – 1,00)

↓ Tableau II

> POURCENTAGE DE DISTRIBUTION DES CONCORDANCES ET DIVERGENCES ENTRE LE CONSENSUS MULTIDISCIPLINAIRE ET O1 PAR CRITÈRE.

Critère	Concordance n (%)	Divergence par excès n (%)	Divergence justifiée n (%)	Divergence par défaut n (%)	Hallucination n (%)
1 (Imprécision)	236 (61,3%)	75 (19,5%)	65 (16,9%)	9 (2,3%)	0 (0,0%)
2 (Doute sur obligation)	175 (45,5%)	117 (30,4%)	90 (23,4%)	3 (0,8%)	0 (0,0%)
3 (Infos hors annexe 4)	297 (77,1%)	38 (9,9%)	34 (8,8%)	16 (4,2%)	0 (0,0%)
4 (Changement de poste)	358 (93,0%)	17 (4,4%)	10 (2,6%)	0 (0,0%)	0 (0,0%)
5 (Rupture secret médical)	370 (96,1%)	11 (2,9%)	4 (1,0%)	0 (0,0%)	0 (0,0%)
Total	**74,60%**	**13,20%**	**10,54%**	**1,45%**	**0%**

DISCUSSION

L'analyse de la classification du CM et du modèle o1 sur 385 préconisations a permis d'évaluer la capacité du modèle à détecter différents types d'erreurs dans la rédaction des préconisations médicales en santé au travail. Les résultats mettent en évidence des divergences statistiquement significatives entre l'évaluation humaine et celle effectuée par le modèle o1. Concernant le critère 0, qui correspond à l'absence totale d'erreurs dans les préconisations, le CM a identifié 21,56 % des préconisations comme étant correctement rédigées, contre seulement 4,68 % pour o1. Cette différence s'accompagne d'un accord faible entre les deux évaluateurs (o1 et CM). Cela suggère que o1 a tendance à considérer comme imparfaites des préconisations que l'humain juge acceptables, probablement en raison d'une interprétation plus stricte des critères de mauvaise qualité.

Pour le critère 1, qui porte sur les formulations imprécises ou difficilement applicables pour l'employeur, o1 détecte des erreurs dans 92,21 % des cas, contre 61,30 % pour le CM. Le faible accord témoigne d'une possible sensibilité du modèle à l'imprécision syntaxique.

Une tendance comparable se retrouve pour les critères 2 et 3. Pour le critère 2, o1 l'identifie dans 70,91 % des cas contre 17,66 % pour le CM, avec un accord faible. Pour le critère 3, les taux sont respectivement de 24,68 % et 8,31 %, avec un accord faible également.

Concernant le critère 4, o1 le détecte dans 9,09 % des cas contre 3,12 % pour le CM, cette fois avec un accord modéré. Ces résultats suggèrent qu'o1 applique des filtres plus stricts dans la détection de ces critères.

Pour le critère 5, qui touche à la rupture du secret médical ou à la divulgation d'informations confidentielles, les résultats sont plus convergents : 16,36 % pour o1 contre 12,99 % pour le CM. Le kappa atteint ici 0,82, traduisant un accord presque parfait entre les deux évaluateurs. Ce critère, plus objectif à évaluer, est moins sujet à interprétation.

L'analyse descriptive montre malgré tout un taux d'accord de 74,6 % entre les évaluations. Une divergence par excès est observée dans 13,2 % des cas, ce qui suggère que le modèle a appliqué des critères de détection plus stricts, voire excessifs. On retrouve 10,54 % de désaccords considérés comme justifiés, reflétant des différences d'interprétation acceptables qui peuvent favoriser une meilleure communication et une approche plus nuancée des recommandations médicales. À noter que la différence observée entre le taux d'accord global relativement élevé (74,6 %) et les valeurs faibles à modérées du Kappa de Cohen s'explique principalement par la prise en compte du hasard et la distribution déséquilibrée des critères évalués. En effet, le Kappa mesure l'accord au-delà de ce qui pourrait survenir par hasard. Ainsi, des divergences même minimes, particulièrement sur des critères rares ou très fréquents, impactent fortement le Kappa sans nécessairement réduire significativement le taux d'accord global. L'approche *«rigoureuse»* d'o1, qui applique des critères de détection souvent plus stricts que l'évaluateur humain, accentue ces divergences. Cela explique pourquoi un pourcentage

d'accord brut élevé ne garantit pas nécessairement un Kappa élevé.
Enfin, 1,45 % des divergences ont été identifiées comme de véritables erreurs de o1. Cependant, aucune de ces erreurs ne concerne le critère 5 relatif à la rupture du secret médical, ce qui souligne la fiabilité du modèle sur les aspects les plus sensibles des préconisations. En revanche, o1 a également détecté, à raison, de véritables ruptures du secret médical qui n'avaient pas été détectées par le CM. Ces préconisations avaient sûrement fait l'objet d'un consensus immédiat entre les deux étudiants et n'ont donc pas été réévaluées par des séniors. De plus, aucune hallucination n'a été observée dans les réponses générées par l'IA, démontrant la qualité du prompt utilisé et la capacité de o1 à fournir des analyses cohérentes et conformes aux attentes.
Ainsi, les divergences observées entre o1 et le CM dans cette étude semblent principalement résulter de l'approche *« rigoureuse »* de o1, qui privilégie la détection des incohérences et des erreurs potentielles par excès de précaution. Cette prudence favorise une meilleure identification des erreurs dans les préconisations médicales et offre un cadre structuré pour leur amélioration, sans nuire à l'indépendance du médecin du travail.

FORCES DE L'ÉTUDE

Une des forces de cette étude est qu'elle est la première à explorer le potentiel des grands modèles de langages à renforcer la précision et la rigueur des préconisations en médecine du travail. Elle a été réalisée sur un échantillon de préconisations suffisamment grand, garantissant une représentativité et une robustesse statistique suffisante. Aussi, plusieurs médecins du travail ont évalué les préconisations afin de diminuer les biais de subjectivité, d'information et de disponibilité. Les préconisations ont également été prélevées de la base de données de 12 SPSTI permettant de diminuer un potentiel effet centre.

LIMITES DE L'ÉTUDE

La principale limite de cette étude réside dans le fait que l'analyse est influencée par l'interprétation des évaluateurs et que certains éléments contextuels ne sont pas pris en compte. En effet, les informations précises sur le poste occupé par le salarié et sur les contraintes spécifiques de son environnement de travail ne sont pas disponibles. Il est donc possible que certaines préconisations, bien que détectées comme problématiques, aient en réalité été adaptées à des situations particulières que o1 et les évaluateurs humains ne pouvaient pas connaître. Une autre limite est que le modèle utilisé, o1, est un modèle de raisonnement qui a ses capacités spécifiques de par les données et le mode d'entraînement qui lui sont propres. Les résultats de cette étude ne sont donc pas extrapolables à d'autres modèles. Une autre limite concerne la rapidité d'évolution des LLMs. Bien que le modèle utilisé dans cette étude soit l'un des plus avancés au moment de l'analyse des données (novembre 2024), les progrès dans ce domaine sont exponentiels. Ainsi, de nouveaux modèles de raisonnement, tels que GPT-5.4, Claude Opus 4.6 (développé par Anthropic), Grok-4.1 (développé par xAI), Gemini 3 (développé par Google) sont déjà parus depuis l'analyse statistique réalisée pour cette étude et surpassent déjà les capacités d'o1 sur de nombreux *benchmarks* [19]. Cette dynamique implique que les résultats obtenus doivent être réévalués périodiquement à mesure que les capacités des LLMs évoluent.

PERSPECTIVES

Les résultats probants de cette étude peuvent présupposer de la pertinence de l'utilisation des LLMs dans d'autres aspects de la santé au travail et notamment l'aide à l'évaluation des risques professionnels ou l'accompagnement dans la rédaction et le suivi des dossiers médicaux des salariés. Au vu de la démographie médicale en berne en médecine du travail, il pourrait alors être pertinent d'intégrer cette technologie dans les pratiques des SPST afin de gagner en efficacité et proposer un meilleur suivi aux salariés. Pour y parvenir, il serait nécessaire d'évaluer rigoureusement les différents LLMs disponibles, à travers un *benchmark* dédié, testant spécifiquement leurs connaissances théoriques en santé au travail, leur raisonnement clinique et leur capacité à prendre en compte les particularités françaises. Cette démarche permettrait de vérifier leurs performances ainsi que leur innocuité avant d'en envisager une diffusion à plus grande échelle.

CONCLUSION

Les performances d'o1 dans cette étude montrent une capacité notable à identifier les erreurs de formulation dans les préconisations médicales, grâce à une rigueur élevée dans leur détection. Bien que o1 ait commis des erreurs d'appréciation, leur faible fréquence, la faible portée de ces erreurs ainsi que l'absence d'hallucination

confirment la fiabilité du modèle et son potentiel en tant qu'outil d'aide à la rédaction des préconisations en médecine du travail. Le modèle a également détecté des erreurs que le CM n'avait pas détectées au préalable. Ces résultats démontrent, *via* l'exemple d'o1, que les LLMs peuvent constituer un atout précieux pour améliorer la clarté et la pertinence des recommandations médicales, tout en renforçant la communication avec l'employeur, et donc la portée de la démarche de prévention, tout en réduisant les risques juridiques pour le médecin et les SPSTI. Par ailleurs, l'utilisation des LLMs pourrait s'avérer prometteuse pour des applications en santé au travail, telles que l'automatisation partielle des consultations, l'évaluation des risques professionnels ou encore le suivi des dossiers médicaux. Cela nécessite toutefois au préalable une évaluation rigoureuse de leurs performances en santé au travail *via* l'élaboration d'un *benchmark* dédié.

POINTS À RETENIR

- Évaluées par un consensus multidisciplinaire (CM), 78 % des préconisations des médecins du travail en Hauts-de-France présentent au moins un critère de mauvaise qualité.
- Le modèle de langage (LLM) o1 a évalué 385 préconisations et ses résultats ont été comparés à ceux du CM.
- Le taux d'accord global entre o1 et le CM a atteint 74,6 %.
- Aucune hallucination n'a été détectée sur l'ensemble des préconisations analysées.
- L'accord est presque parfait (κ = 0,82) pour la détection des ruptures du secret médical.
- Le modèle tend à surévaluer les erreurs (13,2 % de discordances par excès).
- Des erreurs initialement non détectées par le CM ont été identifiées à raison par le modèle.
- Le taux de discordances par défaut reste faible (1,45 %), sans aucune erreur sur le critère *« rupture du secret médical ou atteinte à la vie privée du salarié »*.
- Les LLMs représentent un outil d'aide prometteur pour la relecture des préconisations en santé au travail.
- Un *benchmark* dédié en santé au travail est nécessaire pour évaluer rigoureusement les performances des différents LLMs pour d'autres applications.

BIBLIOGRAPHIE

1 | N'Guessan C – Analyse critique de la rédaction actuelle de préconisations chez les médecins du travail en Hauts de France. Thèse pour le diplôme d'État de docteur en médecine. Lille : Université de Lille, Faculté de Médecine Henri Warembourg ; 2024 : 46 p.

2 | Un premier bilan mitigé du contentieux des avis du médecin du travail. Communiqué de presse. AvoSial, LexisNexis, 2022 (https://www.avosial.fr/medias/org-1522/shared/inaptitude-au-travail-etude-avosial-20.02.2022.pdf).

3 | Telenti A, Auli M, Hie BL, Maher C et al. – Large language models for science and medicine. *Eur J Clin Invest*. 2024 ; 54 (6) : e14183.

4 | Ouyang S, Zhang JM, Harman M, Wang M – An Empirical Study of the Non-determinism of ChatGPT in Code Generation. *ACM Trans Softw Eng Methodol*. 2025 ; 34 (2) : 1-28.

5 | Vaswani A, Shazeer N, Parmar N, Uszkoreit J et al. – Attention Is All You Need. In: ArXiv. Cornell University (Cornell Tech), 2017 (https://arxiv.org/abs/1706.03762).

6 | Sajjadi Mohammadabadi SM, Kara BC, Eyupoglu C, Uzay C et al. – A Survey of Large Language Models: Evolution, Architectures, Adaptation, Benchmarking, Applications, Challenges, and Societal Implications. *Electronics*. 2025 ; 14 (18) : 3580.

7 | Sharma S, Mittal P, Kumar M, Bhardwaj V – The role of large language models in personalized learning: a systematic review of educational impact. *Discov Sustain*. 2025 ; 6 : 243.

FIN DE LA BIBLIOGRAPHIE
PAGE SUIVANTE

L'intelligence artificielle en santé au travail : évaluation de son potentiel pour renforcer la précision et la clarté des préconisations médicales

BIBLIOGRAPHIE (suite)

8 | Huang L, Yu W, Ma W, Zhong W et al. – A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans Inf Syst*. 2025 ; 43 (2) : 1-55.

9 | L'apprentissage du raisonnement avec les LLM. OpenAI, 2024 (https://openai.com/index/learning-to-reason-with-llms/).

10 | Xu N, Ma X – DecoPrompt: Decoding Prompts Reduces Hallucinations when Large Language Models Meet False Premises. In: ArXiv. Cornell University (Cornell Tech), 2025 (https://arxiv.org/abs/2411.07457).

11 | Chen B, Zhang Z, Langrené N, Zhu S – Unleashing the potential of prompt engineering for large language models. *Patterns (N Y)*. 2025 ; 6 (6) : 101260.

12 | White J, Fu Q, Hays S, Sandborn M et al. – A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. In: ArXiv. Cornell University (Cornell Tech), 2023 (https://arxiv.org/abs/2302.11382).

13 | Levy M, Jacoby A, Goldberg Y – Same Task, More Tokens: the Impact of Input Length on the Reasoning Performance of Large Language Models. In : Ku LW, Martins A, Srikumar V (Eds) – Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). ACL Anthology, 2024 (https://aclanthology.org/2024.acl-long.818.pdf).

14 | Nori H, King N, McKinney SM, Carignan D et al. – Capabilities of GPT-4 on Medical Challenge Problems. In: ArXiv. Cornell University (Cornell Tech), 2023 (https://arxiv.org/abs/2303.13375).

15 | Sorin V, Brin D, Barash Y, Konen E et al. – Large Language Models and Empathy: Systematic Review. *J Med Internet Res*. 2024 ; 26 : e52597.

16 | Goh E, Gallo RJ, Strong E, Weng Y et al. – GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. *Nat Med*. 2025 ; 31 (4) : 1233-38.

17 | Havrilla A, Du Y, Raparthy SC, Nalmpantis C et al. – Teaching Large Language Models to Reason with Reinforcement Learning. In: ArXiv. Cornell University (Cornell Tech), 2024 (https://arxiv.org/abs/2403.04642).

18 | Wei J, Wang X, Schuurmans D, Bosma M et al. – Chain-of-Thought Prompting Elicits Reasoning in Large Language Models In: ArXiv. Cornell University (Cornell Tech), 2023 (https://arxiv.org/abs/2201.11903).

19 | Humanity's Last Exam. Center for AI Safety & Scale, 2026 (https://agi.safe.ai/).