ECOS augmentés : l'IA entre dans le jeu

Évaluation du rôle des grands modèles de langage dans la simulation et la génération des ECOS

Grégoire Micicoi & Élise Lupon

CCA-AHU, chirurgie orthopédique et plastique

Institut Universitaire Locomoteur et du Sport, Hôpital Pasteur 2, Université Côte d'Azur, Nice, France

Introduction

- **IA dans l'enseignement :**

- GPT-4 / USMLE : taux de précision 80 - 100 %

Brin D et al., Discov Appl Sci, 2024

- GPT-4 / examens de type internat : comparables aux étudiants en médecine de dernière année

Kung TH et al., Plos Digit Health, 2023

- IA / examens médicaux nationaux : réussite, performances diminuent pour les tâches cliniques plus dépendantes du contexte.

Liu M et al., Int J Med Inf. 2025

- ***ECOS :***

- Référence en matière d'évaluation basée sur la performance dans l'enseignement Clinique

- Ressources **considérables**..

Peu d'études sur le rôle LLM et ECOS publiées

Hypothèse

LLM peut jouer un double rôle dans l'enseignement basé sur les ECOS :

1/ comme évaluateur capable d'interagir avec les grilles d'évaluation

2/ comme générateur de stations d’ECOS complètes

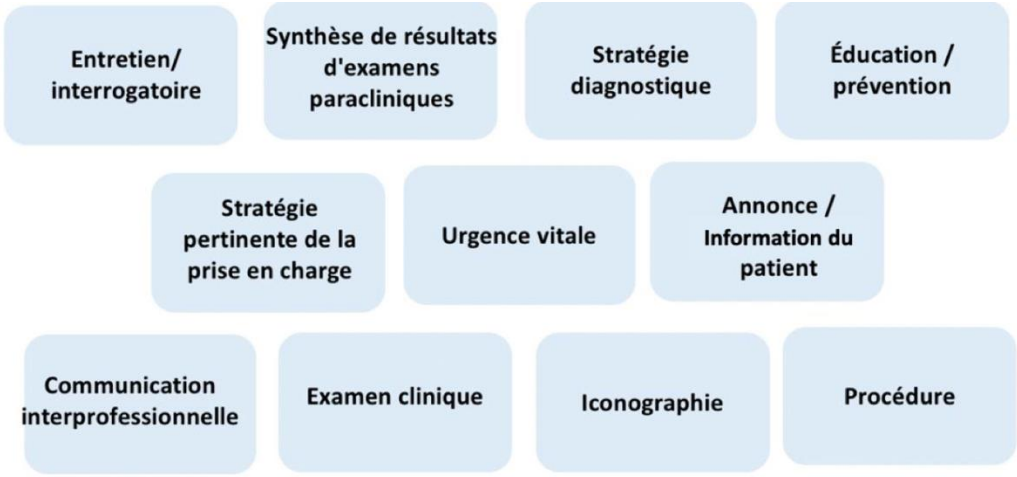
Objectifs

1/ Simuler les réponses des candidats dans des stations d’ECOS validées et mesurer leurs performances

2/ Générer de manière autonome des stations ECOS et évaluer leurs performances

Méthodes

- **11 domaines de compétences**



- **356 Situations cliniques de Départ (SDD)**
- **367 Items du programme LiSA**
- **3 838 Objectifs d'apprentissage (OIC)**

uness livret LiSA

- **10 stations : couvrent l'ensemble des domaines – dont procédure ; pédiatrie ; gériatrie**

ECOS joués par LLM

Utilisation ChatGPT-4o

ECOS Facultaires (validés)

Étudiants 6ème année

10 stations
11 domaines explorés
Items Rang A et B *(référentiel national)*

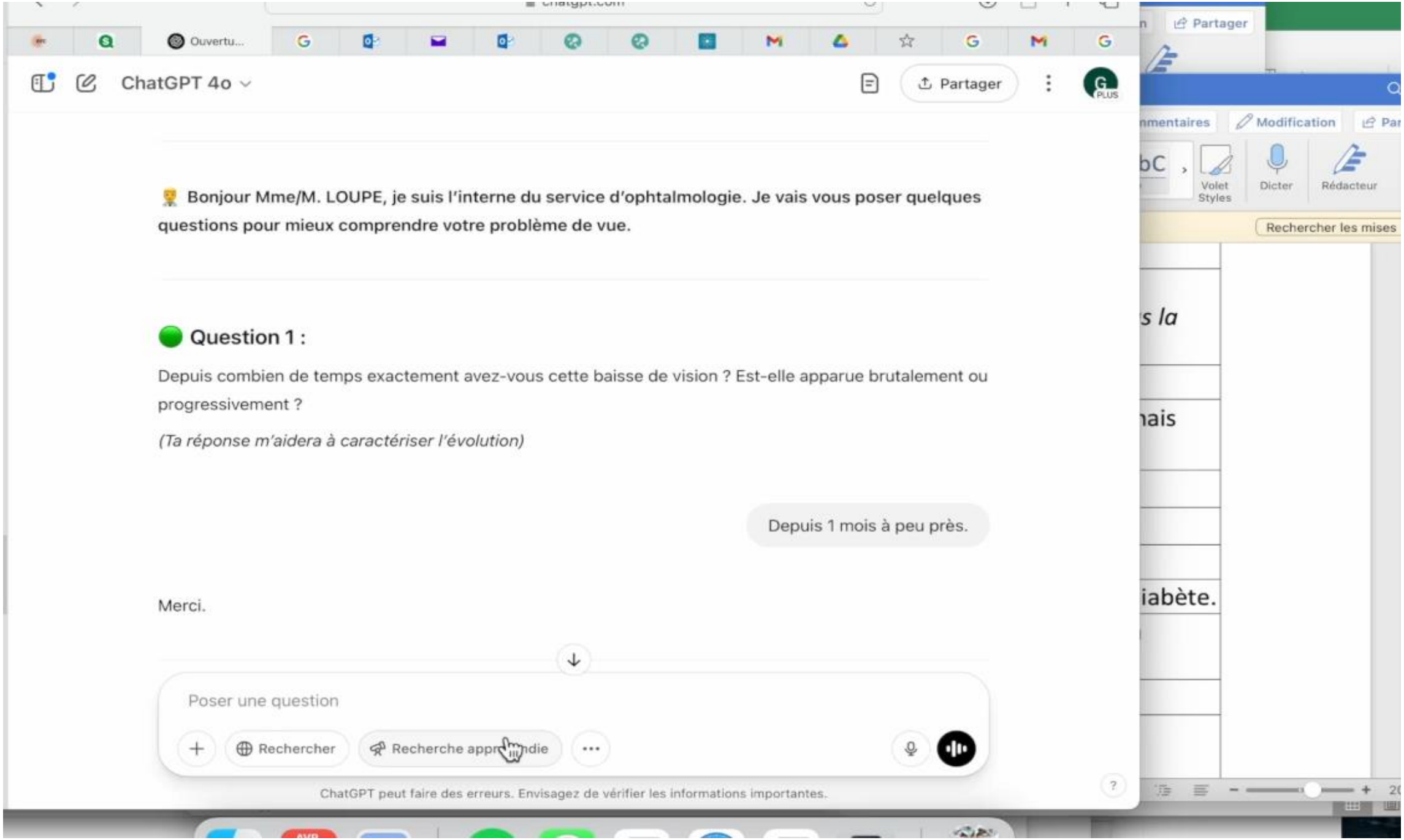
Score Global (& corrigé)

Items A omis – items supplémentaires

Hallucinations

IA comme étudiant évalué : Caractéristiques des 10 stations d'ECOS jouées par le LLM

No.	SDD	Spécialité	Domaine de compétence primaire	Domaine de compétence secondaire	Item(s) LISA concerné(s)	Patient simulé	Mannequin	Document de support	Nombre d'items évalués
1	15 –Anomalie de la couleur des extrémités	Dermatologie / Médicine vasculaire	iconographie	Synthèse de résultats d'examens paracliniques	Acrosyndromes ; Néphropathies vasculaires	NON	NON	Photo d'un pied	14
2	138 – Anomalie de la vision	Ophtalmologie	Entretien / interrogatoire	Stratégie diagnostique	Anomalie de la vision	OUI	NON	Aucun	15
3	118 – Céphalée	Neurologie	Examen clinique	Stratégie pertinente de la prise en charge	Hémorragie méningée	NON	OUI	Aucun	15
4	249 – Prescrire des AINS	Thérapeutique	Éducation / prévention	Stratégie pertinente de la prise en charge	Traitements analgésiques, approches médicamenteuses et alternatives	NON	NON	Aucun	15
5	140 – Baisse de l'audition / surdité	ORL	Procédures	Examen clinique	Otite chez les enfants et les adultes	NON	OUI	Aucun	12
6	13 – Nausées / vomissements	Soins palliatifs	Examen clinique	Stratégie pertinente de la prise en charge	Obstruction intestinale chez l'adulte et l'enfant ; vomissements palliatifs	OUI	NON	Aucun	13
7	44 – Hyperthermie / Fièvre	Oncologie	Entretien / interrogatoire	Urgence vitale	Fièvre chez les patients immunodéprimés ; prise en charge de la neutropénie fébrile	OUI	NON	Résultat d'un bilan sanguin	14
8	208 – Hyperglycémie	Endocrinologie / Diabétologie	Annonce / Information du patient	Éducation / prévention	Diabète de type 1 et 2 chez les enfants et les adultes	OUI	NON	Résultat d'un bilan sanguin	14
9	108 – Anomalie des bourses	Pédiatrie	Communication interprofessionnelle	Stratégie pertinente de la prise en charge	Pathologie génito-scrotale chez le garçon et chez l'homme	OUI	NON	Photo d'un scrotum d'enfant	15
10	298 – Consultation et suivi d'un patient présentant des troubles cognitifs	Gériatrie	Stratégie diagnostique	Procédures	Confusion, démence	OUI	NON	Aucun	13


ChatGPT 4o
Partager
Bonjour Mme/M. LOUPE, je suis l'interne du service d'ophtalmologie. Je vais vous poser quelques questions pour mieux comprendre votre problème de vue.
Question 1 :
Depuis combien de temps exactement avez-vous cette baisse de vision ? Est-elle apparue brutalement ou progressivement ?
(Ta réponse m'aidera à caractériser l'évolution)
Depuis 1 mois à peu près.
Merci.
Poser une question
Rechercher
ChatGPT peut faire des erreurs. Envisagez de vérifier les informations importantes.
Partager
Modification
Dicter
Rédacteur
Rechercher les mises

ECOS générés par LLM

Utilisation ChatGPT-4o

Entrainement et réalisation d'un *prompt*
> 60 heures

Étudiants 6ème année

Génération de 6 ECOS

Relecture Tut'ECOS

- Modification > 5 items = Nouvel ECOS
- < 5 items : modifications manuelles

Joués par Tut'ECOS

> 120 étudiants

Mesure des performances :

Score global, discrimination, effet plafond, perception IA

Performances évaluées – ECOS générés

$$CV = \frac{\text{écart-type}}{\text{moyenne}} \times 100$$

CV (%)	Interprétation
< 10%	Très faible variabilité
10–15%	Faible variabilité
15–30%	Variabilité modérée
> 30%	Forte variabilité (discriminant)

ECOS Global

$$\text{ID} = \frac{N_s + N_i}{2n} \qquad \text{DPI} = \frac{N_s - N_i}{n}$$

N_s = nombre de bonnes réponses dans le groupe **haut score** (top 27%)

N_i = nombre de bonnes réponses dans le groupe **bas score** (bottom 27%)

n = nombre d'étudiants dans chaque groupe (27% de ton échantillon total)

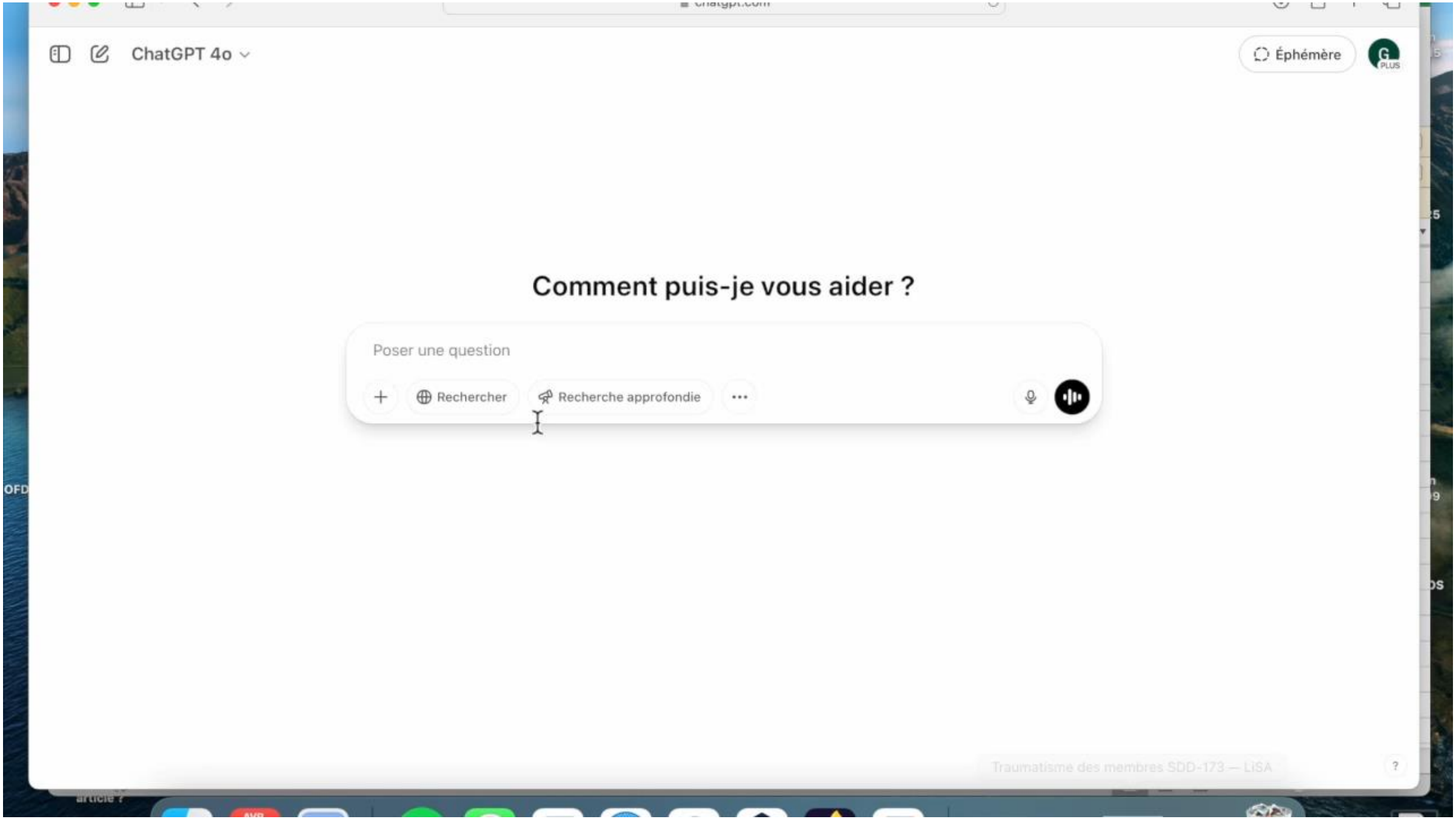
Valeur	Interprétation
> 0.8	Item **très facile**
0.3 – 0.8	Item **équilibré**
< 0.3	Item **très difficile**

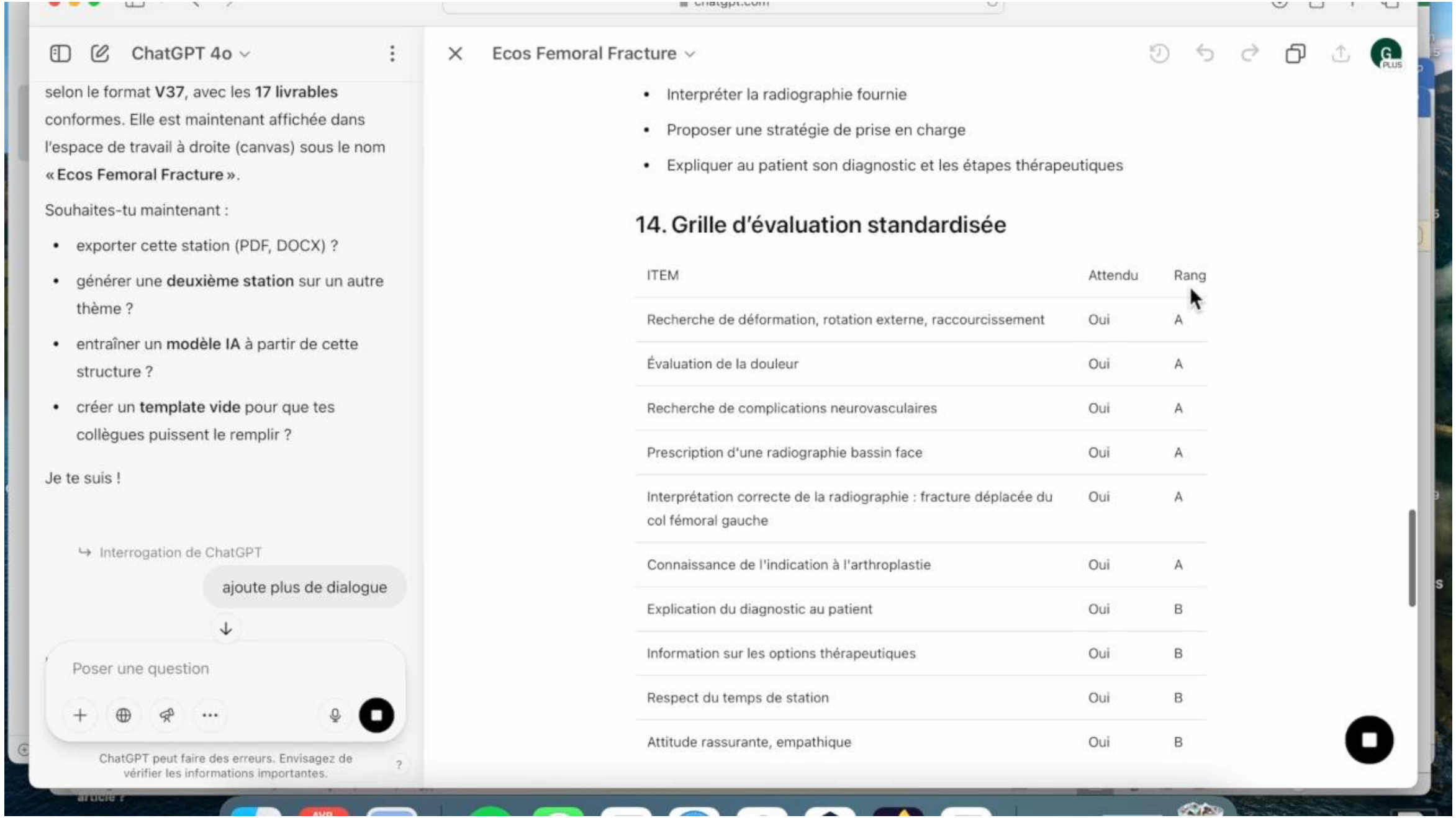
Valeur	Interprétation
> 0.4	Très bon item : **discrimine bien** entre bons et faibles
0.2 – 0.4	Acceptable
< 0.2	Peu discriminant
< 0	Discrimine à l'envers

Items grille

IA générative : Caractéristiques des stations ECOS générées

No.	SDD	Spécialité	Domaine de compétence primaire	Domaine de competence secondaire	Related LiSA Item(s) LiSA concerné(s)	Patient simulé	Mannequin	Documents supplémentaires
1	40 – Écoulement mamelonnaire	Gynécologie	Examen clinique	Procédures	Tumeurs du sein	Non	Oui	
2	173 – Traumatisme des membres	Chirurgie Orthopédique	Examen clinique	Interprétation iconographique	Fractures courantes chez les adultes et les personnes âgées	Oui	Non	Radiographie pelvienne (Garden IV)
3	10 – Méléna / Rectorragies	Médecine générale	Examen clinique	Procédures	Tumeurs colorectales	Non	Oui	
4	170 – Plaie (suture)	Médecine d'urgence / Chirurgie mineure	Procédures	Aucun	Soins préhospitaliers ; Gestion de la douleur	Non	Oui	
5	38 – État de mort apparente	Soins intensifs	Procédures	Urgence	États de choc	Non	Oui	Tracé ECG
6	140 – Baisse de l'audition / Surdité	ORL	Procédures	Aucun	Troubles auditifs	Non	Oui	Audiogramme


ChatGPT 4o
Éphémère
Comment puis-je vous aider ?
Poser une question
Rechercher
Recherche approfondie


ChatGPT 4o
selon le format V37, avec les 17 livrables conformes. Elle est maintenant affichée dans l'espace de travail à droite (canvas) sous le nom « Ecos Femoral Fracture ».
Souhaites-tu maintenant :
exporter cette station (PDF, DOCX) ?
générer une deuxième station sur un autre thème ?
entraîner un modèle IA à partir de cette structure ?
créer un template vide pour que tes collègues puissent le remplir ?
Je te suis !
Interrogation de ChatGPT
ajoute plus de dialogue
Poser une question
ChatGPT peut faire des erreurs. Envisagez de vérifier les informations importantes.
Ecos Femoral Fracture
Interpréter la radiographie fournie
Proposer une stratégie de prise en charge
Expliquer au patient son diagnostic et les étapes thérapeutiques
14. Grille d'évaluation standardisée
ITEM
Attendu
Rang
Recherche de déformation, rotation externe, raccourcissement
Oui
A
Évaluation de la douleur
Oui
A
Recherche de complications neurovasculaires
Oui
A
Prescription d'une radiographie bassin face
Oui
A
Interprétation correcte de la radiographie : fracture déplacée du col fémoral gauche
Oui
A
Connaissance de l'indication à l'arthroplastie
Oui
A
Explication du diagnostic au patient
Oui
B
Information sur les options thérapeutiques
Oui
B
Respect du temps de station
Oui
B
Attitude rassurante, empathique
Oui
B

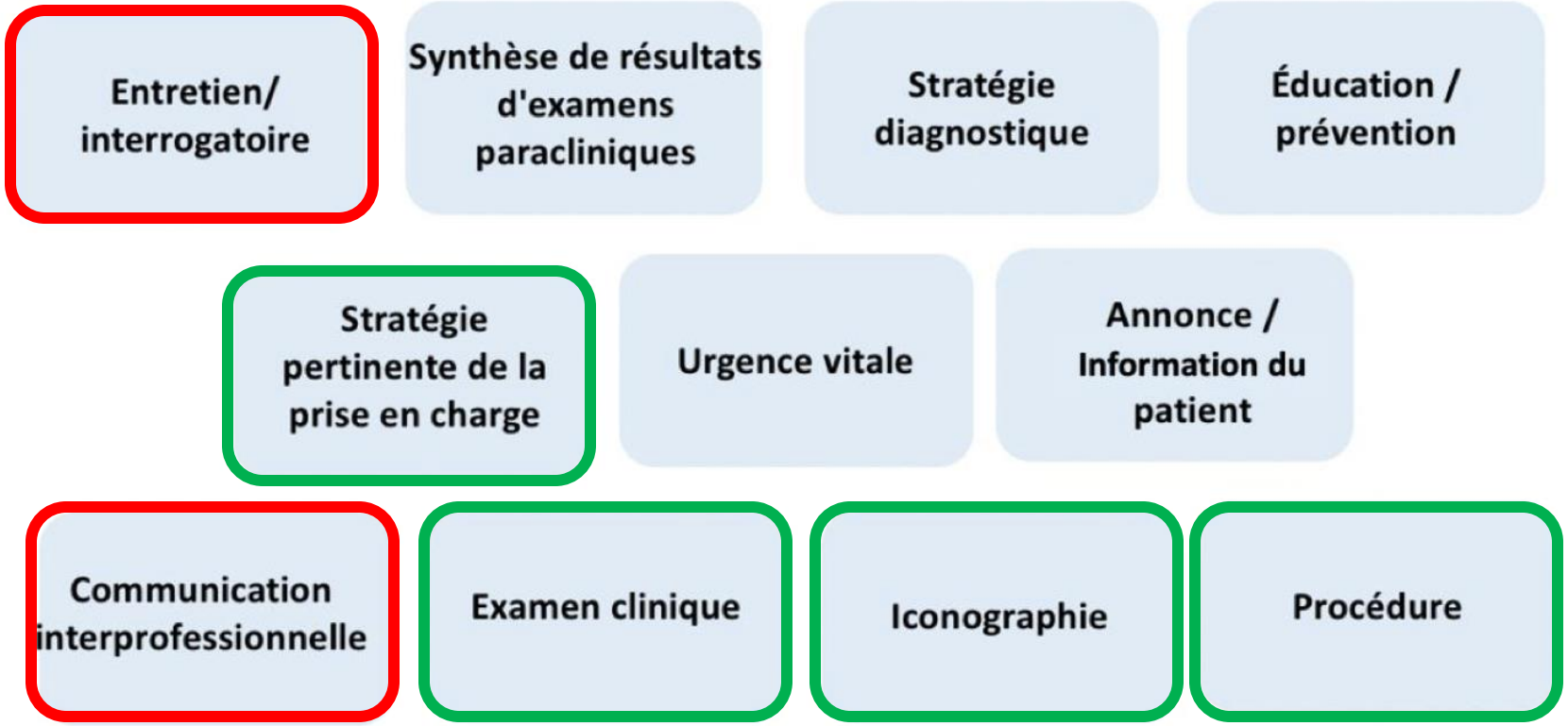
Résumé des indicateurs de performance pour le LLM dans les 10 stations ECOS évaluées

No.	Situation Clinique de Départ (SCD)	Nombre d'items évalués	Score global (%)	Omission d'items de niveau A, % (n items)	Performances LLM avec ajout référentiel, % (n items ajoutés)	Items supplémentaires ajoutés, % (n)	Pertinence des items supplémentaires %, (n, parmi les items supplémentaires)
1	15 –Anomalie de la couleur des extrémités	14	92,8	0 (0)	100 (1)	14.3 (2)	100,0 (2)
2	138 – Anomalie de la vision	15	73,3	13,3 (2)	100 (4)	20 (3)	100,0 (3)
3	118 – Céphalée	15	93,3	6,6 (1)	100 (1)	13,3 (2)	100,0 (2)
4	249 – Prescrire des AINS	15	73,3	13,3 (2)	93,3 (3)	13,3 (2)	100,0 (2)
5	140 – Baisse de l'audition / Surdité	12	91,7	0 (0)	100 (1)	33,3 (4)	75,0 (3)
6	13 – Nausées / vomissements	13	76,9	7,6 (1)	84,6 (1)	30,7 (4)	100,0 (4)
7	44 – Hyperthermie / Fièvre	14	85,7	14,3 (2)	92,8 (1)	28,6 (4)	100,0 (4)
8	208 – Hyperglycémie	14	85,7	7,1 (1)	92,8 (1)	21,4 (3)	100,0 (3)
9	108 – Anomalie des bourses	15	66,7	13,3 (2)	93,3 (4)	20,0 (3)	100,0 (3)
10	298 – Consultation et suivi d'un patient présentant des troubles cognitifs	13	69,2	23,0 (3)	92,3 (3)	23,0 (3)	100,0 (3)
			80,8 % ± 10,8	**9,9 % ± 7,2**	**94,7 % ± 6,2**	**+ 20,2 % ± 6,7** **1 hallucination**	

7 oublis :

- 3 non dans documents officiels
- 2 incohérences de grilles
- 2 non trouvé

Scores globaux de performance en fonction du domaine de compétence



73,7 % ± 7,4

91,6 % ± 4,1

***p* < 0.01**

ECOS générées par IA : résultats des étudiants et analyse des items (discrimination – difficulté)

No.	SCD	Spécialité	Nombre d'étudiants participants	Résultat moyen ± EC (/20)	Min – Max (/20)	Coefficient de Variation	DIFI moyen	DISI moyen
1	40 – Écoulement mamelonnaire	Gynécologie	136	14,1 ± 2,6	5,0 – 17,5	18,6 %	0,75 ± 0,07	0,40 ± 0,07
2					19,5	21,3 %	0,70 ± 0,07	0,39 ± 0,07
3					18,6	17,2 %	0.66 ± 0,08	0,42 ± 0,06
4					16,4	19,3 %	0,79 ± 0,06	0,37 ± 0,08
5	38 – État de mort apparente	Soins intensifs	128	17,7 ± 2,1	10,0 – 20,0	11,7 %	0,82 ± 0,06	0,38 ± 0,08
6	140 – Baisse de l'audition / Surdité	ORL	130	14,9 ± 3,9	5,0 – 20,0	26,7 %	0.78 ± 0,05	0,35 ± 0,09
				14,9 ± 2,9		**19,1 ± 5,8 %**	**0,75 ± 0,07**	**0,39 ± 0,07**

CV (%)	Interprétation
< 10%	Très faible variabilité
10–15%	Faible variabilité
15–30%	Variabilité modérée
> 30%	Forte variabilité (discriminant)

Difficulté & Discrimination des items (n = 56)

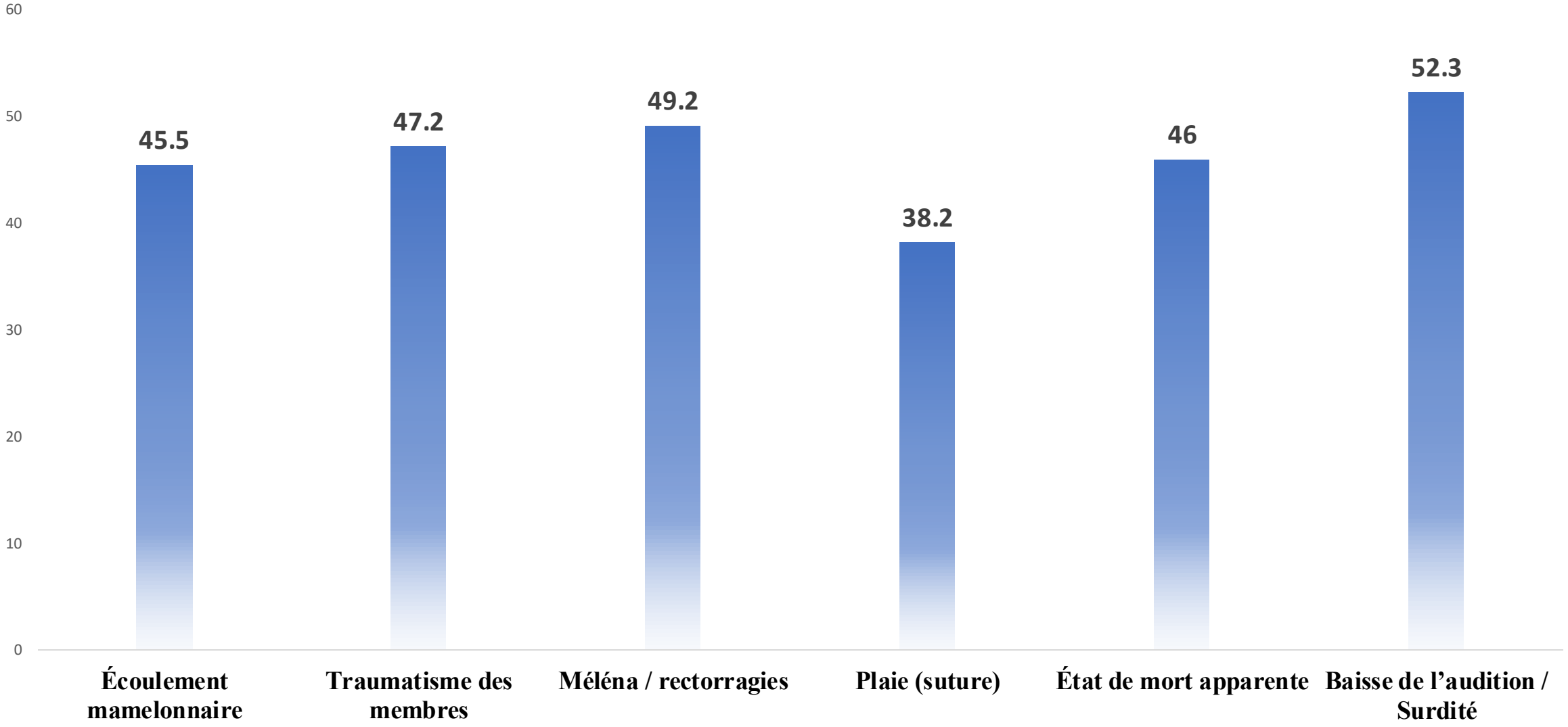
Catégorie	Valeur cible	Nombre d'items n (%)
Item très facile	> 0.8	8 (14,3%)
Item équilibré	0.3 – 0.8	42 (75%)
Item très difficile	< 0.3	6 (10,7%)

Catégorie	Valeur cible	Nombre d'items (n)
Très bon item	> 0.4	14 (25%)
Item acceptable	0.2 – 0.4	33 (58,9%)
Item peu discriminant	< 0.2	9 (16,1%)
Discrimine à l'envers	< 0	0

Effet plafond (> 25 % note maximale)

- **Etat de Mort apparente : 28,2%**
- **Suture plaie : 27,8 %**

Pensez-vous que cette station a été générée par une IA ?



Discussion

A propos d'un cas d'OCSE gynécologie-obstrétique :

Score ChatGPT = 77,2% vs Étudiants (n=26) = 73,7%

Li et al, Am J Obstet Gynecol, 2023

Intérêt *prompt refinement* : améliore les capacités du LLM de 10 à 43%

Experts identifiaient une IA dans 61% des cas *(vs. étudiant)* – **correction des grilles**

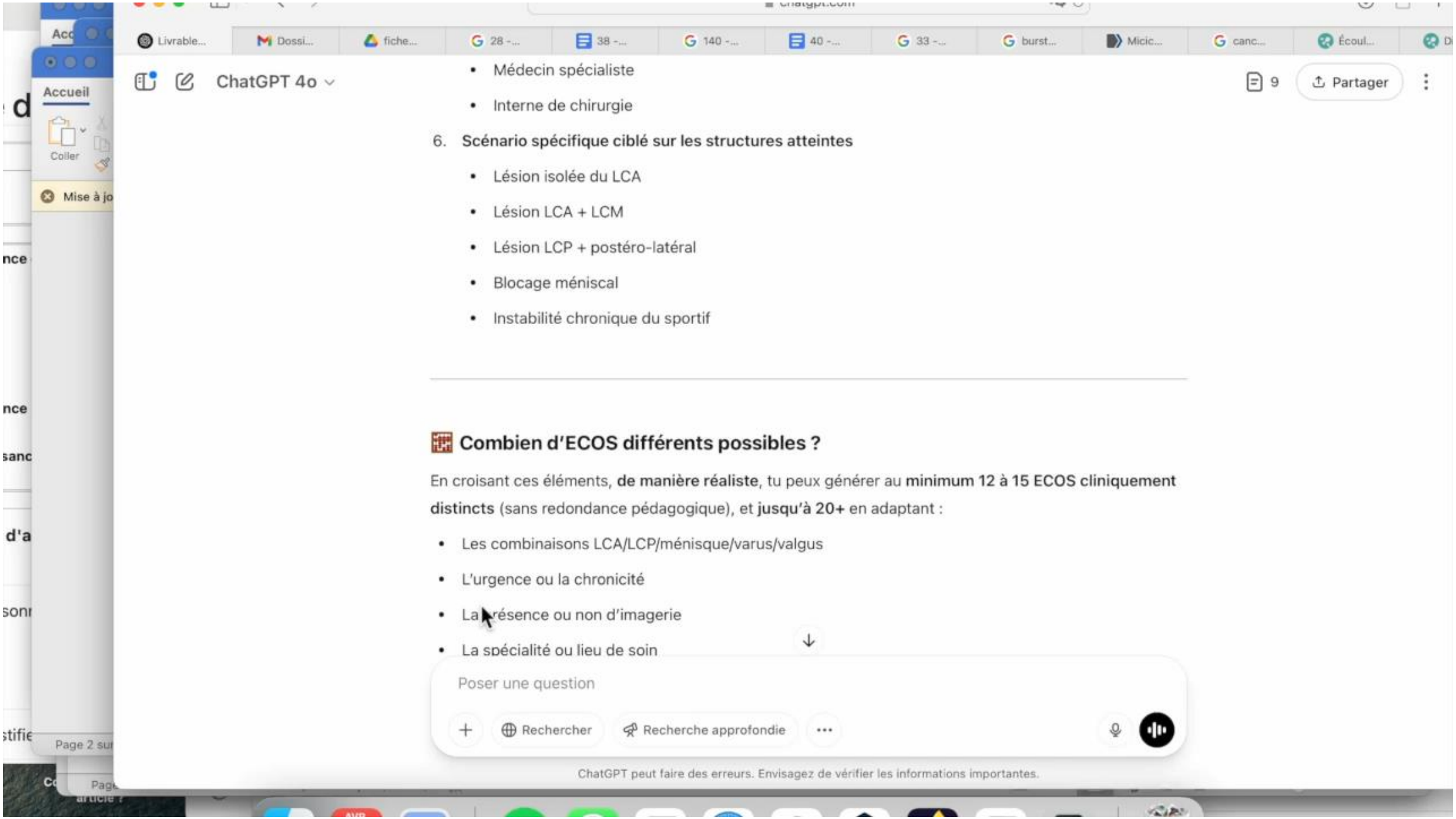
Nayak et al, JAMA Intern Med, 2023

2,027 vidéos OCSE analysés IA : LLM identifie étapes cliniques et standardisées

(-) ''nuances'' raisonnement clinique **–gain de temps et aide à la correction**

Shakur et al, ArXiv, 2024

Supervision humaine reste indispensable – Nécessité d'un Jugement expert


ChatGPT 4o
9
Partager
Médecin spécialiste
Interne de chirurgie
6. Scénario spécifique ciblé sur les structures atteintes
Lésion isolée du LCA
Lésion LCA + LCM
Lésion LCP + postéro-latéral
Blocage méniscal
Instabilité chronique du sportif
Combien d'ECOS différents possibles ?
En croisant ces éléments, de manière réaliste, tu peux générer au minimum 12 à 15 ECOS cliniquement distincts (sans redondance pédagogique), et jusqu'à 20+ en adaptant :
Les combinaisons LCA/LCP/ménisque/varus/valgus
L'urgence ou la chronicité
La présence ou non d'imagerie
La spécialité ou lieu de soin
Poser une question
Rechercher
Recherche approfondie
ChatGPT peut faire des erreurs. Envisagez de vérifier les informations importantes.

Limites

- Peu d’ECOS générés
- Coefficient de Variation peu adapté
- ECOS essentiellement procéduraux
- Nécessite prompt engineering
- Validation Test de Turing

Take-Home Messages

Excellentes performances :

- simulation des candidats aux ECOS
- aide à l'évaluation par des experts (incohérences, items manquants)

ECOS générées par l'IA : valides, pouvoir discriminant limité, gamme de difficultés restreinte

Gain de temps potentiel, supervision humaine indispensable