

Correction d'ECOS avec l'IA : une expérimentation empirique

Marc Cuggia*, Boris Delange, Zine-dine Khene

Professeur des Universités - Praticien Hospitalier

Informatique médicale, Université de Rennes / CHU de Rennes

CMA action 1 : EDSAN - Education en SANTé Numérique



Introduction

- **Contexte :**

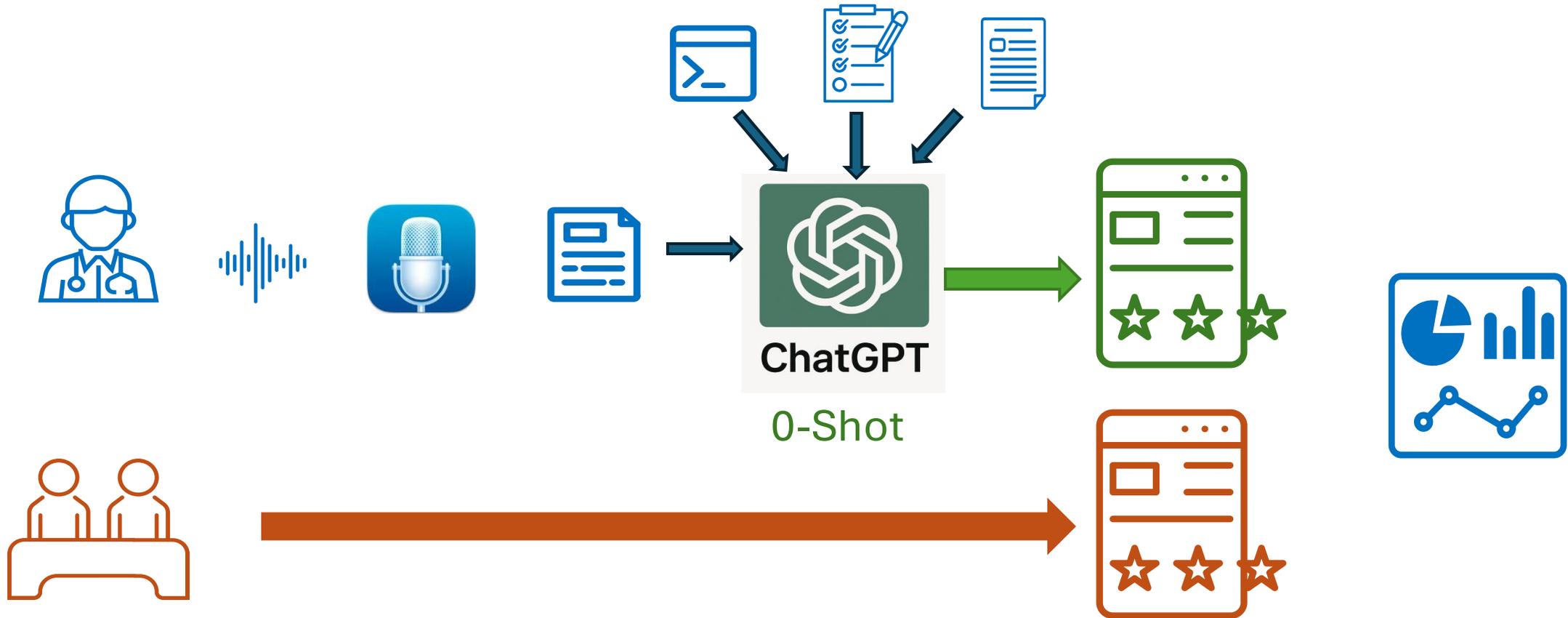
- Expérimentation menée en mars 2025 avec des étudiants de DFASM3
- Utilisation d'un LLM (GPT-4) pour analyser des réponses audio à une station réelle d'ECOS
- Objectif : explorer les **apports et limites** de l'IA générative pour l'évaluation des Ecos

Matériel et méthodes

Station ECOS utilisée pour l'évaluation

- **Nom** : Station 2 – ECOS DFASM3 (26 mars 2025)
- **Contexte clinique simulé** :
 - Homme de 65 ans
 - Colique néphrétique droite documentée 24h plus tôt (calcul 6 mm, AINS prescrits)
 - Nouvelle consultation pour douleur persistante, fièvre, frissons
 - Constantes : PA 85/45, FC 120, T° 39.5°C, FR 20/min, Glasgow 15

Plan expérimental





Axe d'évaluation	Critères	Barème
1. Synthèse des données cliniques	Capacité à sélectionner, justifier, hiérarchiser les éléments pertinents	0 à 1
2. Proposition de prise en charge	Pertinence, justification, absence d'inutile, ordre des priorités	0 à 1

N°	Item évalué	Barème
1	Diagnostic : pyélonéphrite obstructive / aiguë grave / colique néphrétique compliquée	2 pts
2	Évoque l'urgence vitale	2 pts
3	Évoque le pronostic potentiellement grave	1 pt
4	Évoque le risque de sepsis / choc septique / qSOFA ≥ 1	1 pt
5	Prescrit des hémocultures	1 pt
6	Prescrit un ECBU	1 pt
7	Prescrit NFS, ionogramme sanguin et créatinine	1 pt
8	Demande une hospitalisation	1 pt
9	Prescrit une antibiothérapie	1 pt
10	Antibiothérapie probabiliste	1 pt
11	Antibiothérapie intraveineuse	1 pt
12	Remplissage vasculaire (expansion volémique)	1 pt
13	Laisse le patient à jeun	1 pt
14	Évoque la nécessité d'une dérivation urinaire en urgence (mention explicite du caractère urgent)	2 pts
15	Cite au moins une technique : double J ou néphrostomie	1 pt

Mise au point du prompt

- Tu es examinateur d'une station ECOS dans le cadre d'une évaluation facultaire de DFASM3 (étudiants en fin de 5e année de médecine).
- Tu dois évaluer une réponse étudiante enregistrée à l'oral et transcrite textuellement.
- **Ta mission est la suivante :**
 - Lire l'énoncé de la station clinique.
 - Lire la grille d'évaluation officielle (15 items notés sur 18 points).
 - Lire la transcription textuelle de la réponse de l'étudiant (en français, sans reformulation).
 - Pour chaque item, dire si l'objectif est atteint (score = 1 ou 2 selon la pondération) ou non (score = 0).
 - Tu dois te baser uniquement sur ce que dit l'étudiant.
 - Si un élément n'est pas dit explicitement ou n'est pas formulé de façon acceptable, le point n'est pas accordé.
 - Calculer le total sur 18.
 - Évaluer les deux critères transversaux :
 - Aptitude à faire la synthèse des données
 - Aptitude à proposer une prise en charge adaptée
 - Pour chacun, attribue un score entre 0 et 1 (par pas de 0,25), **avec justification.**
- **Tu dois être aussi rigoureux qu'un examinateur humain, sans interprétation généreuse.**
- **Conserve la langue française pour toute ta réponse.**



Evaluation

- Deux examinateurs ont évalué les réponses étudiantes selon la grille ECOS, avec concertation pour harmonisation.
- GPT-4 a reçu les transcriptions brutes et la grille standardisée, sans guidage supplémentaire.
- Comparaison item par item avec mesure de l'accord (même score = accord).



Item de la grille	Score examinateur humain	Score GPT-4	Accord
Diagnostic	2	2	O
Urgence vitale	2	2	O
Hémocultures	1	0	N
Dérivation urinaire urgente évoquée	2	0	N
Antibiothérapie IV probabiliste	2	2	O

Résultats - Exemple



- Ma principale hypothèse diagnostique pour Monsieur Y, c'est une pyélonéphrite aiguë compliquée par obstruction de l'uretère. Donc c'est une pyélonéphrite aiguë compliquée, et qui semble être grave devant l'hypotension de Monsieur Y à 85/45. Donc, son pronostic, c'est qu'il risque d'être en choc septique, et donc avec un risque vital.
- Les examens que je vais prescrire : on va faire une BU, un ECBU, un prélèvement des urines. On peut faire une échographie rénale, mais dans le contexte de pyélonéphrite aiguë compliquée, on va plutôt aller vers un urosanner, non injecté, pour vérifier que le calcul est toujours enclavé, même s'il y a peu de doute, et chercher des foyers de néphrite.
- On va faire des hémocultures, deux paires d'hémocultures en urgence devant son hyperthermie. On va faire un bilan sanguin complet avec une NFS, une CRP, un ionogramme sanguin, urée, créatinine, gaz du sang, lactates. On peut faire un bilan hépatique, devant le risque de... il est en sepsis. Je reviendrai peut-être après.
- Et concernant la prise en charge détaillée en urgence : déjà, il va falloir hospitaliser Monsieur Y en urgence. Il va falloir le mettre en condition, lui poser deux voies veineuses périphériques. Pour le moment, il ne semble pas avoir besoin d'oxygène, mais garder l'oxygène à proximité s'il désature davantage.
- Appeler le réanimateur pour lui faire part de notre inquiétude concernant Monsieur Y, qui risque de décompenser rapidement devant son uretère bouché. Il va falloir appeler le chirurgien urologue pour lui dire qu'il va falloir faire une désobstruction en urgence, dévier les urines, aller enlever ce calcul, cette lithiase urinaire, pour que les urines puissent se déverser et arrêtent de stagner dans le rein droit.
- Il va falloir mettre une antibiothérapie intraveineuse en urgence, probabiliste, à base de C3G et aminoside, par exemple céfotaxime et amikacine. Lui mettre des antalgiques, il présente une douleur évaluée à 9 sur 10, donc on peut faire une titration morphinique IV, associée à du paracétamol. Il faut surtout arrêter les AINS.
- Scoper Monsieur Y, donc surveiller ses constantes : fréquence cardiaque, fréquence respiratoire, saturation, tension artérielle en continu. L'urgence, c'est vraiment de le drainer : par une néphrostomie ou une sonde JJ. On peut lui faire un ECG aussi, comme il est aux urgences, pour vérifier que le cœur n'est pas atteint.
- L'antibiothérapie, on l'adaptera dans un second temps, selon les résultats des hémocultures et de l'ECBU prélevé. Et j'ai oublié : il faut lui prescrire un remplissage, parce qu'il est quand même hypotendu, donc un remplissage par NaCl en débit libre. Et si sa tension remonte, on pourra lui laisser une hydratation. Si ça ne remonte pas, il faudra prévoir d'administrer de la noradrénaline.
- Voilà. Merci.



Grille d'observation (18 points)

N°	Critère	Points possibles	Évaluation stricte
1	Diagnostic (pyélonéphrite obstructive, pyélonéphrite aiguë grave, colique néphrétique compliquée/fébrile)	2	2 (pose clairement le diagnostic de « pyélonéphrite aiguë compliquée, obstructive »)
2	Évoque l'urgence vitale	2	2 (mention explicite de la gravité et de la nécessité d'agir « en urgence »)
3	Pronostic potentiellement grave	1	1 (mention d'un risque vital « risque de décès »)
4	Risque de choc septique / sepsis / qSOFA	1	1 (mention claire « choc septique »)
5	Prescrit des hémocultures	1	1 (dit « deux paires d'hémocultures en urgence »)
6	Prescrit un ECBU	1	1 (« BU ECBU »)
7	Prescrit NFS + ionogramme + créatinine	1	1 (NFS, CRP, iono, urée/créat clairement évoqués)
8	Demande une hospitalisation	1	1 (hospitalisation « en urgence »)
9	Prescrit une antibiothérapie	1	1 (C3G + aminoside)
10	Antibiothérapie probabiliste	1	1 (antibiothérapie probabiliste, adaptée aux ECBU/hémocultures)
11	Voie intraveineuse	1	1 (AB IV clairement indiquée)
12	Remplissage vasculaire	1	1 (mention du remplissage IV pour hypotension, NaCl en débit libre)
13	Laisse le patient à jeun	1	0 (non mentionné)
14	Évoque la nécessité d'une dérivation urinaire en urgence	2	2 (appel urologue pour « dérivation en urgence »)
15	Évoque au moins une technique (sonde JJ ou néphrostomie)	1	1 (évoque « néphrostomie » ou « sonde JJ »)

Résultats globaux (cohorte de 10 étudiants)

- Moyenne des scores attribués par les évaluateurs humains : 15,3 / 20
- Moyenne des scores attribués par GPT-4 : 14,4 / 20
- Écart-type des scores humains : 1,32
- Écart-type des scores GPT-4 : 1,49
- Accord exact item par item (grille sur 18 points) : 96 % (144 accords sur 150 comparaisons)
- Accord exact sur les 2 critères transversaux (synthèse + prise en charge) : 90 % (18 accords sur 20)

Discordances (en faveur de ChatGPT)

1. Item : Prescription d'antibiothérapie probabiliste

- **Évaluation humaine** : 1 - **GPT-4** : 0
- **Discours de l'étudiant** : « Il faudra faire une antibiothérapie à base d'amoxicilline. »
- **Analyse** : L'antibiotique cité (amoxicilline seule) n'est pas adapté à une pyélonéphrite grave. GPT n'accorde pas le point, car la prescription ne respecte pas les recommandations de bithérapie probabiliste en IV avec une C3G et un aminoside

2. Item : Évoquer le risque de choc septique

- **Évaluation humaine** : 1 **GPT-4** : 0
- **Discours de l'étudiant** : « C'est grave, il a une infection, il faut agir vite. »
- **Analyse** : L'évaluateur humain a parfois valorisé des formulations vagues suggérant une urgence grave. GPT-4, de son côté, cherche explicitement les termes « choc septique », « sepsis », ou un score qSOFA.

3. Item : Drainage urgent de la voie excrétrice

- **Évaluation humaine** : 1 **GPT-4** : 0
- **Discours de l'étudiant** : « Il faut appeler l'urologue pour voir s'il faut poser une sonde. »
- **Analyse** : L'intention est présente, mais pas l'expression claire du caractère urgent ou de la nécessité de drainage chirurgical. GPT ne donne pas le point, ce qui respecte strictement la consigne de l'item.

Discordances (en faveur des humains)

1. Item : Hémocultures

- Score humain : 1 / GPT-4 : 0
- Discours étudiant : « On va faire un bilan biologique complet avec une NFS, CRP, ionogramme, urée, créatinine... »
- Analyse : L'étudiant ne mentionne pas les hémocultures. Les évaluateurs humains ont peut-être valorisé l'intention générale d'un bilan en contexte infectieux grave. GPT-4 a raison de ne pas accorder le point en l'absence explicite de la prescription.

2. Item : Dérivation urinaire en urgence (caractère urgent obligatoire)

- Score humain : 2 / GPT-4 : 0
- Discours étudiant : « Il faudra appeler l'urologue pour prendre en charge le calcul... peut-être envisager une néphrostomie... »
- Analyse : L'étudiant mentionne une technique, mais sans insister sur l'urgence. Or, la grille exige que le **caractère urgent** soit explicitement exprimé. GPT-4 a donc respecté la consigne, contrairement aux évaluateurs humains.

3. Item : Laisser le patient à jeun

- Score humain : 1 / GPT-4 : 0
- Discours étudiant : « Il ne prend rien par la bouche pour le moment, on envisage une chirurgie. »
- Analyse : L'intention est présente, mais l'expression « on le garde à jeun » est absente. GPT-4 applique strictement la consigne, tandis que les humains infèrent l'intention. GPT-4 est ici plus rigoureux.

Discussion : limites et enseignements

- **Limites de l'expérimentation**

- Analyse menée sur un seul ECOS, dans un seul contexte clinique
- Petit effectif (10 étudiants), sans sélection aléatoire
- Comparaison avec une évaluation humaine souvent collégiale (donc non strictement comparable à une IA évaluatrice autonome)
- Prompt non optimisé pour chaque item de la grille (pas de calibration fine des critères implicites)
- Multimodalités non expérimentée (video, son, prosodie)

- **Enseignements clés**

- Accord global très élevé (96 % sur les items, 90 % sur les critères transversaux)
- GPT-4 tend à appliquer strictement les consignes explicites de la grille
- Les divergences surviennent surtout sur des interprétations implicites ou contextuelles (ex. “à jeun”, “urgence” non verbalisée)

Perspectives :

- **Améliorations possibles**

- Entraînement ou calibration du modèle avec des critères plus flexibles
- Intégration d'un raisonnement plus nuancé sur l'intention clinique
- Couplage avec une interface permettant à l'étudiant de consulter un feedback ciblé par item

- **Usages pédagogiques envisagés**

- **Outil d'auto-entraînement pour les étudiants**, avec retour immédiat et personnalisé
- **Création assistée d'ECOS** via LLM, incluant des stations enrichies en thématiques transversales (ex. santé numérique)

- **Etude à plus grande échelle : ANR ?**