Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Vers un raisonnement médical français : Revue des ressources d'apprentissage et d'évaluation des grands modèles de langage

Titouan Ravard[1]
(1) Nantes Université, CNRS, LS2N UMR 6004, France
titouan.ravard@etu.univ-nantes.fr

RÉSUMÉ

L'intégration des modèles de langage génératifs dans les systèmes de santé soulève des enjeux majeurs de souveraineté technologique, de protection des données et de conformité éthique. Dans ce contexte, la capacité des modèles à produire des raisonnements explicites, notamment via les chaînes de pensée (*Chain-of-Thought*, CoT), constitue un levier important pour l'interprétabilité et l'évaluation médicale. Cet article propose une revue des ressources existantes dédiées au raisonnement médical des grands modèles de langage génératifs (LLMs), en analysant à la fois les jeux de données, les méthodes de construction et de structuration des raisonnements et les stratégies d'évaluation existantes. Il met également en évidence le manque de ressources françaises et discute des implications pour le développement de modèles fiables adaptés au contexte médical français.

ABSTRACT

Toward French Medical Reasoning : A Review of Learning and Evaluation Resources for Large Language Models

The integration of generative language models into healthcare systems raises major challenges related to technological sovereignty, data protection, and ethical compliance. In this context, the ability of models to produce explicit reasoning, particularly through Chain-of-Thought (CoT), constitutes an important lever for interpretability and medical evaluation. This article presents a review of existing resources dedicated to medical reasoning for generative large language models (LLMs), analyzing datasets, methods for constructing and structuring reasoning processes, and current evaluation strategies. It also highlights the scarcity of French resources and discusses the implications for developing reliable models tailored to the French medical context.

MOTS-CLÉS : Raisonnement médical, Évaluation du raisonnement, Modèles médicaux.

KEYWORDS: Medical reasoning, Reasoning evaluation, French medical models .

1 Introduction

Si les grands modèles de langage génératifs (*Large Language Models*, ou LLMs) démontrent aujourd'hui des performances impressionnantes sur différentes tâches médicales (Shool *et al.*, 2025), leurs limites en matière de raisonnement demeurent importantes. En particulier, leurs capacités à produire des chaînes de pensée (*Chain-of-Thought*, CoT) (Wei *et al.*, 2022) nécessaires à l'explicitation de justifications restent encore limitées (Nachane *et al.*, 2024; Qiu *et al.*, 2025a). Cette absence limite à la fois l'apprentissage et l'évaluation des capacités de raisonnement des LLMs en contexte médical (Griot *et al.*, 2025a).

Actes des 19e Rencontres Jeunes Chercheurs en RI (RJCRI) et 28ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL), pages 202–218.

Par ailleurs, la majorité des jeux de données médicaux disponibles sont issus de données anglophones, ce qui restreint leur utilisation dans d'autres environnements linguistiques et médicaux (Yan *et al.*, 2025). Ces ressources reflètent en effet implicitement la culture, les pratiques pédagogiques et le cadre réglementaire propres à leur contexte d'origine. Ainsi, une simple traduction peut conduire à des interprétations ou décisions médicales inadaptées (Bazoge, 2025).

Ces constats soulèvent le besoin de concevoir et d'évaluer des ressources permettant le développement de capacités de raisonnement médical explicite [1] chez les LLMs, en particulier dans des contextes culturels non anglophones tels que la France.

Cet article présente une revue structurée des ressources existantes relatives au raisonnement médical des LLMs génératifs, avec une attention particulière portée aux enjeux français, afin d'identifier les lacunes et proposer des axes de recherche. La Section 2 analyse les différentes ressources d'apprentissage et d'évaluation existantes. La Section 3 discute des principales approches de construction et de structuration des chaînes de pensées. La Section 4 offre un aperçu des modèles et stratégies de raisonnement. La Section 5 évoque les limites des métriques actuelles pour évaluer les raisonnements produits. Enfin, la Section 6 parle des principaux défis à résoudre et des futures perspectives de recherche.

2 Ressources d'entraînement et d'évaluation existantes

Cette section analyse les principales ressources utilisées pour l'apprentissage et l'évaluation des modèles génératifs, en mettant en évidence l'évolution des ressources contenant des raisonnements structuré, ainsi que le retard des ressources (Figure 1).

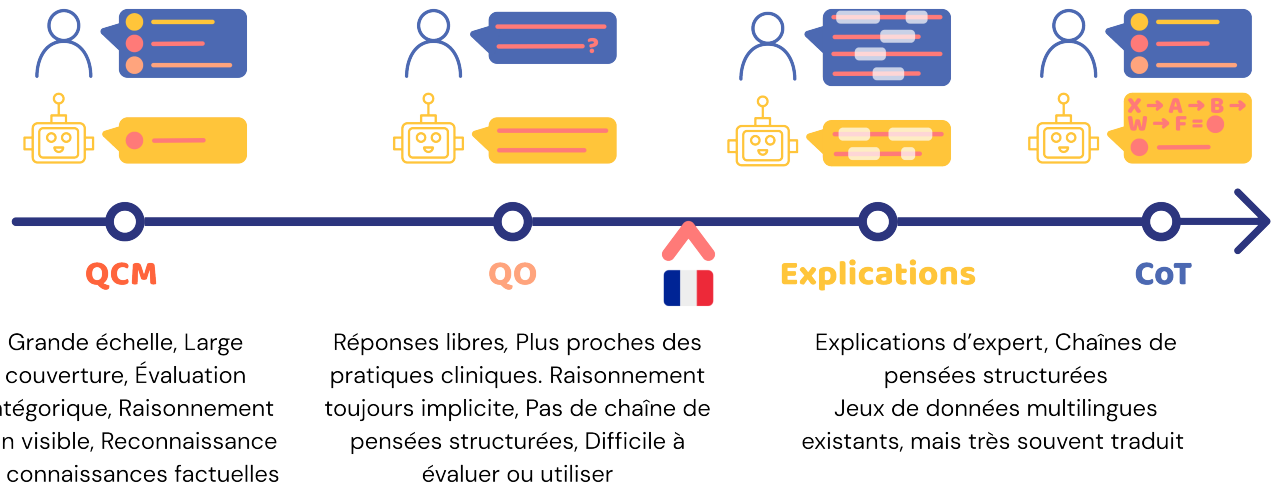

FIGURE 1 – Évolution de la complexité des ressources pour le raisonnement médical. La progression illustre le passage de l'évaluation d'un raisonnement factuel (QCM) vers des structures logiques formalisées (CoT). Le curseur indique l'état actuel des ressources natives francophones.

1. Dans la littérature, le raisonnement explicite est défini comme désignant le raisonnement généré par le modèle lors de l'inférence en même temps que la réponse. Il est à différencier du raisonnement implicite qui est vu comme le raisonnement interne du modèle pour parvenir à générer une réponse.

2.1 Premières approches : amorce par chaînes de pensée

À l'origine, les méthodes visant à induire des chaînes de pensée reposent principalement sur des stratégies d'amorçage (ou *prompting*), généralement regroupées en trois catégories.

"L'amorçage manuel" (*manual prompting*) (Wei *et al.*, 2022) consiste à fournir au modèle des démonstrations annotées servant d'exemples de raisonnement. Cette méthode offre un contrôle fin de la structure argumentative (Fu *et al.*, 2023), mais demeure coûteuse et difficilement reproductible à grande échelle, en particulier en contexte médical où l'expertise requise est élevée.

"L'amorçage automatique" (*automatic prompting*) permet d'induire un raisonnement explicite sans annotation préalable, par exemple via des instructions génériques telles que "*Let's think step by step*" (Kojima *et al.*, 2022). Bien que peu coûteuse, ces stratégies restent sensibles aux formulations employées et peuvent produire des raisonnements instables ou hallucinés (Huang *et al.*, 2025).

Enfin, les approches dites "d'amorçage semi-automatique" (*semi-automatic prompting*) combinent un nombre limité d'exemples annotés induisant une génération automatique (Shao *et al.*, 2023). Elles constituent un compromis intéressant, mais leur efficacité dépend fortement de la qualité des démonstrations initiales et des mécanismes de validation associés.

Si ces méthodes permettent d'expliciter le raisonnement à l'inférence, elles ne garantissent ni sa validité logique ni sa pertinence clinique (Wang *et al.*, 2025), limitant ainsi leur reproductibilité. Ces limites ont conduit à considérer les chaînes de pensée non plus comme une stratégie d'inférence, mais comme un objet de données à part entière. L'apprentissage supervisé à partir de ressources explicitement structurées apparaît comme une voie prometteuse pour améliorer la qualité et l'évaluation du raisonnement médical au-delà de la seule exactitude des réponses finales (Lu *et al.*, 2022).

2.2 Passage à des jeux de données de QCM médicaux

La majorité des benchmarks médicaux existants reposent sur des questions à choix multiples (QCM), en particulier issues d'examens médicaux ou de la littérature médicale. Ces benchmarks sont constitués de questions fermées accompagnées d'une liste de réponses à choisir. Des jeux de données tels que PubMedQA (Jin *et al.*, 2019), MedQA (Jin *et al.*, 2021) et MedMCQA (Pal *et al.*, 2022) constituent encore aujourd'hui des références pour l'évaluation des performances des LLMs en médecine. Des ressources plus récentes, comme MedBullets (Chen *et al.*, 2025a), Humanity Last Exam (Phan *et al.*, 2025), ou BioMed-R1-Eval (Thapa *et al.*, 2025), prolongent cette approche en proposant des questions issues d'experts, ou bien distillé de jeux existants, afin d'extraire des questions jugées plus complexes et nécessitants des capacités de raisonnement plus élevées.

Ces formats offrent des volumes importants et une évaluation standardisée, mais ils restent limités à l'exactitude de la réponse finale, sans accès explicite au raisonnement. Ils favorisent ainsi l'évaluation de connaissances factuelles ou de capacités de récupération d'information, voire de reconnaissance de motifs (Griot *et al.*, 2025b), plutôt que la construction d'un raisonnement clinique structuré.

2.3 Vers des formats plus ouverts et enrichis

Afin de dépasser ces limites, plusieurs travaux proposent des formats plus ouverts ou enrichis. CareQA (Arias-Duart *et al.*, 2025) et MORQA (wai Yim *et al.*, 2025) introduisent ainsi des questions ouvertes (QO) représentant des situations plus proches des scénarios cliniques réels. Ces questions,

contrairement au QCM, sont libres et comportent en général plusieurs réponses valides possibles. D'autres initiatives, comme MedR-Bench (Qiu *et al.*, 2025b), s'appuient sur des cas patients structurés générés à partir de rapports cliniqueVerss, couvrant par exemple la recommandation d'examens, le diagnostic ou la planification thérapeutique.

Des approches orientées *évaluation qualitative* émergent également. HealthBench (Arora *et al.*, 2025) propose une évaluation basée sur des rubriques expertes pour juger la pertinence des réponses, tandis que R2MED (Li *et al.*, 2025) explore l'évaluation du raisonnement via des tâches de recherche d'information clinique. Les initiatives d'OpenMed (OpenMed, 2025) proposent plusieurs jeux de données couvrant un large éventail de tâches médicales, majoritairement générés par des LLMs spécialisés selon les tâches considérées. Ces formats permettent une évaluation plus fine, mais restent souvent dépourvus d'annotations explicites de chaînes de pensée, utile en termes de comparaison ou lors de l'apprentissage.

2.4 Jeux de données intégrant des explications ou des chaînes de pensée

Une évolution récente consiste à associer aux réponses des explications ou des chaînes de pensée de manière explicites. MedExQA (Kim *et al.*, 2024) propose ainsi des explications annotées par des experts, tandis que ThoughtSource (Ott *et al.*, 2023) agrège plusieurs jeux médicaux enrichis en raisonnements issus d'experts ou de production participative (*crowdsourcing*).

Des travaux plus récents cherchent à structurer explicitement ces raisonnements. MedReason (Wu *et al.*, 2025) génère des chaînes de pensée à partir de graphes de connaissances médicaux, RadRBench-CXR (Fan *et al.*, 2025) exploite des rapports radiologiques réels, ReasonMed (Sun *et al.*, 2025) construit un corpus à grande échelle via un processus multi-agents incluant génération, vérification et raffinement des raisonnements, MedCOT-Bench (Ding *et al.*, 2025) amplifie des QCM médicaux avec des chaînes de pensée générées par LLM puis validées par des experts. D'autres benchmarks, comme MedThink-Bench (Zhou *et al.*, 2025a) ou MedXpertQA (Zuo *et al.*, 2025), combinent annotations expertes, multimodalité et évaluation automatique.

Malgré ces récentes avancées, ces ressources restent majoritairement de pays anglophones, parfois spécialisées sur certaines modalités, ou encore trop limitées pour l'apprentissage généralisé d'un raisonnement clinique robuste et fiable.

2.5 Et les ressources françaises ?

Les jeux de données médicaux français pour évaluer les modèles génératifs demeurent eux relativement rares (Servan *et al.*, 2025). FrenchMedMCQA (Labrak *et al.*, 2022) et MediQAl (Bazoge, 2025) constituent des contributions importantes pour l'évaluation des modèles en français. MediQAl couvre notamment plusieurs disciplines médicales et distingue explicitement des questions relevant du raisonnement.

Des initiatives comme CareMedEval (Bonzi *et al.*, 2025) proposent des évaluations sous forme de QCM ciblé sur la lecture critique d'articles scientifiques médicaux. Les corpus conversationnels tels que MedDialog-FR (Liu *et al.*, 2024) offrent également un contexte clinique riche avec des questions ouvertes, mais sans formalisation explicite du raisonnement.

Des jeux "originaux" multilingues, comme notamment MMedBench (Qiu *et al.*, 2024), incluent parfois des données en français, souvent traduite de manière automatique et généré de manière synthétique, restreignant ainsi leur utilisation concrète.

2.6 Synthèse

On observe ainsi une évolution progressive des ressources : des benchmarks QCM qui évaluent principalement les connaissances factuelles, aux formats ouverts avec des questions au contexte enrichi visant des scénarios cliniques plus réalistes, à des jeux de données intégrants des explications structurées ou des chaînes de pensée explicites.

Cependant, les ressources dédiées aux tâches médicales de manière générale restent majoritairement issues de pays anglophone et chinois, ce qui limite leur portée à d'autre pays (Yan *et al.*, 2025). L'absence de ressources francophones structurées pour l'apprentissage et l'évaluation du raisonnement médical explicite constitue ainsi un verrou important pour le développement de LLMs médicaux adaptés au contexte français. La Table 1 structure ces ressources en fonction de différents critères.

Jeu de données	**Langue**	**Tâche**	**Taille**	**Type de raisonnement**
MedBullets 2025a	EN	QCM	308	Explications cliniques expertes non structurées
HLE 2025	EN	QCM,QO	275	Questions expertes nécessitant un raisonnement avancé
BioMed-R1-Eval 2025	EN	QCM	4138	Questions distillées ciblant un raisonnement complexe
CareQA 2025	EN	QO	2769	Reformulation automatique en réponse ouverte (raisonnement implicite)
MORQA 2025	EN/CN	QO	2275	Scénarios experts augmentés par LLM
MedR-Bench 2025b	EN	QO	1453	Cas patients structurés générés par LLM
HealthBench 2025	EN	QO	5000	Évaluation multi-critères experts de raisonnements de dialogue modèle/experts
R2MED 2025	EN	QO	876	Raisonnement évalué via récupération d'évidence
OpenMed 2025	EN	QO	∼375k	Raisonnements/discussions générés par LLM (multi-modèles)
MedExQA 2024	EN	QCM	965	Annotations expertes associées aux réponses
ThoughtSource 2023	EN	QCM	165k	Collection de CoT explicites experts et LLM
MedReason 2025	EN	QO	32k	Raisonnement guidé par graphe de connaissances médical
RadRBench-CXR 2025	EN	QO	59k	Raisonnement diagnostique à partir de rapports réels
ReasonMed 2025	EN	QO	370k	CoT générées via processus multi-agent
MedCOT-Bench 2025	CN	QCM	∼30k	CoT générées par LLM puis validées par experts
MedThink-Bench 2025a	EN	QO	500	Justifications expertes structurées, évaluation automatique
MedXpertQA 2025	EN	QCM	4460	Justifications d'experts structurés (raisonnement multimodal)
MMedBench 2024	Multi.	QCM	∼53k	Questions générées par LLM à partir de QO ; traduction automatique en plusieurs langues
FrenchMedMCQA 2022	FR	QCM	3105	Pas d'explicitation du raisonnement
MediQAl 2025	FR	QCM,QO	32k	Annotation automatique du niveau de raisonnement
CareMedEval 2025	FR	QCM	534	Lecture critique d'article nécessitant raisonnement analytique
MedDialog-FR 2024	FR	QO	15k	Dialogues traduits automatiquement et annotés

TABLE 1 – Synthèse comparative des ressources ouvertes intégrant différentes formes d'explicitation du raisonnement médical. La table met en évidence les approches de supervision par LLM avec un passage à l'échelle beaucoup plus grand, ainsi que le déplacement progressif du raisonnement d'une compétence implicite vers un objet de données structuré. Elle souligne également la forte domination des ressources de culture anglophone et le manque de ressources françaises pour le raisonnement explicite.

3 Structuration du raisonnement dans ces ressources

Après avoir évoqué les ressources suivant leur évolution en termes de format, cette section propose une analyse des principales approches de construction et de structuration des chaînes de pensées dans ces ressources, allant des méthodes entièrement basées sur des experts vers des approches totalement supervisées par des modèles génératifs.

3.1 Annotation experte de raisonnements

Une première approche consiste à annoter explicitement des raisonnements par des experts humains. Les ressources comme MedExQA (Kim *et al.*, 2024) ou MedThink-Bench (Zhou *et al.*, 2025a) fournissent ainsi des justifications textuelles structurées associées aux réponses correctes. De même, certaines ressources issues d'examens médicaux comme MedBullets (Chen *et al.*, 2025a), incluent des discussions explicatives rédigées par des spécialistes.

Ces annotations offrent généralement un haut niveau de fiabilité clinique et reflètent les pratiques professionnelles. Toutefois, la variabilité inter-annotateurs et les contraintes d'accès aux experts limitent leur passage à l'échelle, comme le montre la taille de ces ressources en Table 1.

3.2 Extraction à partir de données cliniques ou scientifiques

Une autre stratégie consiste à extraire des raisonnements à partir de traces existantes, comme des rapports cliniques ou des publications scientifiques. RadRBench-CXR (Fan *et al.*, 2025) exploite par exemple des comptes rendus radiologiques réels pour reconstruire des chaînes diagnostiques, tandis que CareMedEval (Bonzi *et al.*, 2025) s'appuie sur des exercices de lecture critique d'articles médicaux.

Ces approches permettent d'ancrer les chaîne de pensée dans des pratiques réelles, mais elles dépendent fortement de l'accès à des données parfois sensibles, et de leur niveau de structuration. Les contraintes réglementaires européennes comme le RGPD peuvent constituer un frein important à la diffusion de telles ressources au regard des droits d'auteur ou de la protection des données.

3.3 Génération synthétique par modèles de langage

Face aux limites de l'annotation experte, de nombreux travaux reposent sur la génération automatique de chaînes de pensée par des LLMs. MMedBench (Qiu *et al.*, 2024) ou ThoughtSource (Ott *et al.*, 2023) enrichissent ainsi des jeux de questions-réponses existants avec des raisonnements générés automatiquement. Des initiatives plus ambitieuses, comme ReasonMed (Sun *et al.*, 2025) ou les corpus d'OpenMed (OpenMed, 2025), utilisent des pipelines multi-agents pour générer, filtrer et raffiner des chaînes de pensée à grande échelle.

Ces approches permettent de constituer rapidement de larges corpus, mais elles introduisent un risque d'hallucination ou de biais liés aux modèles générateurs (Huang *et al.*, 2025). Plusieurs travaux proposent donc des validations expertes a posteriori, comme MedCOT-Bench (Ding *et al.*, 2025), afin d'améliorer la qualité clinique des raisonnements.

3.4 Approches hybrides et structurées

Certaines ressources combinent génération automatique, validation humaine et structuration explicite du raisonnement. MedReason (Wu *et al.*, 2025) s'appuie sur des graphes de connaissances médicaux pour produire des chaînes de pensée structurées, tandis que MedR-Bench (Qiu *et al.*, 2025b) génère des cas patients synthétiques à partir de rapports cliniques réels. D'autres benchmarks, comme R2MED (Li *et al.*, 2025), considèrent les étapes de recherche d'information clinique comme une forme de raisonnement observable.

Ces approches hybrides visent un compromis entre passage à l'échelle et qualité clinique, mais restent

dépendantes des ressources de connaissance disponibles et des procédures de validation mises en œuvre.

3.5 Implications pour la construction de ressources françaises

Globalement, la littérature montre une transition progressive d'annotations expertes coûteuses vers des pipelines hybrides, combinant génération automatique et validation par des sources humaines. Cependant, la majorité de ces ressources restent dépendantes de leur pays d'origine et sont plus difficilement adaptables aux spécificités linguistiques et réglementaires du contexte médical français.

La constitution de jeux de données français intégrant des chaînes de pensée fiables constitue ainsi un enjeu clé pour le développement de LLMs médicaux interprétables et utilisables dans les systèmes de santé français.

4 Modèles génératifs entraînés pour le raisonnement : évolution et stratégies

L'amélioration des capacités de raisonnement des modèles génératifs repose sur l'évolution conjointe des architectures, des stratégies d'entraînement et des ressources mobilisées. Cette section retrace l'évolution de grands modèles fortement supervisés, vers des approches plus économes en données exploitant davantage l'inférence pour parvenir à un raisonnement médical explicite.

4.1 Premiers modèles spécialisés : raisonnement à grande échelle

Les premiers modèles médicaux orientés raisonnements reposent principalement sur de l'ajustement supervisé (*fine-tuning*) à partir d'annotations d'experts. Med-PaLM 2 (Singhal *et al.*, 2023) constitue une référence, combinant *fine-tuning* et apprentissage par renforcement avec un retour humain pour produire des réponses médicales justifiées.

Des modèles plus récents cherchent à expliciter davantage les trajectoires de raisonnement. HuatuoGPT-o1 (Chen *et al.*, 2024) combine ainsi génération automatique de raisonnements, vérification par LLM et apprentissage avec retour d'IA. FineMedLM-o1 (Yu *et al.*, 2025) adopte un apprentissage progressif, allant de connaissances factuelles à des raisonnements causaux complexes.

4.2 Contraindre ou structurer le raisonnement

Une seconde direction consiste à contraindre explicitement la structure du raisonnement produit. MedReason (Wu *et al.*, 2025) impose par exemple des trajectoires compatibles avec un graphe de connaissances médicales, tandis que DiagnosisGPT (Chen *et al.*, 2025b) formalise une chaîne diagnostique reflétant un raisonnement clinique réel.

D'autres travaux explorent des stratégies de sélection ou d'évaluation du raisonnement plutôt que sa génération directe. MedAdapter (Shi *et al.*, 2024) entraîne un adaptateur léger chargé de classer plusieurs raisonnements candidats produits par un LLM. ClinRaGen (Niu *et al.*, 2025) étend cette idée à des contextes multimodaux, intégrant dossiers médicaux électroniques et paramètres physiologiques.

4.3 Auto-amélioration et génération synthétique de raisonnement

De nombreux travaux explorent la génération synthétique de données de raisonnement afin de leur servir de base d'entraînement. BioMed-R1 (Thapa *et al.*, 2025) et UltraMedical (Jiang *et al.*, 2025) exploitent des pipelines de génération et filtrage automatique pour sélectionner des exemples nécessitant un raisonnement complexe, complétés par des stratégies de recherche du meilleur raisonnement durant l'inférence. ReasonMed (Sun *et al.*, 2025) illustre également cette tendance avec un apprentissage supervisé sur de larges volumes de chaînes de pensée générées puis raffinées automatiquement.

4.4 Apports des modèles généralistes de raisonnement

L'arrivée de modèles généralistes de raisonnement plus compacts comme les différents modèles DeepSeek-R1-distill (DeepSeek-AI, 2025), QwQ (Team, 2025) ou encore Kimi K2 Thinking (Kimi-AI, 2025) suggèrent l'efficacité de stratégies de renforcement, distillation ou raisonnement agentique, semblable à de plus grands modèles comme o1 (OpenAI, 2024). Magistral (Mistral-AI, 2025) met notamment en évidence l'importance de stratégies de post-entraînement et de l'adaptation linguistique, notamment pour des contextes multilingues ou francophones. Ces modèles suggèrent ainsi qu'un *fine-tuning* ciblé sur des données de qualité peut suffire à adapter efficacement des modèles généralistes à un domaine spécialisé.

4.5 Vers des modèles plus compacts et économes en données

Des travaux comme m1 (Huang *et al.*, 2026) montrent qu'un ajustement de la taille du raisonnement produit à l'inférence peut améliorer les performances sans supervision lourde. En mathématique, LIMO (Ye *et al.*, 2025) ou Pensez (Hoang Ha, 2025) mettent en évidence qu'un petit nombre d'exemples de raisonnement soigneusement sélectionnés peut suffire à obtenir de bonnes capacités ainsi qu'une bonne généralisation. Ces travaux suggèrent alors une voie viable pour créer des modèles de raisonnement plus légers, plus efficaces, mais néanmoins très performants (Wang *et al.*, 2025).

5 Évaluation du raisonnement médical

L'évaluation des capacités de raisonnement des modèles de langage constitue un enjeu central pour leur intégration sûre dans les systèmes de santé. Deux grandes approches dominent actuellement la littérature : l'évaluation implicite du raisonnement fondée sur la réponse finale uniquement ; et l'évaluation du raisonnement explicite produit par le modèle.

5.1 Évaluation du raisonnement implicite par la réponse finale

La majorité des travaux évaluent le raisonnement des modèles sous forme implicite sur des tâches objectives comme les QCM médicaux en mesurant la justesse de leur prédiction. Ces approches sont ainsi largement adoptées dans les benchmarks médicaux, car fournit une mesure simple et reproductible des performances du modèle (Chen *et al.*, 2025a; Phan *et al.*, 2025; Thapa *et al.*, 2025).

Cependant, ces approches sont limitées, car elles n'évaluent que la validité de la réponse finale donnée et non la qualité du raisonnement du modèle ayant conduit à cette réponse. Un modèle peut donc produire une réponse correcte en ayant eu un raisonnement erroné ou incomplet, rendant discutable la

pertinence de la réponse en contexte médical (Griot *et al.*, 2025a). Ce type d'évaluation est également critiqué, car sujette à des biais quant au choix des réponses en fonction de leur distribution (Zheng *et al.*, 2024) leur ordre (Pezeshkpour & Hruschka, 2024) ou leur structuration (Yang *et al.*, 2025).

5.2 Évaluation du raisonnement explicite produit par le modèle

Afin d'évaluer directement le raisonnement produit, plusieurs approches ont recours à des métriques automatiques de similarité textuelle telles que BLEU (Papineni *et al.*, 2002), ROUGE (Lin, 2004) ou METEOR (Banerjee & Lavie, 2005), mesurant le recouvrement lexical à partir de références données. Des métriques sémantiques comme BERTScore (Zhang *et al.*, 2020) sont souvent utilisées pour tenter également de capturer une proximité au-delà de la simple surface textuelle.

Toutefois, ces métriques sont peu robustes aux paraphrases médicalement valides et peinent à détecter des erreurs factuelles subtiles et significatives en contexte médical (Abbasian *et al.*, 2024). Plusieurs études montrent également une corrélation limitée entre ces métriques automatiques et une évaluation humaine dans des tâches génératives complexes (Zhou *et al.*, 2025b; wai Yim *et al.*, 2025).

Ces difficultés sont amplifiées dans le domaine médical, caractérisé par une terminologie spécialisée, des nuances conceptuelles délicates et des dépendances fortes au contexte clinique (Berger *et al.*, 2025).

5.3 Entre expertise humaine et évaluation automatisée

Face à ces limites, l'évaluation experte reste le standard de référence. Des initiatives telles que HealthBench (Arora *et al.*, 2025) s'appuient sur des grilles d'évaluation définies par des médecins pour juger la validité clinique, la cohérence logique et la pertinence des réponses. Néanmoins, ce type d'approche reste coûteuse et difficilement reproductible à grande échelle.

Dans ce contexte, des approches utilisant des modèles juges (dite *LLM-as-a-Judge*) constituent une alternative à des experts humains (Zheng *et al.*, 2023). Un modèle de langage distinct est chargé d'évaluer la qualité du raisonnement selon des critères explicitement définis (validité clinique, cohérence des étapes, complétude, absence d'erreurs critiques) (Zhou *et al.*, 2025b). Cette méthode offre ainsi un meilleur passage à l'échelle et peut intégrer des connaissances médicales implicites (Gu *et al.*, 2025). Toutefois, elle reste exposée aux biais du modèle évaluateur, aux phénomènes d'hallucination et peine à évaluer des catégories plus subjectives telles que l'empathie, l'évaluation des risques et le raisonnement clinique contextuel (Diekmann *et al.*, 2025).

Ainsi, l'évaluation du raisonnement médical explicite demeure un problème ouvert. Elle nécessite des méthodes hybrides combinant métriques automatiques, validation experte ciblée et méthodes d'évaluation assistées par LLM, afin de concilier passage à l'échelle et exigence clinique.

6 Discussion, défis et futurs axes de recherche

L'analyse des ressources existantes, des stratégies de génération de chaînes de pensée et des modèles orientés raisonnement met en évidence plusieurs défis scientifiques majeurs pour le développement de LLMs médicaux souverains, fiables et interprétables.

6.1 Disponibilité et qualité des ressources de raisonnement

Un premier défi concerne la constitution de ressources de raisonnement médical explicite. Si les approches récentes permettent de générer à grande échelle des chaînes de pensée synthétiques, la qualité clinique et la fidélité factuelle de ces raisonnements restent variables. Les annotations expertes constituent le standard de référence, mais leur coût limite leur extensibilité. La tension entre passage à l'échelle et fiabilité clinique demeure donc centrale. De futurs travaux devront explorer des protocoles hybrides combinants génération automatique, validation experte ciblée et mécanismes de contrôle qualité systématiques.

6.2 Évaluation du raisonnement : un problème encore ouvert

L'évaluation du raisonnement explicite reste un problème méthodologique non résolu. Les métriques objectives basées sur la réponse finale ne capturent pas la qualité du processus décisionnel, tandis que les métriques de similarité textuelle sont insuffisantes pour refléter la validité clinique. L'évaluation humaine demeure indispensable mais difficilement reproductible à grande échelle, et les approches de type *LLM-as-a-Judge* soulèvent des questions de biais et de robustesse. Dans ce contexte, l'intégration de mesures d'interprétabilité et d'explicabilité, s'appuyant sur l'analyse de l'attention ou sur des méthodes d'influence (Wen *et al.*, 2025), permettrait d'analyser les réponses générées au regard du contexte fourni ou des données d'entraînement. Le développement de protocoles d'évaluation hybrides, combinant critères cliniques explicites et méthodes automatiques fiables, constitue ainsi un axe important pour la communauté.

6.3 Enjeux culturels et adaptation au contexte français

La majorité des ressources et des modèles étudiés restent de pays anglophones. Or, le raisonnement médical est fortement dépendant du contexte culturel, réglementaire et pédagogique. Une simple traduction de jeux de données existants ne garantit ni l'adéquation clinique ni la validité pédagogique des raisonnements produits (Genovese *et al.*, 2024; Vera, 2026). En effet, au-delà de la barrière linguistique, un modèle nativement francophone doit entre autre intégrer les spécificités de la sémiologie médicale française, les cadres éthiques et juridiques locaux, ainsi que l'organisation propre au parcours de soin français. Le développement de ressources françaises intégrant des chaînes de pensée explicites ancrées dans les référentiels académiques nationaux représente ainsi un enjeu, tant pour la souveraineté technologique, que pour l'adaptation aux pratiques cliniques locales.

6.4 Vers des modèles de raisonnements médicaux plus efficaces

Enfin, l'évolution récente des modèles montre qu'il est possible d'obtenir des capacités de raisonnement avancées à partir de volumes de données plus restreints, à condition de disposer d'exemples de haute qualité et de stratégies d'entraînement adaptées. On voit alors trois directions complémentaires émerger : (i) adapter des modèles généralistes de raisonnement via un *fine-tuning* ciblé sur des données médicales structurées (Mistral-AI, 2025) ; (ii) renforcer des modèles médicaux compacts à l'aide d'un nombre limité d'exemples de raisonnement soigneusement sélectionnés (Wang *et al.*, 2025) ; et (iii) exploiter des architectures multi-agents où un modèle orchestre différents agents experts spécialisés, simulant ainsi le raisonnement clinique collaboratif entre différents spécialistes (You *et al.*, 2026). Ces approches ouvrent la voie à des systèmes nécessitant moins de données et potentiellement plus adaptés aux contraintes des systèmes de santé.

6.5 De la performance technique à l'impact clinique

Au-delà de leurs performances techniques, l'intégration du raisonnement explicite dans les modèles médicaux soulève la question de l'usage réel et de la sécurité de ces outils en pratique clinique. En effet, la production d'un raisonnement structuré, s'il est erroné, peut induire un biais d'automatisation où le clinicien accorde une confiance excessive à l'outil au détriment de sa propre analyse (SKITKA *et al.*, 2000). Ce phénomène rendrait l'outil potentiellement plus néfaste qu'un modèle ne fournissant aucune justification. Dès lors, l'impact clinique de tels modèles et des différentes approches actuelles du traitement du langage appliquées à la médecine constitue un enjeu central et devrait être envisagé comme la finalité principal de tels outils. Une analyse des besoins des praticiens et une évaluation de l'applicabilité de ces modèles en environnement hospitalier sont nécessaires pour s'assurer que l'explicabilité soutient la décision médicale sans introduire de nouveaux risques pour la sécurité des soins.

6.6 Conclusion

Le raisonnement médical explicite représente aujourd'hui un enjeu central pour l'intégration responsable des LLMs en santé. Si des progrès significatifs ont été réalisés dans la génération et la structuration des chaînes de pensée, des défis subsistent quant à leur fiabilité, leur évaluation et leur adaptation culturelles. La convergence entre ressources de qualité, stratégies d'apprentissage efficaces et protocoles d'évaluation robustes apparaît comme une condition nécessaire au développement de modèles médicaux interprétables, sûrs et adaptés au contexte médical français.

Remerciements

Je remercie Richard Dufour, Solen Quiniou et Ikram Belmadani pour leurs retours et leur encadrement. Merci également à petit canard, Yannis Chupin, pour ses relectures et conseils. Je remercie également toute l'équipe TALN du LS2N pour l'ambiance et le magnifique cadre de travail qu'elle apporte.

Références

ABBASIAN M., KHATIBI E., AZIMI I., ONIANI D., SHAKERI HOSSEIN ABAD Z., THIEME A., SRIRAM R., YANG Z., WANG Y., LIN B., GEVAERT O., LI L.-J., JAIN R. & RAHMANI A. M. (2024). Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. *npj Digital Medicine*, **7**(1), 82. DOI : 10.1038/s41746-024-01074-z.

ARIAS-DUART A., MARTIN-TORRES P. A., HINJOS D., BERNABEU-PEREZ P., GANZABAL L. U., MALLO M. G., GURURAJAN A. K., LOPEZ-CUENA E., ALVAREZ-NAPAGAO S. & GARCIA-GASULLA D. (2025). Automatic evaluation of healthcare llms beyond question-answering. DOI : 10.48550/arXiv.2502.06666.

ARORA R. K., WEI J., HICKS R. S., BOWMAN P., QUIÑONERO-CANDELA J., TSIMPOURLAS F., SHARMAN M., SHAH M., VALLONE A., BEUTEL A., HEIDECKE J. & SINGHAL K. (2025). HealthBench : Evaluating Large Language Models Towards Improved Human Health. arXiv :2505.08775 [cs], DOI : 10.48550/arXiv.2505.08775.

BANERJEE S. & LAVIE A. (2005). METEOR : An automatic metric for MT evaluation with improved correlation with human judgments. In J. GOLDSTEIN, A. LAVIE, C.-Y. LIN & C. VOSS, Éds., *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, p. 65–72, Ann Arbor, Michigan : Association for Computational Linguistics.

BAZOGE A. (2025). Mediqal : A french medical question answering dataset for knowledge and reasoning evaluation. DOI : 10.48550/arXiv.2507.20917.

BERGER A., KHANNA S., BERGHAUS D. & SIFA R. (2025). Reasoning LLMs in the Medical Domain : A Literature Survey. arXiv :2508.19097 [cs], DOI : 10.48550/arXiv.2508.19097.

BONZI D., GUIGGI A., BÉCHET F., RAMISCH C. & FAVRE B. (2025). CareMedEval dataset : Evaluating Critical Appraisal and Reasoning in the Biomedical Field. arXiv :2511.03441 [cs], DOI : 10.48550/arXiv.2511.03441.

CHEN H., FANG Z., SINGLA Y. & DREDZE M. (2025a). Benchmarking Large Language Models on Answering and Explaining Challenging Medical Questions. In L. CHIRUZZO, A. RITTER & L. WANG, Éds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers)*, p. 3563–3599, Albuquerque, New Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2025.naacl-long.182.

CHEN J., CAI Z., JI K., WANG X., LIU W., WANG R., HOU J. & WANG B. (2024). Huatuogpt-o1, towards medical complex reasoning with llms. DOI : 10.48550/arXiv.2412.18925.

CHEN J., GUI C., GAO A., JI K., WANG X., WAN X. & WANG B. (2025b). CoD, towards an interpretable medical agent using chain of diagnosis. In W. CHE, J. NABENDE, E. SHUTOVA & M. T. PILEHVAR, Éds., *Findings of the Association for Computational Linguistics : ACL 2025*, p. 14345–14368, Vienna, Austria : Association for Computational Linguistics. DOI : 10.18653/v1/2025.findings-acl.740.

DEEPSEEK-AI (2025). Deepseek-r1 : Incentivizing reasoning capability in llms via reinforcement learning. DOI : 10.48550/arXiv.2501.12948.

DIEKMANN Y., FENSORE C., CARRILLO-LARCO R., CASTEJON ROSALES E., SHIROMANI S., PAI R., SHAH M. & HO J. (2025). LLMs as Medical Safety Judges : Evaluating Alignment with Human Annotation in Patient-Facing QA. In D. DEMNER-FUSHMAN, S. ANANIADOU, M. MIWA & J. TSUJII, Éds., *Proceedings of the 24th Workshop on Biomedical Language Processing*, p. 217–224, Viena, Austria : Association for Computational Linguistics. DOI : 10.18653/v1/2025.bionlp-1.19.

DING C., BIAN M., CHEN P., ZHANG H., LI T., LIU L., CHEN J., LI Z., ZHONG Y., LIU Y., HUANG H., SHAN D., HE J. & XU J. (2025). Building a Human-Verified Clinical Reasoning Dataset via a Human LLM Hybrid Pipeline for Trustworthy Medical AI. arXiv :2505.06912 [cs], DOI : 10.48550/arXiv.2505.06912.

FAN Z., LIANG C., WU C., ZHANG Y., WANG Y. & XIE W. (2025). Chestx-reasoner : Advancing radiology foundation models with reasoning through step-by-step verification. DOI : 10.48550/arXiv.2504.20930.

FU Y., PENG H., SABHARWAL A., CLARK P. & KHOT T. (2023). Complexity-based prompting for multi-step reasoning. DOI : 10.48550/arXiv.2210.00720.

GENOVESE A., BORNA S., GOMEZ-CABELLO C. A., HAIDER S. A., PRABHA S., FORTE A. J. & VEENSTRA B. R. (2024). Artificial intelligence in clinical settings : a systematic review of its role in language translation and interpretation. *Annals of Translational Medicine*, **12**(6), 117. DOI : 10.21037/atm-24-162.

GRIOT M., HEMPTINNE C., VANDERDONCKT J. & YUKSEL D. (2025a). Large language models lack essential metacognition for reliable medical reasoning. *Nature Communications*, **16**. DOI : 10.1038/s41467-024-55628-6.

GRIOT M., VANDERDONCKT J., YUKSEL D. & HEMPTINNE C. (2025b). Pattern recognition or medical knowledge ? the problem with multiple-choice questions in medicine. In W. CHE, J. NABENDE, E. SHUTOVA & M. T. PILEHVAR, Éds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 5321–5341, Vienna, Austria : Association for Computational Linguistics. DOI : 10.18653/v1/2025.acl-long.266.

GU J., JIANG X., SHI Z., TAN H., ZHAI X., XU C., LI W., SHEN Y., MA S., LIU H., WANG S., ZHANG K., WANG Y., GAO W., NI L. & GUO J. (2025). A survey on llm-as-a-judge. DOI : 10.48550/arXiv.2411.15594.

HOANG HA H. (2025). Pensez : Moins de données, meilleur raisonnement – repenser les LLM français. In F. BECHET, A.-G. CHIFU, K. PINEL-SAUVAGNAT, B. FAVRE, E. MAES & D. NURBAKOVA, Éds., *Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux*, p. 573–598, Marseille, France : ATALA \\& ARIA.

HUANG L., YU W., MA W., ZHONG W., FENG Z., WANG H., CHEN Q., PENG W., FENG X., QIN B. & LIU T. (2025). A survey on hallucination in large language models : Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, **43**(2), 1–55. DOI : 10.1145/3703155.

HUANG X., WU J., LIU H., TANG X. & ZHOU Y. (2026). m1 : Unleash the potential of test-time scaling for medical reasoning with large language models. DOI : 10.48550/arXiv.2504.00869.

JIANG S., LIAO Y., CHEN Z., ZHANG Y., WANG Y. & WANG Y. (2025). Meds^3 : Towards medical slow thinking with self-evolved soft dual-sided process supervision. DOI : 10.48550/arXiv.2501.12051.

JIN D., PAN E., OUFATTOLE N., WENG W.-H., FANG H. & SZOLOVITS P. (2021). What disease does this patient have ? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, **11**(14). DOI : 10.3390/app11146421.

JIN Q., DHINGRA B., LIU Z., COHEN W. & LU X. (2019). PubMedQA : A dataset for biomedical research question answering. In K. INUI, J. JIANG, V. NG & X. WAN, Éds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 2567–2577, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1259.

KIM Y., WU J., ABDULLE Y. & WU H. (2024). MedExQA : Medical Question Answering Benchmark with Multiple Explanations. arXiv :2406.06331 [cs], DOI : 10.48550/arXiv.2406.06331.

KIMI-AI (2025). Introducing kimi k2 thinking. https://huggingface.co/moonshotai/Kimi-K2-Thinking.

KOJIMA T., GU S. S., REID M., MATSUO Y. & IWASAWA Y. (2022). Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA : Curran Associates Inc.

LABRAK Y., BAZOGE A., DUFOUR R., DAILLE B., GOURRAUD P.-A., MORIN E. & ROUVIER M. (2022). FrenchMedMCQA : A French multiple-choice question answering dataset for medical domain. In A. LAVELLI, E. HOLDERNESS, A. JIMENO YEPES, A.-L. MINARD, J. PUSTEJOVSKY & F. RINALDI, Éds., *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, p. 41–46, Abu Dhabi, United Arab Emirates (Hybrid) : Association for Computational Linguistics. DOI : 10.18653/v1/2022.louhi-1.5.

LI L., ZHOU X. & LIU Z. (2025). R2MED : A Benchmark for Reasoning-Driven Medical Retrieval. arXiv :2505.14558 [cs], DOI : 10.48550/arXiv.2505.14558.

LIN C.-Y. (2004). ROUGE : A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, p. 74–81, Barcelona, Spain : Association for Computational Linguistics.

LIU X., SEGONNE V., MANNION A., SCHWAB D., GOEURIOT L. & PORTET F. (2024). MedDialog-FR : A French version of the MedDialog corpus for multi-label classification and response generation related to women's intimate health. In D. DEMNER-FUSHMAN, S. ANANIADOU, P. THOMPSON & B. ONDOV, Éds., *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, p. 173–183, Torino, Italia : ELRA and ICCL.

LU P., MISHRA S., XIA T., QIU L., CHANG K.-W., ZHU S.-C., TAFJORD O., CLARK P. & KALYAN A. (2022). Learn to explain : Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, **35**, 2507–2521.

MISTRAL-AI (2025). Magistral. DOI : 10.48550/arXiv.2506.10910.

NACHANE S. S., GRAMOPADHYE O., CHANDA P., RAMAKRISHNAN G., JADHAV K. S., NANDWANI Y., RAGHU D. & JOSHI S. (2024). Few shot chain-of-thought driven reasoning to prompt llms for open ended medical question answering. DOI : 10.48550/arXiv.2403.04890.

NIU S., MA J., LIN H., BAI L., WANG Z., XU Y., SONG Y. & YANG X. (2025). Knowledge-augmented multimodal clinical rationale generation for disease diagnosis with small language models. In W. CHE, J. NABENDE, E. SHUTOVA & M. T. PILEHVAR, Éds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 11011–11024, Vienna, Austria : Association for Computational Linguistics. DOI : 10.18653/v1/2025.acl-long.540.

OPENAI (2024). Openai o1 system card. DOI : 10.48550/arXiv.2403.04890.

OPENMED (2025). Medical-reasoning. https://huggingface.co/datasets/OpenMed/.

OTT S., HEBENSTREIT K., LIÉVIN V., HOTHER C. E., MORADI M., MAYRHAUSER M., PRAAS R., WINTHER O. & SAMWALD M. (2023). ThoughtSource : A central hub for large language model reasoning data. *Scientific Data*, **10**(1), 528. DOI : 10.1038/s41597-023-02433-3.

PAL A., UMAPATHI L. K. & SANKARASUBBU M. (2022). Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. In G. FLORES, G. H. CHEN, T. POLLARD, J. C. HO & T. NAUMANN, Éds., *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 de *Proceedings of Machine Learning Research*, p. 248–260 : PMLR.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, p. 311–318, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073135.

PEZESHKPOUR P. & HRUSCHKA E. (2024). Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. In *Findings of the Association for Computational Linguistics : NAACL 2024*, p. 2006–2017, Mexico City, Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2024.findings-naacl.130.

PHAN L., GATTI A., HAN Z., LI N. & HU J. E. A. (2025). Humanity's Last Exam. arXiv :2501.14249 [cs], DOI : 10.48550/arXiv.2501.14249.

QIU P., WU C., LIU S., FAN Y., ZHAO W., CHEN Z., GU H., PENG C., ZHANG Y., WANG Y. & XIE W. (2025a). Quantifying the reasoning abilities of llms on clinical cases. *Nature Communications*, **16**. DOI : 10.1038/s41467-025-64769-1.

QIU P., WU C., LIU S., ZHAO W., CHEN Z., GU H., PENG C., ZHANG Y., WANG Y. & XIE W. (2025b). Quantifying the Reasoning Abilities of LLMs on Real-world Clinical Cases. arXiv :2503.04691 [cs], DOI : 10.48550/arXiv.2503.04691.

QIU P., WU C., ZHANG X., LIN W., WANG H., ZHANG Y., WANG Y. & XIE W. (2024). Towards building multilingual language model for medicine. DOI : 10.48550/arXiv.2402.13963.

SERVAN C., GROUIN C., NÉVÉOL A. & ZWEIGENBAUM P. (2025). Comment évaluer un grand modèle de langue dans le domaine médical en français ? In F. BECHET, A.-G. CHIFU, K. PINEL-SAUVAGNAT, B. FAVRE, E. MAES & D. NURBAKOVA, Éds., *Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*, p. 51–67, Marseille, France : ATALA \\& ARIA.

SHAO Z., GONG Y., SHEN Y., HUANG M., DUAN N. & CHEN W. (2023). Synthetic prompting : generating chain-of-thought demonstrations for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23 : JMLR.org.

SHI W., XU R., ZHUANG Y., YU Y., SUN H., WU H., YANG C. & WANG M. D. (2024). MedAdapter : Efficient test-time adaptation of large language models towards medical reasoning. In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, Éds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 22294–22314, Miami, Florida, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2024.emnlp-main.1244.

SHOOL S., ADIMI S., SABOORI AMLESHI R., BITARAF E., GOLPIRA R. & TARA M. (2025). A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Medical Informatics and Decision Making*, **25**(1), 117. DOI : 10.1186/s12911-025-02954-4.

SINGHAL K., TU T., GOTTWEIS J., SAYRES R., WULCZYN E., HOU L., CLARK K., PFOHL S., COLE-LEWIS H., NEAL D., SCHAEKERMANN M., WANG A., AMIN M., LACHGAR S., MANSFIELD P., PRAKASH S., GREEN B., DOMINOWSKA E., Y ARCAS B. A., TOMASEV N., LIU Y., WONG R., SEMTURS C., MAHDAVI S. S., BARRAL J., WEBSTER D., CORRADO G. S., MATIAS Y., AZIZI S., KARTHIKESALINGAM A. & NATARAJAN V. (2023). Towards expert-level medical question answering with large language models. DOI : 10.48550/arXiv.2305.09617.

SKITKA L. J., MOSIER K. & BURDICK M. D. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies*, **52**(4), 701–717. DOI : https ://doi.org/10.1006/ijhc.1999.0349.

SUN Y., QIAN X., XU W., ZHANG H., XIAO C., LI L., ZHAO D., HUANG W., XU T., BAI Q. & RONG Y. (2025). ReasonMed : A 370K Multi-Agent Generated Dataset for Advancing Medical Reasoning. In C. CHRISTODOULOPOULOS, T. CHAKRABORTY, C. ROSE & V. PENG, Éds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, p. 26446–26467, Suzhou, China : Association for Computational Linguistics. DOI : 10.18653/v1/2025.emnlp-main.1344.

TEAM Q. (2025). Qwq-32b : Embracing the power of reinforcement learning. https://qwenlm.github.io/blog/qwq-32b/.

THAPA R., WU Q., WU K., ZHANG H., ZHANG A., WU E., YE H., BEDI S., ARESH N., BOEN J., REDDY S., ATHIWARATKUN B., SONG S. L. & ZOU J. (2025). Disentangling Reasoning and Knowledge in Medical Large Language Models. arXiv :2505.11462 [cs], DOI : 10.48550/arXiv.2505.11462.

VERA A. L. (2026). Ethical Risks and Structural Implications of AI-Mediated Medical Interpreting - AUtomatic translation biases. *JMIR AI*, **5**(1), e88651. DOI : 10.2196/88651.

WAI YIM W., ABACHA A. B., YU Z., DOERNING R., XIA F. & YETISGEN M. (2025). Morqa : Benchmarking evaluation metrics for medical open-ended question answering. DOI : 10.48550/arXiv.2509.12405.

WANG W., MA Z., DING M., ZHENG S., LIU S., LIU J., JI J., CHEN W., LI X., SHEN L. & YUAN Y. (2025). Medical reasoning in the era of llms : A systematic review of enhancement techniques and applications. DOI : 10.48550/arXiv.2508.00669.

WEI J., WANG X., SCHUURMANS D., BOSMA M., ICHTER B., XIA F., CHI E. H., LE Q. V. & ZHOU D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA : Curran Associates Inc.

WEN F., LU X., YU H., LU C., LI H. & SHI X. (2025). Dynamic asymmetric attention for enhanced reasoning and interpretability in llms. *Symmetry*, **17**(8). DOI : 10.3390/sym17081303.

WU J., DENG W., LI X., LIU S., MI T., PENG Y., XU Z., LIU Y., CHO H., CHOI C.-I., CAO Y., REN H., LI X., LI X. & ZHOU Y. (2025). Medreason : Eliciting factual medical reasoning steps in llms via knowledge graphs. DOI : 10.48550/arXiv.2504.00993.

YAN L. K. Q., NIU Q., LI M., ZHANG Y., YIN C. H., FEI C., PENG B., BI Z., FENG P., CHEN K., WANG T., WANG Y., CHEN S., LIU M., LIU J., SONG X., BAO R., JIANG Z. & QIN Z. (2025). Large language model benchmarks in medical tasks. DOI : 10.48550/arXiv.2410.21348.

YANG Z., JIAN P. & LI C. (2025). Option Symbol Matters : Investigating and Mitigating Multiple-Choice Option Symbol Bias of Large Language Models. In L. CHIRUZZO, A. RITTER & L. WANG, Éds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers)*, p. 1902–1917, Albuquerque, New Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2025.naacl-long.95.

YE Y., HUANG Z., XIAO Y., CHERN E., XIA S. & LIU P. (2025). Limo : Less is more for reasoning. DOI : 10.48550/arXiv.2502.03387.

YOU Z., CHEN X., VASHISHTHA A., DU S., ERION-BARNER G., MEI H., PENG H. & GUO Y. (2026). Improving clinical diagnosis with counterfactual multi-agent reasoning.

YU H., CHENG T., WANG Y., HE W., WANG Q., CHENG Y., ZHANG Y., FENG R. & ZHANG X. (2025). Finemedlm-o1 : Enhancing medical knowledge reasoning ability of llm from supervised fine-tuning to test-time training. DOI : 10.48550/arXiv.2501.09213.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2020). Bertscore : Evaluating text generation with bert. DOI : 10.48550/arXiv.1904.09675.

ZHENG C., ZHOU H., MENG F., ZHOU J. & HUANG M. (2024). Large language models are not robust multiple choice selectors. DOI : 10.48550/arXiv.2309.03882.

ZHENG L., CHIANG W.-L., SHENG Y., ZHUANG S., WU Z., ZHUANG Y., LIN Z., LI Z., LI D., XING E. P., ZHANG H., GONZALEZ J. E. & STOICA I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. https://arxiv.org/abs/2306.05685.

ZHOU S., XIE W., LI J., ZHAN Z., SONG M., YANG H., ESPINOZA C., WELTON L., MAI X., JIN Y., XU Z., CHUNG Y.-H., XING Y., TSAI M.-H., SCHAFFER E., SHI Y., LIU N., LIU Z. & ZHANG R. (2025a). Automating expert-level medical reasoning evaluation of large language models. *npj Digital Medicine*, **9**(1), 34. DOI : 10.1038/s41746-025-02208-7.

ZHOU S., XIE W., LI J., ZHAN Z., SONG M., YANG H., ESPINOZA C., WELTON L., MAI X., JIN Y., XU Z., CHUNG Y.-H., XING Y., TSAI M.-H., SCHAFFER E., SHI Y., LIU N., LIU Z. & ZHANG R. (2025b). Automating expert-level medical reasoning evaluation of large language models. *npj Digital Medicine*, **9**(1), 34. DOI : 10.1038/s41746-025-02208-7.

ZUO Y., QU S., LI Y., CHEN Z., ZHU X., HUA E., ZHANG K., DING N. & ZHOU B. (2025). MedXpertQA : Benchmarking Expert-Level Medical Reasoning and Understanding. arXiv :2501.18362 [cs], DOI : 10.48550/arXiv.2501.18362.