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

SÛRE : Supervision du Triage aux Ûrgences par Raisonnement Étapé

Mohamed Imed Eddine Ghebriout[1] Gaël Guibon[1, 2] Thomas Laurenceau[6]
Richard Chocron[6] Christophe Cerisara[1] Emmanuel Vincent[1] Ivan Lerner[3, 4, 5]
(1) Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
(2) Université Sorbonne Paris Nord, CNRS, Laboratoire d'Informatique de Paris Nord, LIPN
(3) Inserm, Centre de Recherche des Cordeliers, Université Paris Cité, Sorbonne Université
(4) HeKA, Inria Paris, F-75012 Paris, France
(5) Département d'informatique médicale, Assistance Publique Hôpitaux de Paris, Hôpital Européen Georges Pompidou, Paris, France
{imed-eddine.ghebriout,gael.guibon,christophe.cerisara}@loria.fr
{ivan.lerner,emmanuel.vincent}@inria.fr

Résumé

Les services d'urgence exigent des décisions rapides et fiables dans des contextes cliniques incertains. Bien que les grands modèles de raisonnement obtiennent de bons résultats sur des benchmarks médicaux, leur usage en conditions réelles reste limité par la faible fiabilité de leurs trajectoires de raisonnement. Nous proposons *SÛRE*, une méthode de supervision du raisonnement clinique pour prédire le degré de gravité à partir de l'anamnèse uniquement. Notre approche repose sur un modèle de récompense de processus, *SÛRE*-PRM, entraîné à partir d'annotations automatiques produites par un juge ouvert évaluant indépendamment chaque étape du raisonnement. *SÛRE*-PRM sert ensuite à sélectionner les trajectoires les plus cohérentes pour spécialiser un modèle de triage, *SÛRE*-LLM. Évaluée sur des données réelles d'urgences afin de prédire le besoin réel de soins urgents dans les 48 heures, *SÛRE* améliore les performances des approches directes tout en apportant une meilleure interprétabilité clinique.

Abstract

***SÛRE* : Supervising Emergency Department Triage via Stepwise Reasoning**

Emergency departments require rapid and reliable decision-making under high clinical uncertainty. Although large reasoning models achieve strong performance on medical benchmarks, their deployment in real-world settings remains limited by the low reliability of their reasoning trajectories. We introduce *SÛRE*, a supervision framework for clinical reasoning that predicts patient severity from anamnesis alone. Our approach relies on a Process Reward Model, *SÛRE*-PRM, trained using automatic annotations produced by an open LLM judge that independently evaluates each reasoning step. The resulting *SÛRE*-PRM is then used to select the most coherent trajectories in order to specialize a triage language model, *SÛRE*-LLM. We evaluate our approach on real-world emergency department data to predict urgent care needs within 48 hours. *SÛRE* improves performance over direct inference baselines while providing better clinical interpretability through step-level reasoning validation.

Mots-clés : Triage aux Urgences, Modèles de Récompense Processus, Fiabilité de la CoT.

Keywords: Emergency Triage, Process Reward Models, CoT Reliability.

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 1 : articles scientifiques originaux, pages 524–543.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

1 Introduction

La répartition juste et efficace des ressources médicales est une mission centrale de la médecine d'urgence. À l'entrée des Services d'Accueil des Urgences (SAU), le tri infirmier organise la priorité de prise en charge médicale. Il repose généralement sur des échelles standardisées telles que l'*Emergency Severity Index* (Wuerz *et al.*, 2000, ESI), le *French Emergency Nurses Classification in Hospital scale* (Taboulet *et al.*, 2009, FRENCH), ou encore le *Manchester Triage System* (Mackway-Jones *et al.*, 1997, 2013, MTS), fondées principalement sur la mesure des constantes vitales (fréquence cardiaque et respiratoire, pression artérielle, état de conscience, etc.) et sur des éléments succincts de l'anamnèse (motif de recours, etc.) (Zachariasse *et al.*, 2019). Ces systèmes reposent sur des recommandations d'experts et sont évalués prospectivement selon leur capacité à prédire un proxy du besoin réel de soins urgents (Moll, 2010). Par exemple, chez l'adulte, la sensibilité de ces échelles pour la détection d'une admission en réanimation varie de 0,58 (95% IC (Intervalle de confiance) 0,48 à 0,68) à 0,88 (95% IC 0,70 à 0.96), tandis que leur capacité à identifier les patients à faible risque pouvant retourner à domicile varie de 0,64 (95% IC 0,62 à 0,66) à 0,98 (95% IC 0,95 à 0,99) (Zachariasse *et al.*, 2019). La consommation de soins et l'hospitalisation sont généralement considérées comme des proxys valides du besoin réel de soins urgents (FitzGerald *et al.*, 2010).

De nombreux travaux visent à développer des outils d'aide à la décision pour le triage afin de soulager les services d'urgence et d'améliorer la détection précoce des patients à haut risque (El Arab & Al Moosa, 2025; Almulihi *et al.*, 2024). Parallèlement, les avancées récentes en matière de grands modèles de raisonnement (*Large Reasoning Models* – LRMs) (Agarwal *et al.*, 2025) suggèrent qu'ils vont au-delà du simple appariement statistique et sont capables d'aborder des tâches cognitivement complexes en les décomposant en plusieurs étapes de raisonnement interprétables. Les LRMs atteignent, voire surpassent la performance humaine dans les benchmarks médicaux conventionnels (Zuo *et al.*, 2025). Cependant, un écart significatif persiste entre les performances obtenues sur des benchmarks médicaux et leur utilisation dans des environnements cliniques réels, notamment en raison des exigences élevées en matière de sécurité, de fiabilité et de validation clinique (Singhal *et al.*, 2023). Par exemple, ChatGPTHealth sous-trie 52% des cas (Ramaswamy *et al.*, 2026). L'une des raisons de cet écart est que les systèmes basés sur de grands modèles de langage s'appuient principalement sur des connaissances médicales a priori pour classer les patients.

L'une des caractéristiques du raisonnement médical, y compris en médecine d'urgence, est l'exploration progressive d'hypothèses diagnostiques, ainsi que l'évaluation de la compatibilité du tableau clinique (signes, symptômes et antécédents) avec ces hypothèses, tout en estimant le niveau d'urgence associé. Les techniques de raisonnement actuellement utilisées avec les grands modèles de langage, telles que la chaîne de pensée (*Chain-of-Thought* – CoT) (Wei *et al.*, 2022), restent limitées pour ce type de tâche. En particulier, le risque de *faux positifs* tel un raisonnement erroné conduisant par chance à la bonne décision clinique, et de *faux négatifs* tel un raisonnement globalement cohérent mais se détériorant à la dernière étape, fragilise la fiabilité du système et en réduit la transparence. Pour pallier ce manque de fiabilité, une validation plus granulaire des étapes de raisonnement est nécessaire.

Les modèles de récompense du processus de raisonnement (*Process Reward Models* – PRMs) (Uesato *et al.*, 2022; Lightman *et al.*, 2024) visent à dépasser cette limite en évaluant les étapes intermédiaires du raisonnement afin de guider le modèle vers une trajectoire de raisonnement valide. Bien qu'efficaces dans des domaines tels que les mathématiques (Zhang *et al.*, 2025), leur application au domaine clinique demeure peu explorée. Ce retard s'explique principalement par le coût élevé de la vérification humaine des étapes du raisonnement, ainsi que par le faible consensus entre experts quant à la

pertinence des étapes intermédiaires et à la validité des hypothèses diagnostiques produites au cours du raisonnement.

Nous proposons *SÛRE*, une approche visant à exploiter les bénéfices des PRMs tout en réduisant le coût de l'annotation experte. Notre méthode s'appuie sur un juge indépendant pour évaluer les étapes intermédiaires du raisonnement au-delà de la seule prédiction finale, puis distille cette supervision dans un PRM compact. En s'ancrant sur les décisions cliniques observées dans les données du monde réel, *SÛRE* vise ainsi à aligner le raisonnement du modèle avec la pratique clinique réelle, sans nécessiter d'annotations humaines détaillées étape par étape.

Nous proposons un outil d'analyse de l'anamnèse pour l'aide au triage, utilisable en complément du tri infirmier. L'objectif de prédiction n'est pas le score de tri, mais un proxy du besoin réel de soins urgents, appelé MED-TRI-H48, défini par la prise en charge thérapeutique ou diagnostique, ou par une hospitalisation en soins conventionnels ou en soins intensifs dans les 48 heures. Le modèle se limite aux informations disponibles dans un cadre conversationnel, comme lors d'un appel pré-hospitalier ou d'une première interaction avec le patient, où les constantes vitales et l'examen clinique ne sont pas disponibles. Sur le plan méthodologique, *SÛRE* (i) introduit une supervision étape par étape du raisonnement clinique via *SÛRE*-PRM ; (ii) réduit la dépendance à l'annotation experte, grâce à un modèle juge ouvert ; et (iii) améliore la prédiction du degré de gravité par rapport à la sélection de raisonnements fondée sur le résultat final (ORM), avec un gain de 43%.

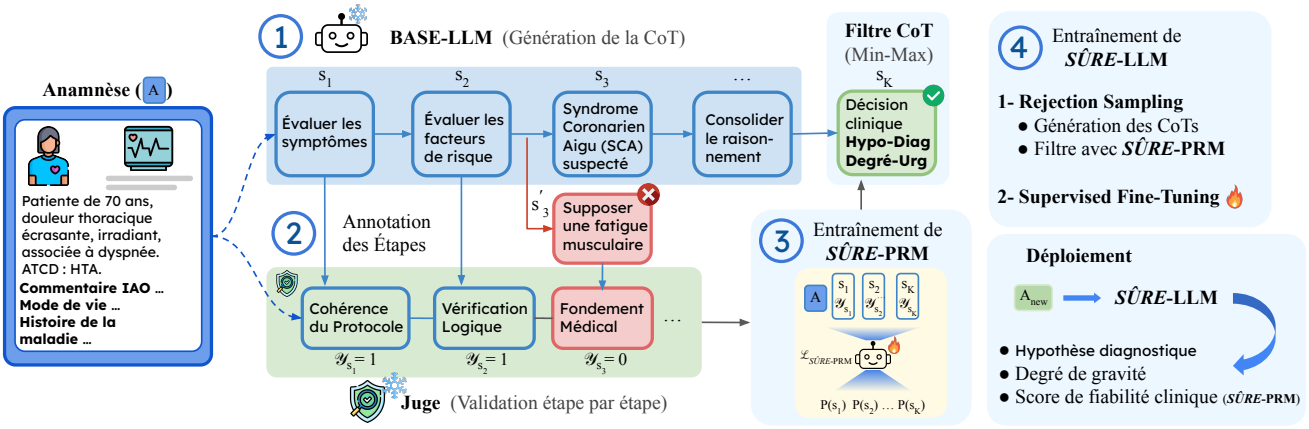

FIGURE 1 – Vue d'ensemble de *SÛRE*. (1) Génération de trajectoires de raisonnement à partir des anamnèses cliniques. (2) Annotation automatique des étapes par le juge. (3) Apprentissage du modèle de récompense de processus (*SÛRE*-PRM). (4) Sélection des trajectoires fiables pour spécialiser le modèle (*SÛRE*-LLM) pour la génération d'hypothèses diagnostiques et du degré de gravité .

2 Travaux connexes

TAL médical et triage.

L'utilisation des modèles de langage pour le triage et d'aide à la décision clinique suscite un intérêt croissant. Certains travaux reposent sur des scénarios simulés afin d'étudier les capacités de raisonnement des LLMs. Par exemple, Hu *et al.* (2024) proposent un jeu de données de scénarios de triage combinant contexte clinique et options de décision, et analysent les performances des modèles en *Zero-Shot* avec des stratégies telles que *Self-Consistency*. Dans la même optique, Lu *et al.* (2024)

introduisent une architecture multi-agents à modèles de langage afin de simuler le raisonnement lors du triage, tandis que Shaposhnikov *et al.* (2025) proposent un assistant clinique reposant sur des agents et une machine à états finis pour analyser les symptômes et orienter les patients. D'autres travaux s'appuient sur des données réelles issues du milieu hospitalier ou pré-hospitalier. Par exemple, He *et al.* (2025) introduisent un corpus d'environ 40 heures d'appels d'urgence destiné à l'extraction d'informations clés pour le triage, tandis que Lansiaux *et al.* (2025) présentent un système de triage basé sur l'analyse automatique des dossiers patients en comparant différentes approches telles que Doc2Vec (Le & Mikolov, 2014), FlauBERT (Le *et al.*, 2020a,b) et XGBoost (Chen & Guestrin, 2016). Enfin, plusieurs travaux évaluent les capacités des grands modèles de langage pour estimer le niveau d'urgence des patients. Par exemple, Williams *et al.* (2024b) montrent que GPT-4 peut prédire l'acuité clinique selon l'échelle ESI (*Emergency Severity Index*) (Wuerz *et al.*, 2000, ESI), avec des performances proches de certaines décisions humaines dans des contextes expérimentaux, sur jeux de données limités et à l'aide de modèles propriétaires. Toutefois, ces travaux soulignent également les limites et les risques associés à l'utilisation des LLMs dans des contextes médicaux critiques, notamment en termes d'influence des modèles de langage sur le raisonnement diagnostique des médecins, montrant que leur utilisation peut modifier les diagnostics proposées par les cliniciens (Goh *et al.*, 2024), tout en soulevant des questions de fiabilité en contexte clinique.

LRMs dans le domaine clinique. Forts du succès des *Large Reasoning Models* (LRMs) dans les domaines des mathématiques et de la programmation, des modèles de fondation médicaux ont adapté ces capacités au contexte clinique (Sellergren *et al.*, 2025; Agarwal *et al.*, 2025). Toutefois, si ces modèles maîtrisent les connaissances médicales théoriques, l'exigence du raisonnement clinique nécessite une synthèse entre cela et les historiques médicaux longitudinaux afin de prendre en compte la nature hautement individualisée des décisions cliniques. Cela reste bien plus difficile que la performance sur bancs d'essai (*benchmarks*) automatisés de type Question-Réponse (QA) et Questions à Choix Multiples (MCQA), dont la difficulté est aujourd'hui jugée insuffisante (Zuo *et al.*, 2025). Pour se hisser à la hauteur de cette complexité, les approches agentiques ont émergé pour tenter de modulariser cette complexité. Dans le cadre du triage, Lu *et al.* (2024) ont utilisé des agents de raisonnement pour récupérer des documents externes et produire des décisions intermédiaires avant d'agréger un consensus final. De manière similaire, dans Li *et al.* (2024), les auteurs simulent l'ensemble de l'environnement hospitalier en spécialisant des agents représentant le médecin, l'infirmier et le patient. Au-delà de ces approches agentiques, Yun *et al.* (2025) utilisent le *Retrieval-Augmented Generation* (Lewis *et al.*, 2020, RAG) et des modèles propriétaires pour générer des données d'entraînement destinées à un modèle de récompense, permettant ainsi la vérification pas à pas du raisonnement de type CoT dans des tâches de question-réponse médicales. D'autres applications émergent également dans la découverte de médicaments (Liu *et al.*, 2025) et la planification de traitements (Qiu *et al.*, 2025). Bien que les cliniciens contribuent largement aux travaux d'évaluation des LRMs (Williams *et al.*, 2024a), leur rythme reste décalé par rapport à l'évolution fulgurante des modèles. En outre, ces évaluations reposent souvent sur des échantillons de petite taille et focalisés sur les modèles propriétaires fermés (Tam *et al.*, 2024). L'intégration directe de ces recherches dans des systèmes médicaux réels, via des solutions souveraines et ouvertes, constitue une voie prometteuse pour combler l'écart entre les benchmarks contrôlés et la complexité brute des situations cliniques réelles.

Modèles de récompense. Les modèles de récompense sont conçus pour capturer les préférences humaines afin d'automatiser l'évaluation et l'alignement des modèles de langage (Ouyang *et al.*, 2022). Toutefois, des travaux récents soulignent la difficulté de développer des modèles de récompense réellement performants, en particulier dans des domaines spécialisés, où l'objectif est d'identifier et de corriger des erreurs de raisonnement au cours du processus d'inférence (Zhang *et al.*, 2025).

Les modèles de récompense de résultat, ou *Outcome Reward Models* (Cobbe *et al.*, 2021, ORM), constituent l'approche la plus connue. Étant donné une question et une trajectoire de raisonnement, ils attribuent un signal de validité global à l'ensemble de la trajectoire via un modèle discriminatif ou génératif. Cependant, pour des tâches complexes, les trajectoires de raisonnement peuvent contenir plusieurs étapes et donc un grand nombre de jetons de raisonnement *(<think>...</think>)*, rendant l'évaluation globale difficile (Luo *et al.*, 2024). Afin de remédier à cette limitation, les modèles de récompense de processus (*Process Reward Models* – PRM) proposent une vérification pas à pas du raisonnement, permettant d'évaluer la validité de chaque étape intermédiaire (Uesato *et al.*, 2022; Lightman *et al.*, 2024). L'annotation des trajectoires de raisonnement constitue une étape critique pour l'apprentissage des PRM et demeure coûteuse et difficile à mettre à l'échelle. Par exemple, Lightman *et al.* (2024) ont recours à des annotateurs humains pour évaluer les étapes de raisonnement, une approche difficilement extensible et généralisable. Dans Wang *et al.* (2024), les auteurs explorent la recherche arborescente de type *Monte Carlo Tree Search* (Kocsis & Szepesvári, 2006; Coulom, 2006, MCTS) associée à un modèle *completer*, chargé de prolonger un raisonnement partiel, tandis que chaque étape est évaluée selon sa probabilité de conduire à une réponse correcte à la fin du raisonnement. Dans le domaine médical, Yun *et al.* (2025) exploitent le RAG et des modèles propriétaires afin de vérifier les étapes de raisonnement, une stratégie incompatible avec les exigences de confidentialité des milieux cliniques. Les modèles de récompense constituent un levier pour la mise à l'échelle à l'inférence (*test-time scaling*). En particulier, les approches *best-of-*N, qui consistent à sélectionner la trajectoire la mieux notée parmi N candidats, surpassent les méthodes de vote majoritaire classiques et améliorent la fiabilité des prédictions finales (Wang *et al.*, 2022).

3 Des données des services d'urgences à la tâche du triage

Nous effectuons une étude rétrospective monocentrique conduite sur des données du monde réel.

Population. La population source de l'étude comprend l'ensemble des patients adultes consultant au service d'accueil des urgences (SAU) de l'Hôpital Européen Georges Pompidou (HEGP, AP-HP) entre le 1er janvier 2014 et le 31 décembre 2023, dont le mode d'entrée est direct depuis le domicile. Les critères d'exclusion comprennent : les patients s'opposant à la réutilisation de leurs données, certaines données manquantes (Échelle Visuelle Analogique - EVA -, motif de recours, mode d'entrée), un nombre de visites supérieur au 99 ème percentile, des filtres qualité (discordance entre la présence d'un transfert depuis le texte libre versus les données structurées, discordance importante entre le tri IAO et l'issue de la consultation d'urgence). Les jeux de données d'entraînement/validation/test sont constitués selon un découpage temporel afin de simuler un scénario prospectif : train, 30 000 visites du 01/01/2014 au 31/12/2020 ; dev, 5000 visites du 01/01/2021 au 30/06/2022 ; test, 5000 visites du 01/07/2022 au 31/12/2023. Ce choix n'exclut pas qu'un même patient puisse apparaître dans plusieurs périodes en cas de visites multiples, ce qui constitue une limite discutée en Section 8.

Données collectées. Les données pseudonymisées sont collectées depuis l'entrepôt de données de santé de l'HEGP (Jannot *et al.*, 2017), conformément à l'avis du CESREES (n°19176515) et à l'autorisation de la CNIL (n°DR-2025-016).

Les données collectées incluent : les données structurées de mouvement du patient (hospitalisation, transfert dans un service ou hôpital), données cliniques (formulaire en texte libre), score de tri IAO (échelle *Manchester Triage System* (Mackway-Jones *et al.*, 1997, 2013, MTS)), décès intra-hospitalier, traitement, résultat de laboratoire et acte diagnostique.

Critères de jugement. Le degré d'urgence a été déterminé rétrospectivement à partir d'un critère composite en quatre classes, `MED-TRI-H48`, reflétant l'urgence d'une prise en charge thérapeutique ou diagnostique, d'une hospitalisation en soins conventionnels ou en soins intensifs, dans les 48 heures. Il est déterminé a posteriori en fonction des décisions médicales prises à l'issue des 48 premières heures et constitue ainsi un standard de référence objectif intégrant l'évolution réelle de l'état clinique du patient. En pratique, il est calculé en assignant le niveau d'urgence le plus élevé en fonction de la présence des événements suivants dans les 48 h : (1) 'Soins intensifs' : décès, hospitalisation en réanimation ou soins intensifs, (2) 'Hospitalisation conventionnelle' : hospitalisation en médecine, gériatrie, chirurgie, service porte médico-chirurgical, (3) 'Acte thérapeutique ou diagnostique' : réalisation d'un examen d'imagerie (radio, scanner, IRM, échographie), prescription d'un traitement aux urgences [1] (médicamenteux, oxygénothérapie, immobilisation), prescription d'un bilan biologique, EVA >=6, durée de passage aux urgences >12 heures, (4) 'Conseil simple' : absence de critères pour (3). Ces évènements sont déterminés à partir de données structurées du dossier patient électronique (DPI), à l'exclusion des transferts (hospitalisation vers d'autres sites), qui font l'objet d'extraction par expression régulière dans le formulaire 'conclusion' du passage aux urgences.

Prédicteurs. Les données d'entrée des modèles pour la prédiction de `MED-TRI-H48` correspondent exclusivement aux informations issues de l'anamnèse, c'est-à-dire les données colligibles dans le cadre d'un échange conversationnel avec le patient, et donc excluant les données nécessitant la réalisation d'examen physique (palpation, auscultation), ou d'examen complémentaire (radiologie, etc). En pratique, ce sont donc les formulaires suivants qui sont inclus et concaténés, pour être donnés en entrée des modèles : âge, sexe, date du jour, commentaire IAO, mode de vie, antécédents médicaux ou chirurgicaux, traitement à l'entrée, histoire de la maladie.

Pré-traitement des données. Les dossiers ont été normalisés afin d'uniformiser l'orthographe (correction des erreurs de saisie) et de standardiser les abréviations médicales, souvent ambiguës et susceptibles de désigner plusieurs domaines ou usages cliniques. Ce choix s'impose compte tenu de la nature bruitée des données issues de situations réelles et vise à garantir que *SÛRE* se concentre sur le cœur du raisonnement clinique plutôt que sur des variations linguistiques. Cette normalisation a été réalisée à l'aide d'une terminologie médicale et d'un pipeline de correction détaillés en Annexe C.

4 Méthodologie

Nous abordons le problème de sélection et d'évaluation automatique des étapes de raisonnement par le biais de phases successives (Figure 1), qui consistent en (1) la génération des trajectoires de raisonnement, (2) la mise en place de données d'entraînement pour le modèle de récompense qui est ensuite (3) appris pour distiller l'évaluation d'une étape de raisonnement en un processus de récompense simple et efficient par un modèle dédié. Enfin, (4) ce processus de récompense sert à trier les trajectoires valides pour affiner et spécialiser notre modèle de langage.

Formulation du problème. Étant donnée une anamnèse brute $\mathcal{A}$, nous cherchons à entraîner un modèle *SÛRE*-LLM capable de prédire le degré d'urgence u tout en générant une trajectoire de raisonnement $S = \{s_i\}_{i=1}^{K}$, composée de K étapes intégrant des hypothèses diagnostiques. La validité de chaque étape est évaluée par un modèle de récompense *SÛRE*-PRM$(\mathcal{A}, S) = \{r_{s_i}\}_{i=1}^{K}$ où $r_{s_i} = P(\text{valide}(s_i) \mid \mathcal{A}, s_1, \ldots, s_{i-1})$ représente la probabilité que l'étape s_i soit cliniquement valide. Les anamnèses $\mathcal{A}$ ont préalablement fait l'objet d'une étape de prétraitement (cf. Annexe D.1)

1. hors paracétamol, métoclopramide, phloroglucinol, métopimazine

visant à standardiser les entrées et réduire le bruit linguistique, afin de garantir que le raisonnement du modèle se concentre principalement sur l'analyse clinique du cas.

(1) Génération des trajectoires de raisonnement. Pour constituer les traces de raisonnement $\mathcal{C}$, nous sollicitons le modèle de base *BASE*-LLM avec différents paramètres de décodage variés (température, top-p). Pour chaque anamnèse $\mathcal{A}$, nous demandons au modèle de générer N trajectoires $\mathcal{C}_{\mathcal{A}} = \{S_j\}_{j=1}^{N}$, afin de favoriser une exploration plus large de l'espace des raisonnements cliniques. Chaque trajectoire intègre des hypothèses diagnostiques et conduit à la prédiction du degré d'urgence, tout en respectant les protocoles de triage en vigueur (cf. Annexe D.2).

(2) Données d'entraînement du PRM. Contrairement aux approches reposant sur des modèles propriétaires dans l'annotation des étapes de raisonnement (Yun *et al.*, 2025) et motivés par les résultats des schémas *LLM-as-a-Judge* ouverts pour cette annotation (Zhang *et al.*, 2025; Yin *et al.*, 2025), nous nous appuyons sur un grand modèle de langage ouvert agissant comme juge. Celui-ci évalue chaque étape de la trajectoire de raisonnement de manière indépendante et détermine si elle est cliniquement valide ou non. Le prompt utilisé pour cette évaluation est fourni en Annexe D.3. Autrement dit, $\text{Juge}(\mathcal{A}, u^*, [S_{\mathcal{A}} : u]) = y_{S_{\mathcal{A}}}$ avec $y_{S_{\mathcal{A}}} = \{y_{s_i}\}_{i=1}^{K} \in \{0,1\}^K$, où $S_{\mathcal{A}}$ désigne la trajectoire contenant les étapes de raisonnement générées sur la base de l'anamnèse $\mathcal{A}$, y_{s_i} le label de validité attribué à l'étape s_i de la trajectoire S et u^* le degré de gravité de référence.

(3) Apprentissage de *SÛRE*-PRM. Étant donnée une étape s_i d'une trajectoire S, *SÛRE*-PRM est entraîné à attribuer un score de validité $r_{s_i} \in [0,1]$. Cet apprentissage est réalisé en insérant un jeton spécial de fin d'étape après chaque étape s_i, tandis que le degré de gravité prédit est concaténé à la dernière étape s_K. Le modèle apprend ainsi à décider si l'étape complétée est valide ou erronée. Le score r_{s_i} est obtenu en appliquant un softmax sur les logits de sortie des deux classes (valide/erronée) pour minimiser l'entropie croisée binaire, définie par rapport aux annotations $\mathbf{y}_{S_{\mathcal{A}}} = \{y_{s_i}\}_1^N$ fournies par le juge : $\mathcal{L}_{\textit{SÛRE}\text{-PRM}} = -\sum_{S\in\mathcal{C}} \sum_{i=1}^{K} (y_{s_i} \log(r_{s_i}) + (1 - y_{s_i}) \log(1 - r_{s_i}))$.

(4) Spécialisation de *SÛRE*-LLM. Une fois le *SÛRE*-PRM optimisé, nous procédons à l'entraînement du modèle de prédiction, *SÛRE*-LLM, à l'aide d'une stratégie d'échantillonnage par rejet (*rejection sampling*). Pour chaque anamnèse $\mathcal{A}$, nous générons un ensemble de trajectoires candidates. Ces dernières sont évaluées par *SÛRE*-PRM, et nous ne conservons que les trajectoires S maximisant la fiabilité minimale de leurs étapes intermédiaires (*max-min*) tout en assurant qu'elles prédisent le bon degré d'urgence : $S^* = \arg\max_{S\in\mathcal{C}} \left(\min_{i=1}^{K} r_{s_i} \right) \cdot \mathbb{I}(u = u^*)$.

Le jeu de données d'entraînement supervisé est ainsi constitué des couples $(\mathcal{A}, S^*)$ les plus robustes. Dans notre cadre, cette stratégie de fine-tuning est privilégiée aux approches basées sur l'apprentissage par renforcement (RL) pour son efficacité computationnelle et sa stabilité de convergence, tout en garantissant un alignement strict avec les protocoles et les trajectoires de triage validées.

5 Protocole expérimental

Pour spécialiser nos modèles *SÛRE*-LLM et *SÛRE*-PRM, nous utilisons comme base (*BASE*-LLM) le modèle de raisonnement Qwen3-8B (Yang *et al.*, 2025). Pour chaque anamnèse, $K = 8$ trajectoires de raisonnement sont générées avant d'être filtrées selon le critère décrit à l'étape (4) de la Section 4. Le modèle juge utilisé pour l'annotation des étapes de raisonnement est GPT-OSS-120B (Agarwal

et al., 2025), choisi pour ses performances solides sur plusieurs benchmarks de raisonnement médical. Afin de s'assurer que le modèle *SÛRE*-PRM soit entraîné sur des étapes de raisonnement pertinentes, c'est-à-dire qui ont une chance de mener à une décision correcte sans être trop évidentes ni aléatoires, nous appliquons un filtrage aux trajectoires générées. Seules les trajectoires de $\mathcal{C}$ dont la précision se situe dans l'intervalle $[0,3\,;\,0,8]$ sont retenues pour l'annotation. Ce filtrage élimine les trajectoires erronées ou triviales, c'est-à-dire celles qui n'apportent pas de raisonnement clinique exploitable (par exemple une répétition de la classe cible ou une justification générique), afin de concentrer l'apprentissage du modèle sur des exemples plus informatifs et difficiles. Les deux modèles sont entraînés à l'aide de la méthode LoRA (Hu *et al.*, 2022) sur une carte NVIDIA H100 NVL (94 GB). Les détails des hyperparamètres sont fournis en Annexe B.

Sélection des baselines. Bien que notre travail se concentre sur les méthodes de raisonnement, nous comparons également notre approche à plusieurs modèles de classification supervisée, notamment *Support Vector Machines* (Cortes & Vapnik, 1995, SVM), les arbres de décision, *XGBoost* (Chen & Guestrin, 2016), ainsi qu'à l'encodeur biomédical spécialisé DrBERT (Labrak *et al.*, 2023) et le modèle *Qwen3-8B* (Yang *et al.*, 2025). Ces méthodes prédisent directement le degré de gravité à partir de l'anamnèse, sans modéliser explicitement un processus de raisonnement clinique. Elles peuvent ainsi exploiter des corrélations statistiques dans les données sans nécessairement capturer la complexité de l'inférence clinique. Le banc de test de *SÛRE* inclut également différentes configurations d'inférence, allant du few-shot jusqu'aux stratégies de mise à l'échelle du temps d'inférence.

Méthode	**QWK ↑**	**MAE ↓**	**F1-W ↑**	**F1-Mac ↑**	**Acc ↑**	**Hospit ↑**
Approches fréquentielles et contextuels (Classification Supervisée)						
TF-IDF + SVM	50,28	0,4716	58,68	47,05	59,12	80,76
FastText + XGBoost	52,66	0,4490	<u>59,99</u>	46,88	<u>60,68</u>	80,78
DrBERT	<u>57,57</u>	0,4478	58,14	44,25	58,92	81,96
BASE-LLM *(Qwen3-8B)*	**60,94**	**0,4002**	**62,09**	**52,46**	**62,32**	83,64
Grands Modèles de Raisonnement (Inférence directe)						
Llama-3.1-8B-Instruct *(0-Shot)*	10,75	0,5980	27,75	19,7	43,26	55,63
BASE-LLM *(Qwen3-8B) (0-Shot)*	31,01	0,7078	34,84	29,55	39,26	65,74
Llama-3.1-8B-Instruct *(5-Shot)*	29,81	0,6856	40,60	31,54	42,72	57,94
BASE-LLM *(Qwen3-8B) (8-Shot)*	37,54	0,6696	42,97	35,31	43,16	68,43
Grands Modèles de Raisonnement (Inférence mise à l'échelle)						
BASE-LLM *(Qwen3-8B) (ORM)*	38,31	0,6399	43,71	36,39	44,30	68,57
SÛRE-LLM *(Qwen3-8B)*	47,38	0,5220	52,07	43,24	53,07	77,75
Maj-Voting	49,60	0,4816	53,83	43,62	55,56	80,84
Best-of-N	54,30	0,4292	57,70	<u>48,30</u>	59,94	<u>84,14</u>
Self-Consistency	54,81	<u>0,4268</u>	57,81	47,17	60,02	**84,32**

TABLE 1 – Prédiction du degré d'urgence. Baselines vs *SÛRE* (scores en % sauf **MAE** en valeur absolue). Les meilleurs et second meilleurs résultats sont en **gras** et <u>soulignés</u>, respectivement.

Évaluation et métriques. Nous considérons deux tâches : prédiction du degré d'urgence et annotation des étapes de raisonnement. Pour évaluer les différentes méthodes sur la tâche d'estimation du degré d'urgence u à partir de l'anamnèse $\mathcal{A}$, nous considérons plusieurs métriques d'évaluation. La première est le κ quadratique pondéré, autrement nommé *Quadratic Weighted Kappa* – **QWK** (Cohen, 1968). Elle sert à mesurer la concordance entre les degrés de gravité prédits par le modèle et les

degrés de gravité de référence, en tenant compte de la nature ordinale des classes et en pénalisant proportionnellement les désaccords selon leur distance. Nous accompagnons cette métrique de l'erreur absolue moyenne (**MAE**), la F1-mesure pondérée (**F1-W**), la F1-mesure macro (**F1-Mac**) ainsi que l'exactitude (*Accuracy* – **Acc**) afin d'obtenir une intuition globale sur les performances de *SÛRE* et ses variantes, comparées aux baselines. Nous considérons une métrique supplémentaire avec le score **Hospit** qui mesure la capacité du modèle à prédire la nécessité d'une hospitalisation dans les 48 heures. Les cas sont regroupés en deux catégories : patients nécessitant une hospitalisation (soins intensifs ou hospitalisation conventionnelle) et patients ne nécessitant pas d'hospitalisation (conseil simple ou acte thérapeutique/diagnostique).
Enfin, pour vérifier la qualité d'annotation des étapes de raisonnement par *SÛRE*, nous considérons la précision et l'aire sous la courbe ROC (**ROC-AUC**) afin de correspondre à la tâche de classification binaire d'annotation des étapes en étape correcte ou non.

Modèle de récompense	**Acc** ↑	**F1-Score** ↑	**Précision** ↑	**ROC-AUC** ↑
BASE-LLM (0-Shot)	44,24	52,15	85,30	55,06
SÛRE-PRM	**90,61**	**92,70**	**91,71**	**92,70**

TABLE 2 – Évaluation intrinsèque de *SÛRE*-PRM au niveau d'annotation des étapes de raisonnement.

6 Résultats

La Table 1 montre les résultats de la tâche de prédiction du degré d'urgence à partir de l'anamnèse. Nous pouvons y remarquer que les approches fréquentielles et contextuelles reposant sur la classification supervisée apprennent les corrélations statistiques ; leur capacité à passer à l'échelle avec un raisonnement complexe reste limitée. À l'inverse, les modèles de langage présentent des performances très faibles en configuration *0-Shot*. Le modèle de base atteint un coefficient de Cohen (QWK) de $31,01$, soit un niveau proche d'une prédiction aléatoire. Cette faible performance en *0-Shot* se reflète également dans l'approche ORM (*Outcome Reward Models* (Cobbe *et al.*, 2021, ORM)), qui évalue les trajectoires de raisonnement uniquement à partir du résultat final. Dans ce cadre, le modèle génère un grand nombre de faux positifs. Lors de l'entraînement, ces trajectoires erronées vont être renforcées, poussant progressivement le modèle à adopter des schémas de raisonnement incorrects et à amplifier ces erreurs au fil des itérations. La spécialisation du modèle *BASE*-LLM permet néanmoins une amélioration significative de $\sim 53\%$ du score QWK. Ce gain est particulièrement notable compte tenu du fait que le modèle est entraîné sur un sous-ensemble restreint des anamnèses, soulignant la qualité du filtrage décrit à l'étape (4) de la Section 4.

Les meilleures performances sont obtenues avec *SÛRE*, qui combine la spécialisation du modèle *SÛRE*-LLM et l'évaluation étape par étape du raisonnement par *SÛRE*-PRM. La spécialisation du modèle permet d'ancrer le raisonnement dans l'espace des hypothèses cliniques plausibles, tandis que *SÛRE*-PRM agit comme un mécanisme de guidage en favorisant les trajectoires de raisonnement les plus cohérentes. Ce dernier atteint une précision supérieure de $90,61\%$ (voir la Table 2) lors de la tâche intermédiaire d'annotation des étapes de raisonnement, ce qui en fait un guide fiable pour orienter la sélection des trajectoires. Couplé à l'utilisation d'une stratégie de *Self-Consistency* même avec un nombre limité de trajectoires ($N = 8$), il permet d'obtenir un gain supplémentaire de $\sim 16\%$ par rapport à *SÛRE*-LLM en QWK. Dans cette configuration, la précision maximale pour la prédiction

des hospitalisations atteint $84,32\%$. Au-delà des performances, *SÛRE* permet de localiser les erreurs au niveau des étapes, contrairement au modèle ORM qui rejette l'ensemble du raisonnement sans fournir d'explication. Cette propriété fournit un signal d'apprentissage plus informatif et facilite l'analyse des défaillances du modèle ; un exemple est présenté en Annexe E. Enfin, bien que ces résultats restent limités à l'échelle de notre expérimentation, ils suggèrent le potentiel de l'approche lorsqu'elle est appliquée à des volumes de données plus importants et à des modèles de plus grande capacité.

7 Conclusion et perspectives

Nous avons présenté *SÛRE*, une méthode de supervision du raisonnement clinique pour l'aide au triage à partir de l'anamnèse. L'idée centrale est que, dans un contexte clinique à haut risque, la fiabilité d'une décision dépend aussi de la validité des étapes intermédiaires qui y conduisent. *SÛRE* s'appuie ainsi sur *SÛRE*-PRM pour évaluer le raisonnement étape par étape et sélectionner les trajectoires les plus cohérentes avant la spécialisation de *SÛRE*-LLM. Nos résultats montrent que cette supervision processuelle améliore la prédiction du degré d'urgence par rapport aux approches directes, tout en offrant une meilleure interprétabilité. Contrairement à un affinage supervisé classique, *SÛRE* cherche à structurer l'inférence clinique autour d'un raisonnement vérifiable. Les perspectives principales concernent l'extension à des données multicentriques, l'exploration de modèles plus larges, ainsi que l'intégration de mécanismes de correction lors des étapes erronées, via des approches inspirées du *Tree of Thoughts* (Yao *et al.*, 2023), ainsi qu'une validation des annotations du juge par des experts médicaux.

8 Limites

Ce travail présente quelques limites. Une limite centrale concerne la qualité du juge automatique utilisé pour annoter les étapes de raisonnement : Même si *SÛRE*-PRM distille efficacement ses décisions, les annotations initiales peuvent refléter les biais, omissions ou erreurs. Par ailleurs, *SÛRE* repose sur la spécialisation d'un modèle de taille modeste ($\sim$8B), ce qui favorise la reproductibilité mais peut limiter ses capacités de raisonnement et de généralisation par rapport à des modèles plus larges. L'exploration des trajectoires reste également restreinte, avec un nombre limité de raisonnements générés par anamnèse, ce qui peut réduire la diversité des hypothèses diagnostiques envisagées, en particulier pour les anamnèses complexes. Cela se reflète notamment dans la performance de base de *SÛRE*-LLM, suggérant que certaines hypothèses cliniques plausibles peuvent ne pas être encore explorées. Enfin, bien que *SÛRE*-PRM atteigne une précision élevée (90,61%) en distinguant les étapes de raisonnement erronées des étapes valides, son entraînement repose principalement sur nos données ; sa robustesse face aux décalages de distribution (*data shift*) devrait donc être confirmée sur un spectre plus large de raisonnements médicaux. De plus, un même patient pouvant apparaître dans différents *splits*, un risque de fuite d'information au niveau patient ne peut être totalement exclu, ce qui pourrait conduire à surestimer les performances de généralisation.

Références

Agarwal S., Ahmad L., Ai J., Altman S., Applebaum A., Arbus E., Arora R. K., Bai Y., Baker B., Bao H. *et al.* (2025). GPT-OSS-120B & GPT-OSS-20B model card. *arXiv preprint arXiv :2508.10925*.

Almulihi Q. A., Alquraini A. A., Almulihi F. A. A., Alzahid A. A., Al Qahtani S. S. A. J., Almulhim M., Alqhtani S. H. S., Alnafea F. M. N., Mushni S. A. S., Alaqil N. A. *et al.* (2024). Applications of artificial intelligence and machine learning in emergency medicine triage-a systematic review. *Medical Archives*, **78**(3), 198.

Chen T. & Guestrin C. (2016). Xgboost : A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, p. 785–794.

Cobbe K., Kosaraju V., Bavarian M., Chen M., Jun H., Kaiser L., Plappert M., Tworek J., Hilton J., Nakano R. *et al.* (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv :2110.14168*.

Cohen J. (1968). Weighted kappa : Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, **70**(4), 213.

Cortes C. & Vapnik V. (1995). Support-vector networks. *Machine learning*, **20**(3), 273–297.

Coulom R. (2006). Efficient selectivity and backup operators in monte-carlo tree search. In *International conference on computers and games*, p. 72–83 : Springer.

El Arab R. A. & Al Moosa O. A. (2025). The role of ai in emergency department triage : An integrative systematic review. *Intensive and Critical Care Nursing*, **89**, 104058.

FitzGerald G., Jelinek G. A., Scott D. & Gerdtz M. F. (2010). Emergency department triage revisited. *Emergency Medicine Journal*, **27**(2), 86–92.

Goh E., Gallo R., Hom J., Strong E., Weng Y., Kerman H., Cool J. A., Kanjee Z., Parsons A. S., Ahuja N. *et al.* (2024). Large language model influence on diagnostic reasoning : A randomized clinical trial. *JAMA Network Open*, **7**(10), e2440969.

He K., Lin Q., Fei H., Chng E. S., Hong D., Ong M. E. H. & Feng M. (2025). InTriage : Intelligent telephone triage in pre-hospital emergency care. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 873–885, Suzhou, China.

Hu B., Ray B., Leung A., Summerville A., Joy D., Funk C. & Basharat A. (2024). Language models are alignable decision-makers : Dataset and application to the medical triage domain. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 6 : Industry Track)*, p. 213–227, Mexico City, Mexico.

Hu E. J., Shen Y., Wallis P., Allen-Zhu Z., Li Y., Wang S., Wang L., Chen W. *et al.* (2022). Lora : Low-rank adaptation of large language models. *Iclr*, **1**(2), 3.

Jannot A.-S., Zapletal E., Avillach P., Mamzer M.-F., Burgun A. & Degoulet P. (2017). The Georges Pompidou University hospital clinical data warehouse : A 8-years follow-up experience. *International Journal of Medical Informatics*, **102**, 21–28.

Kocsis L. & Szepesvári C. (2006). Bandit based monte-carlo planning. In *European conference on machine learning*, p. 282–293 : Springer.

Labrak Y., Bazoge A., Dufour R., Rouvier M., Morin E., Daille B. & Gourraud P.-A. (2023). Drbert : A robust pre-trained model in french for biomedical and clinical domains. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 16207–16221.

Lansiaux E., Azzouz R., Chazard E., Vromant A. & Wiel E. (2025). Development and comparative evaluation of three artificial intelligence models (NLP, LLM, JEPA) for predicting triage in emergency departments : A 7-month retrospective proof-of-concept : Triage intelligent à l'entrée des urgences (TIAEU). In *Proceedings of the IEEE/ACM 12th International Conference on Big Data Computing, Applications and Technologies*, p. 1–10.

Le H., Vial L., Frej J., Segonne V., Coavoux M., Lecouteux B., Allauzen A., Crabbé B., Besacier L. & Schwab D. (2020a). FlauBERT : des modèles de langue contextualisés pré-entraînés pour le français (FlauBERT : Unsupervised language model pre-training for French). In C. Benzitoun, C. Braud, L. Huber, D. Langlois, S. Ouni, S. Pogodalla & S. Schneider, Éds., *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, p. 268–278, Nancy, France : ATALA et AFCP.

Le H., Vial L., Frej J., Segonne V., Coavoux M., Lecouteux B., Allauzen A., Crabbé B., Besacier L. & Schwab D. (2020b). FlauBERT : Unsupervised language model pre-training for French. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis, Éds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 2479–2490, Marseille, France : European Language Resources Association.

Le Q. & Mikolov T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, p. 1188–1196 : PMLR.

Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W.-t., Rocktäschel T., Riedel S. & Kiela D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan & H. Lin, Éds., *Advances in Neural Information Processing Systems*, volume 33, p. 9459–9474 : Curran Associates, Inc.

Li J., Wang S., Zhang M., Li W., Lai Y., Kang X., Ma W. & Liu Y. (2024). Agent hospital : A simulacrum of hospital with evolvable medical agents (2024). *arXiv preprint arXiv :2405.02957*.

Lightman H., Kosaraju V., Burda Y., Edwards H., Baker B., Lee T., Leike J., Schulman J., Sutskever I. & Cobbe K. (2024). Let's verify step by step. In *12th International Conference on Learning Representations*.

Liu S., Lu Y., Chen S., Hu X., Zhao J., Fu T. & Zhao Y. (2025). DrugAgent : Automating AI-aided drug discovery programming through LLM multi-agent collaboration. In *2nd AI4Research Workshop : Towards a Knowledge-grounded Scientific Research Lifecycle*.

Lu M., Ho B., Ren D. & Wang X. (2024). TriageAgent : Towards better multi-agents collaborations for large language model-based clinical triage. In *Findings of the Association for Computational Linguistics : EMNLP 2024*, p. 5747–5764.

Luo L., Liu Y., Liu R., Phatale S., Guo M., Lara H., Li Y., Shu L., Zhu Y., Meng L. *et al.* (2024). Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv :2406.06592*.

Mackway-Jones K., Marsden J. & Windle J. (1997). Manchester triage group. *Emergency triage*.

Mackway-Jones K., Marsden J. & Windle J. (2013). *Emergency Triage : Manchester Triage Group*. John Wiley & Sons.

Moll H. A. (2010). Challenges in the validation of triage systems at emergency departments. *Journal of Clinical Epidemiology*, **63**(4), 384–388.

Ouyang L., Wu J., Jiang X., Almeida D., Wainwright C., Mishkin P., Zhang C., Agarwal S., Slama K., Ray A. *et al.* (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, **35**, 27730–27744.

Qiu P., Wu C., Liu S., Fan Y., Zhao W., Chen Z., Gu H., Peng C., Zhang Y., Wang Y. *et al.* (2025). Quantifying the reasoning abilities of LLMs on clinical cases. *Nature Communications*, **16**(1), 9799.

Ramaswamy A., Tyagi A., Hugo H., Jiang J., Jayaraman P., Jangda M., Te A. E., Kaplan S. A., Lampert J., Freeman R. *et al.* (2026). ChatGPT Health performance in a structured test of triage recommendations. *Nature Medicine*, p. 1–1.

Sellergren A., Kazemzadeh S., Jaroensri T., Kiraly A., Traverse M., Kohlberger T., Xu S., Jamil F., Hughes C., Lau C. *et al.* (2025). Medgemma technical report. *arXiv preprint arXiv :2507.05201*.

Shaposhnikov V., Nesterov A., Kopanichuk I., Bakulin I., Egor Z., Abramov R., Olegovna T. E., Bespalov I. R., Dylov D. V. & Oseledets I. (2025). CLARITY : Clinical assistant for routing, inference, and triage. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing : Industry Track*, p. 1805–1821, Suzhou, China.

Singhal K., Azizi S., Tu T., Mahdavi S. S., Wei J., Chung H. W., Scales N., Tanwani A., Cole-Lewis H., Pfohl S. *et al.* (2023). Large language models encode clinical knowledge. *Nature*, **620**(7972), 172–180.

Taboulet P., Moreira V., Haas L., Porcher R., Braganca A., Fontaine J.-P. & Poncet M.-C. (2009). Triage with the french emergency nurses classification in hospital scale : reliability and validity. *European Journal of Emergency Medicine*, **16**(2), 61–67.

Tam T. Y. C., Sivarajkumar S., Kapoor S., Stolyar A. V., Polanska K., McCarthy K. R., Osterhoudt H., Wu X., Visweswaran S., Fu S. *et al.* (2024). A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ digital medicine*, **7**(1), 258.

Uesato J., Kushman N., Kumar R., Song F., Siegel N., Wang L., Creswell A., Irving G. & Higgins I. (2022). Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv :2211.14275*.

Wang P., Li L., Shao Z., Xu R., Dai D., Li Y., Chen D., Wu Y. & Sui Z. (2024). Math-shepherd : Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 9426–9439.

Wang X., Wei J., Schuurmans D., Le Q. V., Chi E. H., Narang S., Chowdhery A. & Zhou D. (2022). Self-consistency improves chain of thought reasoning in language models. In *11th International Conference on Learning Representations*.

Wei J., Wang X., Schuurmans D., Bosma M., ichter b., Xia F., Chi E., Le Q. V. & Zhou D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, p. 24824–24837.

Williams C. Y., Miao B. Y., Kornblith A. E. & Butte A. J. (2024a). Evaluating the use of large language models to provide clinical recommendations in the Emergency Department. *Nature Communications*, **15**(1), 8236.

Williams C. Y., Zack T., Miao B. Y., Sushil M., Wang M., Kornblith A. E. & Butte A. J. (2024b). Use of a large language model to assess clinical acuity of adults in the emergency department. *JAMA Network Open*, **7**(5), e248895.

Wuerz R. C., Milne L. W., Eitel D. R., Travers D. & Gilboy N. (2000). Reliability and validity of a new five-level triage instrument. *Academic emergency medicine*, **7**(3), 236–242.

Yang A., Li A., Yang B., Zhang B., Hui B., Zheng B., Yu B., Gao C., Huang C., Lv C. *et al.* (2025). Qwen3 technical report. *arXiv preprint arXiv :2505.09388*.

Yao S., Yu D., Zhao J., Shafran I., Griffiths T., Cao Y. & Narasimhan K. (2023). Tree of thoughts : Deliberate problem solving with large language models. *Advances in neural information processing systems*, **36**, 11809–11822.

Yin Z., Sun Q., Zeng Z., Cheng Q., Qiu X. & Huang X.-J. (2025). Dynamic and generalizable process reward modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 4203–4233.

Yun J., Sohn J., Park J., Kim H., Tang X., Shao D., Koo Y. H., Minhyeok K., Chen Q., Gerstein M. *et al.* (2025). Med-PRM : Medical reasoning models with stepwise, guideline-verified process rewards. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, p. 16565–16582.

Zachariasse J. M., van der Hagen V., Seiger N., Mackway-Jones K., van Veen M. & Moll H. A. (2019). Performance of triage systems in emergency care : a systematic review and meta-analysis. *BMJ Open*, **9**(5), e026471.

Zhang Z., Zheng C., Wu Y., Zhang B., Lin R., Yu B., Liu D., Zhou J. & Lin J. (2025). The lessons of developing process reward models in mathematical reasoning. In *Findings of the Association for Computational Linguistics : ACL 2025*, p. 10495–10516.

Zuo Y., Qu S., Li Y., Chen Z.-R., Zhu X., Hua E., Zhang K., Ding N. & Zhou B. (2025). MedXpertQA : Benchmarking expert-level medical reasoning and understanding. In *International Conference on Machine Learning*, p. 80961–80990.

A Considérations éthiques

Bien que ce travail de rétrospective constitue une étape vers le développement de modèles de raisonnement vérifiable dans des environnements cliniques réels, l'utilisation d'un tel système doit être envisagée avec prudence. C'est pourquoi ce travail a été effectué dans le cadre d'un accord du comité éthique et scientifique pour les recherches, les études et les évaluations dans le domaine de la santé, plus communément nommé CESREES.

B Détails de l'implémentation

B.1 Configuration de génération

Dans l'étape (1) de la Section 4, les trajectoires de raisonnement sont générées à partir du modèle *BASE*-LLM en utilisant l'échantillonnage par rejet, avec les paramètres de décodage : température $= 0.7$, $top_p = 0.95$ et $top_k = 50$, favorisant la diversité des trajectoires générées.

Le modèle juge utilisé pour l'annotation des trajectoires de raisonnement est quant à lui exécuté de manière déterministe afin d'assurer la stabilité de l'évaluation. Les paramètres de décodage utilisés sont : température $= 0.0$, $top_p = 1$ et $top_k = 0$. Le nombre maximal de jetons générés est fixé à 4096. Cela permet de disposer de suffisamment d'espace pour analyser et annoter l'ensemble de la trajectoire de raisonnement.

B.2 Configuration d'entraînement

Le modèle de récompense *SÛRE*-PRM est entraîné à l'aide de la méthode LoRA. Nous utilisons un rang $r = 8$, un facteur d'échelle $\alpha = 16$ et un taux de dropout de 0.05, en ciblant les modules de projections d'attention `q_proj`, `v_proj` et `o_proj`. L'entraînement est réalisé avec un taux d'apprentissage de 2×10^{-4}.

Pour la phase de spécialisation du modèle *SÛRE*-LLM via *Rejection Sampling*, nous utilisons également LoRA avec $r = 16$, $\alpha = 32$ et un taux de dropout de 0.1. L'adaptation est appliquée à l'ensemble des couches linéaires du modèle. L'entraînement est effectué avec un taux d'apprentissage de 2×10^{-4} et un *warmup* linéaire de 10%. Afin de préserver les capacités de raisonnement du modèle, les trajectoires de raisonnement sélectionnées sont insérées entre les balises `<think>` et `</think>`, puis concaténées avec la réponse finale contenant l'hypothèse diagnostique et le score de triage.

C Prétraitement des données

Les anamnèses $\mathcal{A}$ issues des SU sont généralement très brutes. Contenant fréquemment des fautes d'orthographe, des informations incomplètes ainsi qu'un usage intensif d'abréviations médicales spécifiques au contexte clinique. Ces abréviations peuvent être ambiguës ou polysémiques, certains termes pouvant désigner plusieurs concepts selon le domaine médical. Dans certains cas, cette

variabilité linguistique peut compliquer l'interprétation automatique des textes. Afin de garantir que le raisonnement dans *SÛRE* se concentre principalement sur l'analyse clinique plutôt que sur la résolution d'ambiguïtés linguistiques, nous avons mis en place une étape de normalisation des anamnèses. Nous avons constitué un lexique d'abréviations médicales à partir d'un glossaire public [2]. Ce glossaire a été collecté automatiquement afin d'obtenir, pour chaque abréviation, une liste de formes développées possibles ainsi que les domaines médicaux associés. Une étape de vérification et de filtrage manuels a ensuite été réalisée afin de conserver uniquement les expansions pertinentes dans le contexte du système de triage français, tout en complétant la liste avec certaines abréviations manquantes. La normalisation des anamnèses est effectuée avec GPT-OSS-120B, utilisé ici comme scribe médical. Il reçoit en entrée l'anamnèse brute ainsi que la liste des abréviations et de leurs expansions possibles, et génère une version réécrite et normalisée du texte. Le prompt utilisé pour cette étape est présenté en Annexe D.1.

D Prompts utilisés

D.1 Prétraitement des anamnèses

System Prompt
Rôle : Tu es un scribe médical expert spécialisé dans la standardisation de dossiers cliniques.
Objectif : Réécrire le dossier patient fourni en respectant strictement les règles suivantes :
Orthographe & Grammaire : Corriger toutes les fautes sans modifier le sens clinique.
Terminologie Médicale : Tous les termes médicaux doivent être écrits en MAJUSCULES.
Abréviations : Utilise le dictionnaire fourni en contexte pour remplacer chaque abréviation par sa forme complète.
Structure : Sexe : Âge : Date de visite : Commentaire infirmier d'accueil : Antécédents allergiques : Antécédents médicaux : Antécédents chirurgicaux : Traitement habituel : Mode de vie : État de conscience : Histoire de la maladie :
Zéro hallucination : Ne génère aucune analyse ou recommandation.
Sécurité : Si un passage est incompréhensible, répondre uniquement par **ERROR**.
À la fin ajouter : "Dossier réécrit avec succès".

Input
Dossier médical : {anamnèse}
Abréviations pertinentes : {anamnèse_abrevs}

D.2 Génération des trajectoires de raisonnement (trois étapes)

System Prompt
Tu es un système expert d'aide au triage médical. Ta mission est d'analyser un dossier patient rédigé en français et de prédire le devenir du patient dans les 48 heures. Tu dois déterminer l'hypothèse diagnostique la plus probable ainsi que le score de triage correspondant.

2. https://abreviationsmedicales.ch

Définition des scores de triage :
0 – **Soins intensifs** : décès dans les 48 heures ou hospitalisation en réanimation / soins intensifs.
1 – **Hospitalisation conventionnelle** : admission en médecine, gériatrie, chirurgie ou dans une unité spécialisée.
2 – **Acte thérapeutique ou diagnostique** : pas d'hospitalisation conventionnelle, mais nécessité d'un examen d'imagerie avancée (scanner, IRM, échographie), d'un bilan biologique, d'un traitement non élémentaire, d'une douleur EVA ≥ 6, ou d'un passage aux urgences supérieur à 12 heures.
3 – **Conseil simple** : retour au domicile en moins de 12 heures, sans imagerie, sans bilan biologique et sans traitement autre qu'un traitement de confort.

Procédure de raisonnement
Tu dois résoudre l'évaluation du triage étape par étape. Chaque étape du raisonnement doit commencer par le format `Step {numéro}:` .

- **Step 1** – Analyser la présentation clinique en examinant les symptômes, les signes cliniques et les antécédents du patient. Une attention particulière doit être portée aux comorbidités, aux hospitalisations récentes, aux refus de soins et à l'évolution chronologique des symptômes.
- **Step 2** – Déterminer l'hypothèse diagnostique la plus probable (en français) à partir de l'analyse effectuée à l'étape 1, en tenant compte des facteurs de risque et des complications possibles.
- **Step 3** – Attribuer le score de triage approprié et justifier cette décision en s'appuyant strictement sur les définitions fournies ci-dessus.

Format de sortie
À la fin du raisonnement, fournir le résultat final strictement au format JSON, selon la structure suivante :

```
## Final Result: {"hypo_diag": "...", "triage": score}
```

Input
Dossier médical : {anamnèse}

D.3 Prompt du juge pour l'annotation des étapes de raisonnement

System Prompt
Tu es un évaluateur médical senior chargé d'analyser des trajectoires de raisonnement produites par un modèle de triage. Pour chaque exemple, tu reçois :

- un dossier médical,
- le score de triage de référence (*Ground Truth*),
- une trajectoire de raisonnement composée de trois étapes.

Ton rôle est d'évaluer chaque étape du raisonnement de manière indépendante et de déterminer si elle est cliniquement valide.

Référence des scores de triage
0 – **Soins intensifs** : décès dans les 48 heures ou hospitalisation en réanimation / soins intensifs.
1 – **Hospitalisation conventionnelle** : admission en médecine, gériatrie, chirurgie ou dans une unité spécialisée.
2 – **Acte thérapeutique ou diagnostique** : pas d'hospitalisation conventionnelle mais nécessité d'un examen d'imagerie avancée (scanner, IRM, échographie), d'un bilan biologique, d'un traitement non élémentaire, d'une douleur EVA ≥ 6, ou d'un passage aux urgences supérieur à 12 heures.
3 – **Conseil simple** : retour au domicile en moins de 12 heures, sans imagerie, sans bilan biologique et sans traitement autre qu'un traitement de confort.

Critères d'évaluation
Attribue un score de **1** si l'étape de raisonnement est cliniquement cohérente, fondée sur des éléments médicaux plausibles et compatible avec les informations du dossier patient et les règles de triage. Attribue un score de **0** si ce n'est pas le cas.
Chaque étape doit être évaluée indépendamment. Par exemple, même si l'étape 1 est incorrecte, l'étape 2 peut recevoir un score de 1 si son raisonnement diagnostique est médicalement valide.

Règles spécifiques pour l'étape 3
L'étape 3 doit recevoir un score de 1 uniquement si :

- le score de triage prédit correspond exactement au score de référence (*Ground Truth*), et
- la justification clinique est cohérente et médicalement fondée.

Si le score de triage est incorrect ou si le raisonnement contient des hypothèses non justifiées, l'étape doit être notée 0.
Le champ `hypo_diag` est fourni pour vérifier que l'hypothèse diagnostique formulée par le modèle est cliniquement compatible avec le score de triage de référence.

Format de sortie
Pour chaque étape, fournis une critique médicale très concise (une phrase), puis indique le score attribué.
Le résultat final doit respecter exactement le format suivant :

```
## Step 1: 0 or 1
## Step 2: 0 or 1
## Step 3: 0 or 1
```

Input
Dossier médical : {anamnèse}
Score de triage de référence : $\{u^*\}$
Trajectoire de raisonnement : {trace}

E Exemple complet d'évaluation de trajectoires de *SÛRE*

Note éthique et confidentialité : L'anamnèse présentée ci-dessous a été entièrement écrite manuellement. Par souci de stricte confidentialité des données de santé, elle ne reflète aucun dossier patient réel, mais constitue une synthèse didactique et élaborée pour illustrer la complexité des situations rencontrées aux urgences.

Anamnèse d'entrée (Cas Complexe : Piège diagnostique)

Sexe : M **Âge :** 62
Mode de vie : Tabagisme actif (30 PA), sédentaire.
Antécédents médicaux : Hypertension artérielle mal équilibrée, dyslipidémie.
Commentaire infirmier d'accueil : Douleur abdo 10/10, patient très algique, agité, sueurs froides.
Histoire de la maladie : Patient se présentant pour une douleur abdominale épigastrique transfixiante irradiant dans le dos, d'apparition brutale il y a 2 heures. Décrit une sensation de « déchirure ». Associée à des nausées, sans vomissement. Pas de fièvre.

Trajectoire 1 : Raisonnement Clinique Valide

Step 1 : Le patient est un homme de 62 ans présentant de multiples facteurs de risque cardiovasculaire. Il consulte pour une douleur abdominale épigastrique d'apparition brutale, transfixiante, irradiant dans le dos. L'intensité maximale (10/10) associée à des sueurs froides évoque d'emblée une urgence vasculaire majeure plutôt qu'une pathologie digestive bénigne.
- *Confiance :* $\mathbf{r_{s_1}} = \mathbf{87.50}\,\%$
Step 2 : Le diagnostic hypothétique le plus probable, devant cette douleur déchirante irradiant dans le dos chez un patient à haut risque cardiovasculaire, est une Dissection Aortique ou une Rupture d'anévrisme de l'aorte abdominale.
- *Confiance :* $\mathbf{r_{s_1}} = \mathbf{96.48}\,\%$
Step 3 : Ces hypothèses diagnostiques sont des urgences chirurgicales et médicales absolues engageant le pronostic vital à très court terme. Le patient nécessite une admission immédiate en salle de déchocage. Le score de triage est donc 0.
Final Result : {"hypo_diag" : "Dissection Aortique", "triage" : 0}
- *Confiance :* $\mathbf{r_{s_1}} = \mathbf{62.50}\,\%$

Analyse : *La trajectoire est cohérente de bout en bout. SÛRE valide le fait que le modèle génératif a déjoué le piège de la "douleur abdominale" pour s'orienter vers l'urgence vasculaire vitale appropriée.*

Trajectoire 2 : Rupture Logique et Mise en Danger

Step 1 : Le patient est un homme de 62 ans consultant pour une douleur abdominale épigastrique transfixiante (10/10) associée à des sueurs froides et une agitation. Ses antécédents incluent

l'HTA et le tabagisme actif. Ces signes traduisent une détresse aiguë sévère.
- *Confiance :* $\mathbf{r_{s_1}} = \mathbf{99.22}\,\%$
Step 2 : Devant l'intensité de la douleur, l'irradiation dans le dos ("déchirure") et le terrain cardiovasculaire, le diagnostic le plus probable est une Dissection Aortique. Il s'agit d'une urgence vitale absolue nécessitant une prise en charge immédiate.
- *Confiance :* $\mathbf{r_{s_1}} = \mathbf{98.44}\,\%$
Step 3 : Puisque le patient ne présente pas de fièvre ni de vomissements, son état est jugé stable. Il peut être renvoyé à son domicile avec une simple prescription d'antalgiques et l'instruction de revenir si la douleur s'aggrave. Le score de triage est 3.
Final Result : {"hypo_diag" : "Dissection Aortique", "triage" : 3}
- *Confiance :* $\mathbf{r_{s_1}} = \mathbf{2.75}\,\%$

Analyse : *SÛRE détecte une rupture logique majeure à l'étape 3. Bien que le diagnostic d'urgence vitale ait été correctement posé à l'étape 2 (validée à 98 %), la déduction clinique qui en découle (retour à domicile) est en contradiction totale avec le bon sens médical. Le modèle sanctionne immédiatement cette incohérence, neutralisant la trajectoire avant l'apprentissage.*