Prédiction du degré d'urgence à partir de l'anamnèse : Apport des grands modèles de langage

Mohamed Imed Eddine Ghebriout, Thomas Laurenceau, Richard Chocron, Emmanuel Vincent, Christophe Cerisara, Gaël Guibon, Ivan Lerner

Prédiction du degré d'urgence à partir de l'anamnèse : Apport des grands modèles de langage.

Mohamed Imed Eddine Ghebriout[1,5,†] Thomas Laurenceau[5] Richard Chocron[5]
Emmanuel Vincent[1] Christophe Cerisara[1] Gaël Guibon[1,2] Ivan Lerner[3,4,5]
(1) Univ. Lorraine, CNRS, Inria, LORIA, Nancy (2) Univ. Sorbonne Paris Nord, LIPN, Villetaneuse
(3) Inserm, CRC, Univ. Paris Cité (4) HeKA, Inria Paris (5) DIM, AP-HP, HEGP, Paris

† *Auteur correspondant* : imed-eddine.ghebriout@loria.fr

ARTICLE : **Accepté à** Journée llm@hopital.fr.

---

1 Introduction

Face à la saturation des Services d'Urgences (SUs) et aux risques de sous-triage impactant le pronostic clinique (O'Connor *et al.*, 2014; Yoon *et al.*, 2003), une hiérarchisation efficace via la Grille FRENCH (Taboulet *et al.*, 2019) est cruciale. S'appuyant sur les performances des LLMs en médecine (Sandmann *et al.*, 2025), nous évaluons ici leur capacité à prédire le degré d'urgence à partir de la seule anamnèse, sans examen clinique ou paraclinique. Cette approche ouvre la voie à des systèmes d'aide au pré-triage ou à des agents conversationnels de recueil d'informations.

Le degré d'urgence reflète à la fois l'intensité du risque vital et/ou fonctionnel encouru par un patient, ainsi que la rapidité avec laquelle ce risque peut s'aggraver. Son évaluation par le médecin ou l'infirmier détermine le niveau de soins requis et le délai maximal avant leur réalisation. Nous proposons de le caractériser par deux objectifs complémentaires : *(1)* ***IAO-TRI-H0***, qui caractérise l'urgence d'une consultation médicale, déterminé rétrospectivement à partir du score de tri établi à l'arrivée du patient (H0) par l'Infirmier d'Accueil et d'Orientation (IAO), sur la base du motif de recours aux urgences et des paramètres physiologiques (fréquence respiratoire, pression artérielle, etc.). *(2)* ***MED-TRI-H48***, qui caractérise l'urgence d'une hospitalisation (soins intensifs ou conventionnelle), ou d'un acte thérapeutique ou diagnostique, déterminé à partir des décisions médicales prises dans les premières 48 heures. Au-delà de ce délai, la prise en charge ne relève plus d'une situation d'urgence initiale.

2 Méthodes

Données. Nous avons exploité une cohorte rétrospective de 575 829 visites aux SUs de l'Hôpital Européen Georges-Pompidou entre 2014 et 2024. Après exclusion des patients mineurs et des dossiers incomplets, la cohorte finale comprend 293 342 visites. Pour chaque visite, nous avons extrait les **entrées** (données démographiques : âge, sexe ; notes de triage : commentaire infirmier, mode de vie, antécédents, histoire de la maladie, traitement habituel, état de conscience), à l'exclusion des constantes vitales, ainsi que les **cibles** : ***IAO-TRI-H0*** ([*Urgence absolue, Urgence vraie, Urgence relative, Consultation urgente, Consultation de médecine*]) et ***MED-TRI-H48*** ([*Soins intensifs, Hospitalisation conventionnelle, Acte thérapeutique ou diagnostique, Conseil simple*]).

Nous présentons des résultats préliminaires sur un échantillon aléatoire 30 000 visites.

Modèles. Nous comparons plusieurs stratégies de modélisation : **Approches fréquentielles :** TF-IDF couplé à des classifieurs (SVM, XGBoost). **Modèles contextuels :** Utilisation de *DrBERT-7GB* (Labrak *et al.*, 2023) pour évaluer l'apport des plongements lexicaux contextuels. **Grands Modèles de Langue :** Exploration des capacités des modèles *Llama-3.1-8B-Instruct* (Grattafiori *et al.*, 2024) et *GPT-OSS-20b* (Agarwal *et al.*, 2025), en Zéro-Shot, Few-Shot, Fine-tuning sur les visites des urgences et les raisonnements cliniques.

Éthique. Cette étude rétrospective a été conduite conformément à l'avis du CESREES (n°19176515) et à l'autorisation de la CNIL (n°DR-2025-016).

3 Résultats

La concordance entre l'évaluation infirmière ($H0$) et le devenir du patient ($H48$) fournit un signal informatif mais reste inférieure aux approches fréquentielles. Concernant les modèles BERT : bien que les tâches soient corrélées, le Multi-Label surpasse l'approche Multi-Tâche, qui impose des décisions strictes (softmax), pénalisant la convergence face au bruit et aux chevauchements entre classes (ex : niveaux 2 vs 3). Malgré leurs performances sur les benchmarks (Sandmann *et al.*, 2025), les LLMs peinent en zéro/few-shot. Le modèle *Llama-3.1-8B-Instruct* fine-tuné atteint la meilleure performance globale de **62,2%** et **61,49%** en Weighted Cohen's kappa pour le triage à ***H0*** et ***H48***, respectivement, ainsi qu'une exactitude de **84,64%** pour la prédiction de l'hospitalisation. Les scores par classe étant rapportés dans la Table 2.

Méthode	*IAO-TRI-H0*					*MED-TRI-H48*				
	QWK ↑	MAE ↓	F1-W ↑	F1-Mac ↑	Acc ↑	QWK ↑	MAE ↓	F1-W ↑	F1-Mac ↑	Acc ↑
Concordance H0 ⇒ H48	-	-	-	-	-	52,74	0,528	50,9	30,51	50,78
Approches fréquentielles										
TF-IDF + SVM	43,93	0,5066	53,36	37,71	54,96	50,54	0,471	58,68	47,03	59,12
FastText + XGBoost	43,47	0,4812	52,98	35,60	56	52,66	0,449	59,99	46,88	60,68
Modèles contextuels										
DrBERT (MultiTâche)	46,78	0,4742	52,89	31,24	55,84	52,74	0,4478	58,14	44,25	58,92
DrBERT (MultiLabel)	51,65	0,437	56,49	37,9	58,98	57,57	0,4196	62,38	48,67	62,96
Grands Modèles de Langue										
Llama 0-Shot	26,21	0,7496	42,83	29,53	43,24	10,75	0,598	27,75	19,7	43,26
OSS 0-Shot	33,25	1,1143	22,38	20,69	23,91	40,84	0,6835	44,63	30,24	44,53
Llama 5-Shot	32,31	0,7826	40,8	30,18	39,12	29,81	0,6856	40,6	31,54	42,72
OSS 5-Shot	39,66	0,8386	35,91	28,24	34,7	42,41	0,6211	47,59	32,55	47,22
Llama (LoRA)	**62,2**	**0,4004**	**61,14**	**49,7**	**62**	**61,94**	**0,3902**	**64,09**	**53,46**	**64,32**

TABLE 1 – Performances des modèles (scores en % sauf *MAE* en valeur absolue).

4 Discussion

Les performances par classe (Table 2) reflètent la distribution des données : les classes *Relative* et *Cons.U* sont prédominantes (74,58 %) pour $H0$, de même que *Acte/Diag* et *Conseil* pour $H48$ (73,62 %), biaisant l'apprentissage au détriment des classes rares. Nos résultats s'alignent sur les travaux de Guerra-Adames *et al.* (2025). Bien que ces derniers rapportent un F1-W supérieur (66,5 %), la comparaison directe reste limitée par l'utilisation de constantes vitales et l'exclusion de cas critiques, ce qui facilite la tâche. Nous confirmons néanmoins la capacité robuste des LLMs à prédire l'hospitalisation à partir de la seule anamnèse textuelle, égalant, voire surpassant, la valeur prédictive des paramètres physiologiques utilisés au tri. Couplés aux résultats sur l'extraction d'informations de régulation médicale (Ghebriout *et al.*, 2025), ces résultats démontrent que les LLMs saisissent les descripteurs cliniques clés, ouvrant la voie à des outils d'assistance au triage sur mesure. Les travaux futures incluront l'évaluation de mécanismes de raisonnement par étapes (Li *et al.*, 2025) pour affiner la détection des urgences vitales.

	IAO-TRI-H0 (F1-Score par classe)				
	Absolue	**Vraie**	**Relative**	**Cons. U**	**Cons. Med**
Llama (LoRA)	36,50	53,13	67,75	60,76	30,38
	MED-TRI-H48 (F1-Score par classe)				
	SI	**Hospit.**	**Acte/Diag**	**Conseil**	
Llama (LoRA)	18,08	66,80	65,08	63,88	

TABLE 2 – Détail des performances par classe du meilleur modèle.

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*.

GHEBRIOUT M. I. E., GUIBON G., LERNER I. & VINCENT E. (2025). Quartz : Qa-based unsupervised abstractive refinement for task-oriented dialogue summarization. In *Findings of the Association for Computational Linguistics : EMNLP 2025*, p. 14689–14706.

GRATTAFIORI A., DUBEY A., JAUHRI A., PANDEY A., KADIAN A., AL-DAHLE A., LETMAN A., MATHUR A., SCHELTEN A., VAUGHAN A. *et al.* (2024). The llama 3 herd of models. *arXiv preprint arXiv :2407.21783*.

GUERRA-ADAMES A., AVALOS-FERNANDEZ M., DORÉMUS O., CELI L. A., GIL-JARDINÉ C. & LAGARDE E. (2025). A counterfactual llm framework for detecting human biases : A case study of sex/gender in emergency triage. *arXiv preprint arXiv :2511.17124*.

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. *arXiv preprint arXiv :2304.00958*.

LI Z.-Z., ZHANG D., ZHANG M.-L., ZHANG J., LIU Z., YAO Y., XU H., ZHENG J., WANG P.-J., CHEN X. *et al.* (2025). From system 1 to system 2 : A survey of reasoning large language models. *arXiv preprint arXiv :2502.17419*.

O'CONNOR E., GATIEN M., WEIR C. & CALDER L. (2014). Evaluating the effect of emergency department crowding on triage destination. *International journal of emergency medicine*, **7**(1), 16.

SANDMANN S., HEGSELMANN S., FUJARSKI M., BICKMANN L., WILD B., EILS R. & VARGHESE J. (2025). Benchmark evaluation of deepseek large language models in clinical decision-making. *Nature Medicine*, p. 1–1.

TABOULET P., MAILLARD-ACKER C., RANCHON G., GODDET S., DUFAU R., VINCENT-CASSY C., YORDANOV Y. & EL KHOURY C. (2019). Triage des patients à l'accueil d'une structure d'urgences. présentation de l'échelle de tri élaborée par la société française de médecine d'urgence : la french emergency nurses classification in hospital (french). *Annales françaises de médecine d'urgence*, **9**(1), 51–59.

YOON P., STEINER I. & REINHARDT G. (2003). Analysis of factors influencing length of stay in the emergency department. *Canadian Journal of Emergency Medicine*, **5**(3), 155–161.