Exploring the Ethical Challenges of Large Language Models in Emergency Medicine: A Comparative International Review

Mahmoud Elbattah [1,2], Emilien Arnaud [1,3], Daniel Aiham Ghazali [3,4,5], Gilles Dequen [1]
[1] Laboratoire MIS, Université de Picardie Jules Verne, Amiens, France
[2] College of Art, Technology and Environment, University of the West of England, Bristol, United Kingdom
[3] Department of Emergency Medicine, Amiens-Picardy University Hospital, Amiens, France
[4] Université de Picardie Jules Verne, Amiens, France
[5] University of Paris Cité, Paris, France

***Abstract*— Large Language Models (LLMs) hold promise for advancing Emergency Medicine by enhancing operational efficiency and supporting decision-making. This scoping review explores the ethical, legal, and global considerations influencing LLM deployment in emergency care. Key ethical concerns, including patient safety, data privacy, and transparency, emphasise the need for explainable AI (XAI) to build trust and prevent biased outputs. Legal challenges highlight the importance of regulatory compliance, especially regarding data protection laws like the GDPR. Significant international variability in LLM adoption further underscores the need for harmonised guidelines to ensure safe and equitable AI integration across diverse healthcare systems. To advance the responsible use of LLMs, future research should prioritise model transparency, consider resource-limited settings, and focus on establishing robust regulatory frameworks.**

Keywords*— *Large Language Models, LLMs, Emergency Medicine, Emergency Department.

I. Introduction

The integration of artificial intelligence (AI) in healthcare, particularly through advancements in Natural Language Processing (NLP), has gained considerable attention due to its potential to transform patient care, streamline clinical workflows, and support decision-making [1, 2]. Healthcare environments, especially Emergency Medicine, rely heavily on the management of vast amounts of unstructured text data, including clinical notes, patient histories, and triage records.

Accordingly, NLP has emerged as a valuable tool to meet these demands, providing methods to process and interpret clinical language, extract relevant medical information, and identify important patterns in patient care. Applications of NLP are increasingly used to automate clinical documentation, summarise patient histories, and identify key symptoms, reducing the cognitive load on healthcare providers who often work under intense time pressures (e.g. [3-5]).

Building on the advancements in NLP, Large Language Models (LLMs), such as GPT models [6, 7] , represent a new frontier offering enhanced capabilities to understand and generate coherent clinical language. In Emergency Medicine, where timely and accurate decision-making is critical, LLMs are being experimented to assist in areas such as triage, documentation, and real-time clinical support [8]. By rapidly processing and analysing large volumes of text, LLMs aim to alleviate administrative burdens, improve accuracy, and support healthcare providers in delivering quality care in emergency settings [9].

However, as with many technological advancements, the implementation of LLMs in emergency settings is not without challenges. Ethical concerns such as patient safety, privacy, and fairness have raised questions about the responsible use of AI in critical care [8, 10]. Legal and regulatory frameworks, which are still evolving, struggle to keep pace with the rapid deployment of AI technologies, leaving gaps in accountability and compliance across different healthcare systems [11]. These concerns are especially pronounced in the high-stakes environment of Emergency Departments, where errors can have significant consequences for patient outcomes [12].

Adding complexity to this landscape is the variability in how LLMs are adopted and regulated internationally. While some countries, particularly those with advanced healthcare infrastructure and robust regulatory systems, are early adopters of LLM technology in healthcare, others face barriers due to resource constraints or lack of regulatory clarity [13, 14] . This disparity highlights the need for a comparative approach to understand how different regions are addressing the ethical and legal challenges posed by applying LLMs in medicine.

In view of that, this scoping review aims to synthesise emerging literature on the ethical, legal, and comparative international aspects of LLM use in Emergency Medicine. By focusing on these dimensions, this review seeks to identify common themes, challenges, and gaps in knowledge, providing a foundation for future research and guiding principles for responsible AI deployment in emergency care.

II. Review Methodology

This scoping review aims to explore the ethical, legal, and international aspects of LLMs in Emergency Medicine by systematically gathering and analysing relevant literature. The methodology was designed to capture key themes and identify gaps in the existing research, rather than to provide a comprehensive review.

A. *Review Questions*

To guide the scope of this review, the following questions were developed:

- What are the primary ethical considerations associated with the use of LLMs in Emergency Medicine?
- What legal and regulatory challenges are identified in the literature regarding LLM deployment in emergency care?
- How do different countries approach the adoption and regulation of LLMs in Emergency Medicine, and what factors influence these approaches?

B. *Search Strategy*

- **Database:** All literature was sourced from PubMed, chosen for its extensive coverage of medical and healthcare-related studies.
- **Keywords and Search Terms:** Keywords used for the search included "Large Language Models", "LLM", "Emergency Medicine", and " Emergency Department". We exclusively employed the "AND" operator to ensure the results were highly relevant to all keywords.
- **Timeframe:** The search was limited to studies published from 2020 onwards to focus on recent advancements in LLMs and their application in healthcare. Table 1 summarises the search strategy.

C. *Inclusion and Exclusion Criteria*

Inclusion Criteria:

- Studies discussing LLM applications in Emergency Medicine.
- Papers covering ethical considerations, legal challenges, or regulatory perspectives in healthcare settings, with a focus on Emergency Departments where available.
- Literature on the international adoption of LLMs, including country-specific or region-specific studies.

Exclusion Criteria:

- Studies focusing on general AI applications in healthcare without a focus on LLMs.
- Research involving traditional NLP methods without the use of large models, to ensure the review centres specifically on recent advances in LLM technology.
- Non-English language studies were excluded.
- Studies that are distant from the context of emergency care or Emergency Departments.

D. *Data Extraction and Analysis*

Relevant information was extracted from each study, with attention to study type, region, and key findings. Each paper was analysed to learn aspects pertaining to ethical considerations, legal implications, and international adoption.

A thematic analysis approach was used to categorise data into three primary theme including: i) Ethical Considerations, ii) Legal and Regulatory Challenges, and iii) Comparative International Adoption. Each theme addresses an aspect of LLMs in Emergency Medicine, identifying both common issues and region-specific challenges. This approach enabled a structured review of recurring concepts, and challenges within each thematic area.

TABLE 1. SUMMARY OF SEARCH STRATEGY.

Digital Library	**PubMed**
Search Terms	Large Language Models AND Emergency Medicine Large Language Models AND Emergency Department LLM AND Emergency Medicine LLM AND Emergency Department
Search Items	Title, Abstract, Keywords
Types of Document	Conference Proceedings, Journal Articles
Timespan	2020-2024
Language	English

III. REVIEW ANALYSIS

A total of 30 papers were considered in this review, selected from an initial retrieval of 196 records identified through a PubMed search. Following a relevance screening, 166 papers were excluded based on the criteria outlined in the methodology. From the remaining 30, we included only studies directly aligned with at least one of the three primary themes defined earlier. This section presents an analysis of findings from the selected literature, organised around these themes. Additionally, Table 2 provides a concise overview of the key aspects of this analysis.

A. *Ethical Considerations*

The ethical deployment of LLMs in Emergency Medicine raises critical concerns surrounding patient safety, data privacy, fairness, and transparency:

1. **Patient Safety**: Studies indicate that while LLMs offer potential benefits for triage and clinical decision support, they carry risks associated with incorrect or oversimplified recommendations. Misinterpretations by LLMs, due to incomplete data or biases in training data, may lead to harmful outcomes in high-stakes emergency settings [10, 15].
2. **Data Privacy and Confidentiality**: The reliance on patient data for LLM training introduces privacy concerns, particularly in adhering to strict healthcare data protection standards. There is a need for secure data handling protocols and anonymisation practices to safeguard patient confidentiality [8].
3. **Bias and Fairness**: Bias in LLMs remains a persistent concern, as these models may inadvertently reinforce existing health disparities, especially for

underrepresented groups. This poses an ethical risk of reinforcing inequities in Emergency Medicine, where diverse populations are treated [9].

4. **Transparency and Explainability**: The complexity of LLMs can result in “black-box” models, making it difficult for healthcare providers to interpret AI-generated outputs. The literature suggests integrating explainable AI (XAI) techniques to enhance trust and accountability in clinical settings [16].

B. *Legal and Regulatory Challenges*

The legal framework for LLMs in healthcare, particularly in Emergency Medicine, remains underdeveloped, with countries facing unique regulatory challenges:

1. **Accountability**: A primary concern is the ambiguity around accountability when LLMs contribute to clinical decisions. The extent of clinician responsibility when following AI-generated recommendations is unclear, introducing potential legal risks in cases of error [12].
2. **Compliance with Data Protection Regulations**: The application of LLMs involves handling sensitive data, requiring adherence to privacy laws like General Data Protection Regulation (GDPR) in the EU and HIPAA in the US. The literature underscores the challenges in ensuring compliance across systems and borders [17].
3. **Guidelines and Standards**: The absence of established guidelines for LLM use in clinical workflows creates legal uncertainty. Calls for standardised guidelines stress the need for clear policies that safeguard patient rights and ensure regulatory alignment [18].

C. *International Adoption and Policy Variations*

There are considerable international differences in the adoption and regulation of LLMs in Emergency Medicine, shaped by healthcare infrastructure, policies, and cultural perceptions:

1. **Variability in Adoption**: High-resource countries, such as the United States and the United Kingdom, lead in adopting LLMs due to advanced infrastructure and research funding. Conversely, low-resource settings face barriers to deploying LLMs, resulting in disparities in AI-driven care access [15, 18].
2. **Policy and Regulatory Differences**: National policies on AI in healthcare vary significantly, with the EU and the US having relatively established frameworks, while other regions may lack comparable guidance, leading to inconsistencies in regulation [10].
3. **Regional Comparisons**: Case studies illustrate how different countries approach the use of LLM in Emergency Medicine, with the US emphasising robust regulatory oversight and European nations prioritising privacy compliance. These examples highlight the need for harmonised standards globally to ensure equitable implementation of LLMs [9, 16].

TABLE 1. ANALYSIS SUMMARY.

Theme	Key Considerations	Examples / Sources
Ethical Considerations	Patient Safety	[10, 15]
	Data Privacy and Confidentiality	[8]
	Bias and Fairness	[9]
	Transparency and Explainability	[16]
Legal and Regulatory Challenges	Accountability and Liability	[12]
	Compliance with Data Protection Regulations	[17]
	Guidelines and Standards	[18]
International Adoption and Policy Variations	Variability in Adoption	[15, 18]
	Policy and Regulatory Differences	[10]
	Case Studies and Regional Comparisons	[9, 16]

IV. DISCUSSION

This section aims to synthesise the key findings, explores their implications for practice and policy, and identifies areas for future research to support the responsible deployment of LLMs in high-stakes medical environments. A visual summary of these considerations and their intersections is presented in Figure 1.

A. *Ethical and Practical Implications*

LLMs offer promising solutions for alleviating workload burdens in emergency settings, particularly in streamlining triage and clinical documentation [10, 15]. However, the ethical implications identified, including patient safety, model bias, and data privacy, suggest a need for careful, context-sensitive implementation [8, 9]. Given the potential for LLMs to inadvertently reinforce healthcare disparities through biased training data, it is essential that these models are developed and monitored with diverse populations in mind [16]. Additionally, the emphasis on XAI methods underscores the importance of transparency to build trust among healthcare providers and ensure the interpretability of AI-generated outputs [17, 22].

The need for XAI is particularly crucial in healthcare due to the field’s rigorous regulatory standards. For example, under the European Union’s GDPR, patients have the right to receive transparent explanations about automated decisions that affect their medical care [19]. This regulatory requirement highlights the critical role of XAI in ensuring that AI-driven healthcare solutions are not only efficient but also legally compliant and ethically sound, fostering a transparent and patient-centred approach to medical care. Without explainability, the black-box nature of advanced AI models may hinder their acceptance in critical care settings. Studies also reveal that LLMs sometimes "hallucinate," or produce incorrect outputs, which could lead to harmful misinformation in medical settings, without rigorous validation [20].

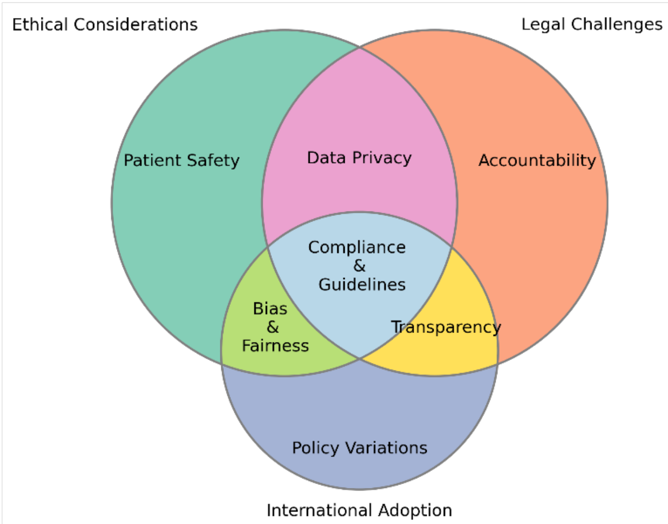

Figure 1: Ethical, Legal, and International Considerations in LLM Deployment for Emergency Medicine. The diagram highlights primary concerns like patient safety, data privacy, and transparency, with overlaps showing shared challenges such as accountability and standardised guidelines critical for responsible LLM integration in healthcare.

B. Legal and Regulatory Challenges

Legal ambiguity around LLM use in medicine introduces potential risks that should be addressed for safe clinical integration [12, 18]. The literature indicates a lack of clarity on accountability when AI recommendations influence patient care decisions, which is further complicated by variations in regulatory frameworks across countries [9]. The need for harmonised guidelines that delineate clear roles and responsibilities, both for clinicians and technology providers, is apparent. Furthermore, robust compliance mechanisms with data protection regulations such as GDPR and HIPAA are necessary to mitigate privacy risks [8]. Establishing clear protocols around data handling and anonymisation is essential to uphold patient confidentiality, particularly as LLMs become more integrated into healthcare systems [17]. Moreover, recent studies argue that the tendency of LLMs to generate variable outputs complicates responsibility when errors arise, reinforcing the necessity of a regulatory framework that balances LLM deployment with accountability measures [21].

C. International Perspectives and Policy Recommendations

The review highlights notable discrepancies in LLM adoption and policy approaches across regions, with high-resource countries leading in AI-driven care solutions due to advanced infrastructure and regulatory readiness [8, 10]. In contrast, low-resource settings face obstacles that hinder LLM implementation, which may exacerbate disparities in healthcare access. To ensure equitable access to AI advancements, international collaboration is needed to develop standards that consider diverse resource contexts. Such standards could provide a foundation for equitable deployment, allowing countries with varying capabilities to benefit from AI-driven innovations without compromising ethical or regulatory standards [16].

V. FUTURE DIRECTIONS

Future research on LLMs should address their broader integration into healthcare, with Emergency Medicine serving as a critical and high-stakes subset of this domain. In general healthcare applications, the development of context-specific models remains a priority. Generic LLMs may not capture the nuances of specialised fields like Emergency Medicine, where time-sensitive and complex decisions require precise, tailored outputs. Future efforts should focus on training models with datasets representative of real-world scenarios in specific healthcare contexts, ensuring their utility and reliability across diverse environments.

The need for XAI remains a critical priority. While LLMs offer advanced capabilities, clinicians require clear and interpretable insights to make informed decisions. Efforts should focus on designing explainability frameworks that are universally applicable across healthcare while being tailored to the fast-paced and high-stakes nature of emergency settings. This may involve developing user-friendly interfaces that provide concise and actionable explanations. Incorporating clinician feedback loops as a human-in-the-loop approach can further enhance explainability. Healthcare professionals, particularly emergency clinicians, play a vital role in improving model accuracy and usability by identifying errors and contributing domain-specific expertise. Embedding feedback mechanisms into healthcare AI systems would establish a cycle of continuous learning and refinement, ensuring that these tools remain responsive to the dynamic needs of clinical practice.

Incorporating guardrail frameworks [23] represents another critical avenue for exploration. Tools, such as NVIDIA NeMo Guardrails [24] and Llama Guard [25], could be utilised to mitigate risks like hallucinations and security vulnerabilities by linking LLM outputs to trusted medical knowledge bases and enforcing safeguards against adversarial prompts. Future studies should evaluate the effectiveness of these guardrails in emergency scenarios, particularly their scalability and reliability in real-time decision support.

Equally important, regulatory innovation must also remain a central focus. As Emergency Medicine operates within broader healthcare systems, harmonised frameworks are necessary to ensure that AI tools are deployed ethically and legally. Comparative studies (e.g. [26]) across regions, examining the evolution of regulations, such as GDPR and HIPAA, could provide useful insights into best practices for balancing innovation with accountability. Addressing these challenges will enable both general healthcare and emergency-specific settings to leverage LLMs effectively while maintaining robust ethical standards.

VI. LIMITATIONS

This review highlights key themes related to the integration of LLMs in Emergency Medicine but is subject to a set of limitations. First, the literature search was confined to PubMed, which may have excluded relevant studies indexed in other databases, such as IEEE Xplore. This constraint might have limited the diversity of perspectives included, particularly those from technical or interdisciplinary journals.

Second, the review focused on recent publications to capture current advancements and challenges. While this approach ensures relevance, it may overlook foundational studies that have shaped the field over time. Future reviews could adopt a broader temporal scope to include earlier work that contextualises ongoing developments.

Additionally, the exclusion of studies on traditional NLP techniques narrows the scope to LLM-based systems. While this focus aligns with the objectives of the review, it excludes potentially valuable insights from systems employing non-LLM-based approaches in Emergency Medicine.

Finally, the analysis centred around three primary themes including ethical considerations, legal challenges, and international adoption, which although comprehensive, may not encompass all relevant aspects of LLM deployment in healthcare. For example, technical implementation challenges, such as computational resource constraints or interoperability with existing clinical systems, were not explored in detail.

VII. Conclusions

The deployment of LLMs in Emergency Medicine introduces both transformative possibilities and significant challenges. This review highlights that while LLMs can enhance operational efficiency and support decision-making in critical care settings, their integration must be approached with caution. Ethical considerations, such as the risks of biased recommendations and data privacy, underscore the need for transparent, explainable models to foster trust and safeguard patient welfare.

Legal and regulatory complexities, particularly concerning accountability and compliance with stringent data protection standards, add further challenges to LLM integration. The variability in global regulatory frameworks highlights the need for harmonised guidelines to support safer and more equitable AI adoption. Addressing these issues requires the development of context-specific models, improving model transparency, and mitigating risks such as hallucinations and security vulnerabilities. Implementing robust guardrail frameworks and clinician feedback mechanisms will be essential to ensure these AI systems are reliable and adaptable to real-world scenarios.

As LLMs continue to evolve, their integration into Emergency Medicine should be guided by a balanced approach that combines technical innovation with ethical and regulatory rigor. By addressing these challenges, healthcare systems can harness the full potential of LLMs to improve care delivery in both emergency-specific and broader healthcare contexts.

References

[1] Locke S, Bashall A, Al-Adely S, Moore J, Wilson A, Kitchen GB. Natural language processing in medicine: a review. Trends in Anaesthesia and Critical Care. 2021 Jun 1;38:4-9.

[2] Elbattah M, Arnaud É, Gignon M, Dequen G. The Role of Text Analytics in Healthcare: A Review of Recent Developments and Applications. Healthinf. 2021 Feb 11:825-32.

[3] Sax DR, Warton EM, Sofrygin O, Mark DG, Ballard DW, Kene MV, Vinson DR, Reed ME. Automated analysis of unstructured clinical assessments improves emergency department triage performance: A retrospective deep learning analysis. Journal of the American College of Emergency Physicians Open. 2023 Aug;4(4):e13003.

[4] Arnaud E, Elbattah M, Ammirati C, Dequen G, Ghazali DA. Use of artificial intelligence to manage patient flow in emergency department during the Covid-19 pandemic: a prospective, single-center study. International Journal of Environmental Research and Public Health. 2022 Aug 5;19(15):9667.

[5] Sterling NW, Patzer RE, Di M, Schrager JD. Prediction of emergency department patient disposition based on natural language processing of triage notes. International journal of medical informatics. 2019 Sep 1;129:184-8.

[6] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pretraining. OpenAI.

[7] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. In Advances in Neural Information Processing Systems (NeurIPS 2020), 33.

[8] Woo KM, Simon GW, Akindutire O, Aphinyanaphongs Y, Austrian JS, Kim JG, Genes N, Goldenring JA, Major VJ, Pariente CS, Pineda EG. Evaluation of GPT-4 ability to identify and generate patient instructions for actionable incidental radiology findings. Journal of the American Medical Informatics Association. 2024 May 23:ocae117.

[9] Patel D, Timsina P, Gorenstein L, Glicksberg BS, Raut G, Cheetirala SN, Santana F, Tamegue J, Kia A, Zimlichman E, Levin MA. Traditional Machine Learning, Deep Learning, and BERT (Large Language Model) Approaches for Predicting Hospitalizations From Nurse Triage Notes: Comparative Evaluation of Resource Management. JMIR AI. 2024 Aug 27;3(1):e52190.

[10] Lansiaux E, Baron MA, Vromant A. Navigating the landscape of medical triage: Unveiling the potential and challenges of large language models and beyond. The American Journal of Emergency Medicine. 2024 Feb 10:S0735-6757.

[11] Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. Heliyon. 2024 Feb 15.

[12] Huang T, Safranek C, Socrates V, Chartash D, Wright D, Dilip M, Sangal RB, Taylor RA. Patient-Representing Population's Perceptions of GPT-Generated Versus Standard Emergency Department Discharge Instructions: Randomized Blind Survey Assessment. Journal of Medical Internet Research. 2024 Aug 2;26:e60336.

[13] Ong JC, Seng BJ, Law JZ, Low LL, Kwa AL, Giacomini KM, Ting DS. Artificial intelligence, ChatGPT, and other large language models for social determinants of health: Current state and future directions. Cell Reports Medicine. 2024 Jan 16;5(1).

[14] Zemplényi A, Tachkov K, Balkanyi L, Németh B, Petykó ZI, Petrova G, Czech M, Dawoud D, Goettsch W, Gutierrez Ibarluzea I, Hren R. Recommendations to overcome barriers to the use of artificial intelligence-driven evidence in health technology assessment. Frontiers in Public Health. 2023 Apr 26;11:1088121.

[15] Lin YT, Deng YX, Tsai CL, Huang CH, Fu LC. Interpretable Deep Learning System for Identifying Critical Patients Through the Prediction of Triage Level, Hospitalization, and Length of Stay: Prospective Study. JMIR Medical Informatics. 2024 Apr 1;12:e48862.

[16] Sezgin E, Sirrianni J, Kranz K. Development and Evaluation of a Digital Scribe: Conversation Summarization Pipeline for Emergency Department Counseling Sessions towards Reducing Documentation Burden. medRxiv. 2023 Dec 7.

[17] Lee S, Lee J, Park J, Park J, Kim D, Lee J, Oh J. Deep learning-based natural language processing for detecting medical symptoms and histories in emergency patient triage. The American Journal of Emergency Medicine. 2024 Mar 1;77:29-38.

[18] Abi-Rafeh J, Mroueh VJ, Bassiri-Tehrani B, Marks J, Kazan R, Nahai F. Complications Following Body Contouring: Performance Validation of Bard, a Novel AI Large Language Model, in Triaging and Managing Postoperative Patient Concerns. Aesthetic Plastic Surgery. 2024 Mar;48(5):953-76.

[19] Mourby M, Cathaoir KÓ, Collin CB. Transparency of machine-learning in healthcare: The GDPR & European health law. Computer Law & Security Review. 2021 Nov 1;43:105611.

[20] Preiksaitis C, Ashenburg N, Bunney G, Chu A, Kabeer R, Riley F, Ribeira R, Rose C. The Role of Large Language Models in Transforming Emergency Medicine: Scoping Review. JMIR Medical Informatics. 2024 May 10;12:e53787.

[21] Ong JC, Chang SY, William W, Butte AJ, Shah NH, Chew LS, Liu N, Doshi-Velez F, Lu W, Savulescu J, Ting DS. Medical Ethics of Large Language Models in Medicine. NEJM AI. 2024 Jun 17:AIra2400038.

[22] Arnaud E, Elbattah M, Moreno-Sánchez PA, Dequen G, Ghazali DA. Explainable NLP model for predicting patient admissions at emergency department using triage notes. In2023 IEEE International Conference on Big Data (BigData) 2023 Dec 15 (pp. 4843-4847). IEEE.

[23] Dong Y, Mu R, Jin G, Qi Y, Hu J, Zhao X, et al. Building guardrails for large language models. arXiv preprint arXiv:2402.01822; 2024.

[24] Rebedea T, Dinu R, Sreedhar M, Parisien C, Cohen J. Nemo guardrails: A toolkit for controllable and safe LLM applications with programmable rails. arXiv preprint arXiv:2310.10501; 2023.

[25] Inan H, Upasani K, Chi J, Rungta R, Iyer K, Mao Y, et al. Llama guard: LLM-based input-output safeguard for human-AI conversations. arXiv preprint arXiv:2312.06674; 2023.

[26] Freyer O, Wiest IC, Kather JN, Gilbert S. A future role for health applications of large language models depends on regulators enforcing safety standards. The Lancet Digital Health. 2024 Sep 1;6(9):e662-72.