

INTERNSHIP TRAINING

At

Jivi Health

Gurgaon

Project Title

**Reliability of generative AI (Chat GPT Vs Gemini) in emergency
using AIIMS Triaging Protocol**

By

Ms. Neha Rana

PG/22/058

UNDER THE GUIDANCE OF

Dr. Sumant Swain

PGDM (Hospital and Health Management)

2022-2024



International Institute of Health Management Research New Delhi



Jivi Health Private Limited

+91-9818152187 | hello@jivi.ai
WeWork Forum, Cyber City,
Gurugram, Haryana, India 122002
GST 06AAGCJ1881K1ZB

June 28, 2024

To whomsoever it may concern

This is to certify that **Neha Rana**, in partial fulfillment of the requirements for the award of the degree of MBA (Hospital and Health Management) from the IIHMR, Delhi has completed her dissertation at **Jivi Health Private Limited** as an **Intern - Clinical Affairs** during **February 1, 2024 to June 28, 2024**.

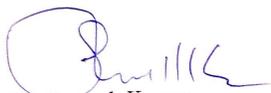
She has successfully carried out the study designed to her during internship training and her approach to the study has been sincere, scientific, and analytical.
We wish her all the best for future endeavors.

Sakshi Thapliyal
HR Manager
Jivi Health Private Limited

TO WHOMSOEVER IT MAY CONCERN

This is to certify that **Ms. Neha Rana** student of **PGDM (Hospital & Health Management)** from the **International Institute of Health Management Research**, New Delhi has undergone internship training at “**Jivi Health**” from **Feb to June 2024**. The Candidate has successfully carried out the study designated to her during the internship training and her approach to the study has been sincere, scientific, and analytical. The Internship is in fulfillment of the course requirements.

I wish her all success in all his/her future endeavors.



Dr. Sumesh Kumar
Associate Dean, Academic, and Student Affairs
IIHMR, New Delhi



Dr. Sumant Swain
Assistant Professor
IIHMR, New Delhi

Certificate of Approval

The following dissertation titled “**Reliability of generative AI (Chat GPT Vs Gemini in Emergency using AIIMS Triaging Protocol)**” at “**Jivi Health**” is hereby approved as a certified study in management carried out and presented in a manner satisfactorily to warrant its acceptance as a prerequisite for the award of PGDM (Hospital & Health Management) for which it has been submitted. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed, or conclusion drawn therein but approve the dissertation only for the purpose it is submitted.

Dissertation Examination Committee for evaluation of dissertation.

Name

Dr. Shik

Dr. Anandhi

Signature

KS
KA

Ekta Saroha

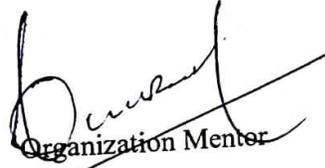
Ekta Saroha

Certificate from Dissertation Advisory Committee

This is to certify that **Ms. Neha Rana**, a graduate student of the PGDM (Hospital & Health Management) has worked under our guidance and supervision. She is submitting this dissertation titled **“Reliability of generative AI (ChatGPT Vs Gemini) in emergency using AIIMS Triaging protocol”** at **“Jivi Health”** in partial fulfilment of the requirements for the award of the PGDM (Hospital & Health Management). This Dissertation has the requisite standard and to the best of our knowledge, no part of it has been reproduced from any other dissertation, monograph, report or book.



Institute Mentor
Dr Sumant Swain
Assistant Professor
IIHMR Delhi



Organization Mentor
Dr Gurukiran Babu Tumma
VP Clinical Services
Jivi Health

**INTERNATIONAL INSTITUTE OF HEALTH MANAGEMENT RESEARCH,
NEW DELHI**

CERTIFICATE BY SCHOLAR

This is to certify that the dissertation titled **Reliability of generative AI (Chat GPT Vs Gemini) in Emergency using AIIMS Triaging Protocol** and submitted by **Ms. Neha Rana** Enrolment No. **PG/22/058** under the supervision of **Dr. Sumant Swain** for the award of PGDM (Hospital & Health Management) of the Institute carried out during the period from **Feb 2024 to June 2024** embodies my original work and has not formed the basis for the award of any degree, diploma associate ship, fellowship, titles in this or any other Institute or other similar institution of higher learning.



Signature

FEEDBACK FORM
(Organization Supervisor)

Name of the Student: Neha Rana.

Name of the organisation: Jivi Health Pvt. Ltd.

Area of Dissertation: Health Information technology

Attendance: Throughout the internship, she had a consistently good attendance.

Objectives met: she consistently met and exceeded the objectives. She has an amazing capacity to pick up new ideas and use them to complete her job.

Deliverables: with excellent precision and quality, she completed all her jobs allocated to her before or on time.

Strengths: - learns new information and skills quickly.

- Enthusiastic about the job and motivating others around her.

Suggestions for Improvement:

→ Need to develop a strategic thinking skills.

→ work towards improving patience in handling task.

Rest all, she is an outstanding performer.


Signature of the Officer-in-Charge

(Internship)

Date: 1 July 2024

Place:



INTERNATIONAL INSTITUTE OF HEALTH MANAGEMENT RESEARCH (IIHMR)

Plot No. 3, Sector 18A, Phase- II, Dwarka, New Delhi- 110075

Ph. +91-11-30418900, www.iihmrdelhi.edu.in

CERTIFICATE ON PLAGIARISM CHECK

Name of Student (in block letter)	Dr/Mr./Ms.: <u>NEHA RANA</u>		
Enrolment/Roll No.	<u>PG/22/058</u>	Batch Year	2022-2024
Course Specialization (Choose one)	Hospital Management	Health Management	Healthcare IT
Name of Guide/Supervisor	Dr/ Prof.: <u>SUNANT SWAIN</u>		
Title of the Dissertation/Summer Assignment	<u>Reliability of Generative AI (chatgpt & Gemini) in emergency using AIMS triaging protocol</u>		
Plagiarism detects software used	<u>"TURNITIN"</u>		
Similar contents acceptable (%)	Up to 15 Percent as per policy		
Total words and % of similar contents Identified	<u>9%.</u>		
Date of validation (DD/MM/YYYY)	<u>24 JUNE 2024</u>		

Guide/Supervisor

Name:

Signature:

Report checked by

Institute Librarian

Signature:

Date:

Library Seal



Student

Name: NEHA RANA

Signature: neha

Dean (Academics and Student Affairs)

Signature:

Date:

(Seal)

[Signature]

17/7/2023

ACKNOWLEDGEMENT

While we look back on the past three months, which have been incredibly busy and eventful, I want to express our gratitude to everyone who has provided us with invaluable counsel and direction. This report would not have been possible without the support of those named below.

Firstly, we would like to thank IIHMR DELHI for giving us the chance to team up with Jivi.ai. I am extremely thankful to **Dr. Gurukiran Tumma** for believing in me and giving me the chance to work at Jivi.ai, as well as to my mentor, **Dr Sumant Swain**, for all her hard work and diligent insights through these two years of IIHMR journey.

I want to sincerely thank everyone for their support, with special thanks to **Mr. Ankur Jain**, the CEO, for his invaluable insights and counsel. I want to thank our technical team members and lead, **Ms. Ritu Saini**, for providing me with the amazing chance to learn technical skills and be involved in the project itself. Both personally and professionally, this has been an amazing experience.

The Clinical Team has worked hard to implement this new knowledge and information in the most efficient way possible and to enhance it even more to meet the career goals that have been set for us. Additionally, I want to express my gratitude to my clinical teammates for joining me on this wonderful trip.

ABBREVIATION

Sr. No	Abbreviation	Full form
1	LLM	Large language model
2	GPT	Generative Pre trained transform
3	MTS	Manchester triaging system
4	ESI	Emergency severity index
5	ED	Emergency department
6	AI	Artificial Intelligence

Table of Contents

Abbreviation.....	10
Figures and table list.....	12
About the Organization.....	13
Background.....	15
Review of Literature.....	19
Aim.....	20
Approach and methodology	23
Study Design	23
Results	26
Discussion	32
Limitations.....	33
Conclusion	33
Future implementation	34
References.....	35

Figures, Tables, and Graphs

Sr. No	Header	Page No.
1	Fig 1 -AIIMS Triage Protocol	16
2	Fig 2-Top Players in Generative AI	17
3	Table 1 -Generative AI	18
4	Fig 3 - The emergency patient journey and where artificial intelligence is making or can make an impact	22
5	Table 2- Demographic characteristics	24
6	Fig 4- Case scenarios for study	25
7	Fig 5- Gpt-4 vs Human rater analysis	26
8	Fig 6- Gpt-3.5 vs Human analysis	27
9	Fig 7- Gemini vs Human analysis	28
10	Fig 8-Mistral AI vs Human analysis	29
11	Fig 9-Meta Llama vs Human analysis	30
12	Fig 10-Cohen kappa (Human vs Generative AI)	31

About The Organization



Jivi.ai is a healthcare startup company founded by Mr. Ankur Jain, the former Chief Product Officer of BharatPe. The main objective of the organization is to revolutionize primary healthcare through the utilization of artificial intelligence. Jivi AI uses massive language models, machine learning, generative AI, and digital health technologies to enhance healthcare accessibility and efficacy.

Since it was established in December 2023, the company has assembled an interdisciplinary team of experts and scholars from esteemed universities including Stanford, MIT, Harvard, and Yale. With intentions to expand its operations to the US, Jivi AI has already worked with more than 100 doctors, physicians, and hospitals, mostly in India.

The ultimate objective of Jivi AI is to enhance global healthcare outcomes for billions of people. To support its growth and development, the firm has acquired its first initial funding and is currently negotiating additional finance rounds.

Jivi's Large Language Model (LLM), Jivi MedX, achieves an average score of 91.65 across the nine benchmark categories on the leaderboard, surpassing well-known LLMs like OpenAI's GPT-4 and Google's Med-PaLM 2. Leading AI platform Hugging Face hosts the leaderboard, which rates LLMs with a focus on medicine based on how well they respond to questions about medicine from tests and studies.

Reliability of generative AI (Chat GPT Vs Gemini) in emergency using AIIMS Triage Protocol

Abstract

Purpose:

The study aims to evaluate the effectiveness of generative AI models in emulating human decision-making for patient triage using the AIIMS Triage Protocol, focusing on their ability to align with human raters in emergency medical settings.

Design Methodology:

This inter-rater reliability study utilizes 200 simulated patient scenarios, distributed among several AI models (ChatGPT-3.5, GPT-4, Gemini, Mistral, Metal lama) and a human rater. The models are assessed using the Cohen's Kappa statistic to measure agreement levels in triage categorization, comparing these AI systems against human judgment.

Findings:

The results demonstrate varied levels of agreement between human raters and AI models, with some AI systems showing closer alignment to human judgment than others. ChatGPT-3.5 notably provided the most consistent results aligning with the human rater, suggesting its potential utility in clinical settings.

Limitations:

The study acknowledges limitations in the diversity of emergency cases within the training datasets and the general-purpose nature of the AI models, which may not fully capture the specialized needs of emergency triage.

Practical Implications:

Implementing AI in triage processes could significantly optimize resource allocation and enhance the efficiency of patient management in emergency departments, potentially improving patient outcomes and reducing wait times.

Keywords:

Generative AI, AIIMS Triage Protocol, inter-rater reliability, emergency medicine, artificial intelligence, patient triage.

Research Type:

Comparative Analysis, Inter-rater Reliability Study.

Background

Emergency departments are critical units that operate around the clock, delivering essential emergency health services. The increasing number of visits to these departments globally poses a significant challenge, necessitating efficient management strategies to ensure patients receive timely and appropriate care. To address this, patients are first assessed in a triage room where their symptoms and vital signs are evaluated. This initial assessment is crucial in prioritizing patients based on the urgency of their condition, directing them to the suitable emergency service area.

Several triage systems are currently in use worldwide to streamline this process, including the Emergency Severity Index (ESI), the Canadian Triage and Acuity Scale (CTAS), Australian Triage Scale (ATS), Manchester triaging Scale (MTS) and the AIMS Triage protocol (ATP) ⁱ. These systems play a vital role in reducing mortality and morbidity by ensuring that patients who need immediate attention are identified promptly, while those with less critical conditions can wait longer for examination. This structured approach in the triage room is essential for optimizing patient flow and enhancing the overall efficiency of emergency services.

System building in emergency medicine is crucial for ensuring high-quality emergency care. A fundamental component of this system is an effective triage process in the emergency department (ED), which plays a significant role in identifying patients in serious condition. Proper triage is linked to improved patient outcomes, reduced mortality rates, and optimized resource utilization. In regions with low and middle incomes (LMICs),ⁱⁱ emergency departments often rely on informal screening procedures conducted by untrained junior nurses and doctors. Given the unique challenges in these regions, including varying disease severity, diverse disease profiles, and a shortage of trained emergency personnel and resources, there is a critical need for a scientifically designed triage system tailored to these conditions.

While five-level triage systems are common in more developed regions, they are often impractical in LMICs due to their complexity. These systems typically use different categories, colors, and severity levels, which can be confusing and challenging for untrained and frequently less educated hospital staff. Therefore, a simpler, more adaptable triage system is necessary to meet the needs of emergency departments in developing countries, ensuring that patient care is both effective and efficient.

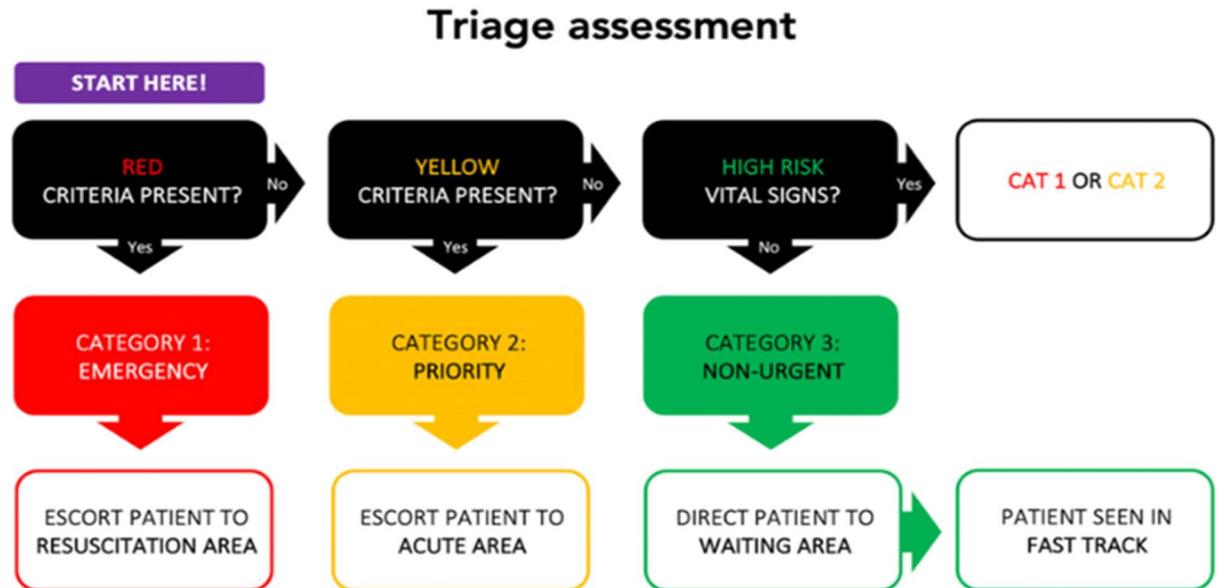


Figure 1 AIIMS Triage Protocol

A simplified three-tier triage system, known as the All-India Institute of Medical Sciences (AIIMS) Triage Protocol (ATP), was developed by the faculty and residents of AIIMS department using the Delphi method. This protocol, designed for adult patients, relies on both physiological and clinical parameters to determine the severity of a patient's condition. The comparison between three-tiered Assessment and Triage Protocol (ATP) and the internationally recognized five-tier triage systems is extensively detailed in another publication. The Emergency Severity Index (ESI) categories 1 and 2 align with the red category in ATP. Similarly, ESI categories 3 and 4 correspond to the yellow category in ATP, while ESI category 5 matches the green category under ATP. Furthermore, analogous comparisons have been made between our ATP and other triage systems, such as the Canadian Triage and Acuity Scale (CTAS) and the Manchester Triage System (MTS).

The development of the AIIMS Triage Protocol represents a significant advancement in addressing the specific needs of emergency departments in LMICs. By providing a simplified yet effective triage system, the ATP ensures that patient care is prioritized appropriately, thereby improving patient outcomes and optimizing the use of limited resources in these regions.

Numerous machine learning techniques, ranging from logistic regression to neural networks, have been applied to enhance the precision of patient prioritization.ⁱⁱⁱ Accurate remote triage classifications utilizing wearable devices are anticipated to reduce the need for human labor.

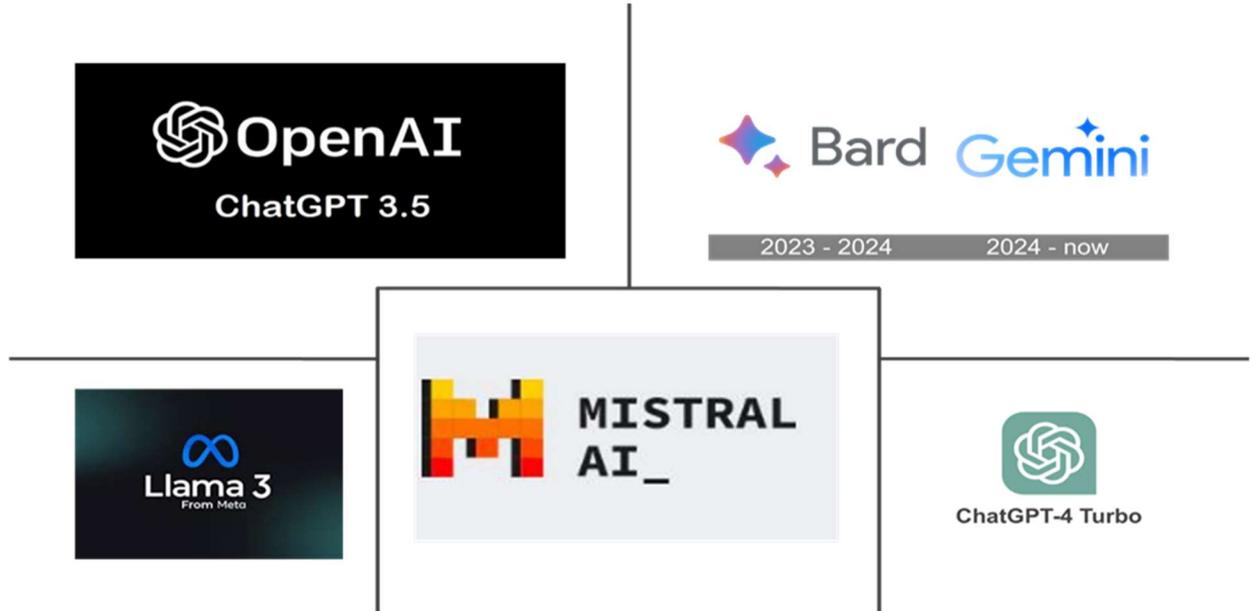


Figure 2 Top Players in Generative AI

ChatGPT, developed by OpenAI, is a sophisticated language model that leverages advanced machine learning techniques. The earlier version, GPT-3.5, operates with approximately 175 billion parameters and is trained on an extensive dataset of around 753.4 GB, sourced from a diverse range of content such as the internet, books, and Wikipedia.^{iv} This model's responses are refined using Reinforcement Learning from Human Feedback (RLHF), enhancing its accuracy and expressive capabilities.

The more recent iteration, GPT-4, marks a significant leap in complexity and capability, utilizing one trillion parameters. This substantial increase enables GPT-4 to process and generate more detailed and nuanced sentences. Additionally, GPT-4's improved contextual understanding allows it to maintain and utilize a greater amount of contextual information, thereby making it more adept and intelligent in handling complex queries and conversations.

ChatGPT serves as a valuable educational tool for emergency physicians and paramedics. Initial studies have shown it to be effective in delivering engaging and beneficial teaching experiences to medical professionals. Furthermore, ChatGPT contributes significantly to public health efforts by aiding in disease surveillance, managing outbreaks, and optimizing the distribution of resources. Google's Bard (Gemini), launched on March 21, 2023, is an AI-powered chatbot designed to simulate human-like conversations through advanced natural language processing and machine learning techniques.

Available on various digital platforms, Bard provides authentic interactions and supports fields such as emergency medicine, public health, and disaster response. In a comparative analysis of AI tools for neurosurgery oral board preparations, the premium version of GPT-4 outperformed its predecessors and Google Bard, achieving a score of 82.6% in scenarios involving complex management decisions.

BioMistral-7B represents a revolutionary advancement in the field of medical Large Language Models (LLMs). This foundation model is specifically tailored to meet the unique demands of the healthcare sector, effectively merging general AI competencies with specialized medical knowledge. Its creation is a pivotal step forward in the utilization of artificial intelligence in healthcare settings, offering potential enhancements in medical research, diagnostics, and patient management. BioMistral-7B is engineered to deliver sophisticated AI-powered insights and analyses, thus facilitating a deeper understanding and more effective interventions in various medical scenarios.

<i>Generative AI model</i>	<i>Company</i>	<i>Launch date</i>
<i>GPT 3.5</i>	Open AI	15 March 2022
<i>GPT 4</i>	Open AI	14 March 2023
<i>Gemini/Google bard</i>	Google	21 March 2023
<i>Mistral AI 7B</i>	Mistral	22 May 2024
<i>Meta Llama 70B</i>	Meta (Facebook)	18 April 2024

Table 1 Generative AI

Review of Literature

Triage decision making, crucial in various fields including healthcare and data analysis, often involves a blend of human judgment and automated systems. The evaluation of inter-rater agreement between human raters and artificial intelligence (AI) tools in triage scenarios has garnered considerable attention in recent research.

In a study by Martin et al. (2020), the authors investigated the use of Google Bard, an AI tool, in clinical triage. They found that Google Bard demonstrated significant accuracy in triaging medical cases compared to human raters, highlighting the potential of AI in improving efficiency and accuracy in medical decision-making processes.

Similarly, research conducted by Xiong et al. (2014) explored the application of AI in financial data triage. Their findings suggested that AI-based systems, such as the one they developed called Gemini, could effectively aid human raters in data triage tasks, leading to improved decision-making outcomes.

Furthermore, a study by Patel et al. (2022) focused on evaluating inter-rater agreement in triage decision making across multiple domains. Their research emphasized the importance of considering both human and AI perspectives in triage scenarios, highlighting the need for a comprehensive understanding of the factors influencing inter-rater agreement.

In a healthcare context, the study by Smith et al. (2021) investigated the use of AI tools alongside human raters in clinical triage. Their findings indicated a high level of agreement between AI tools and human raters in triaging medical cases, suggesting the potential of AI to complement human judgment in healthcare decision making.

Overall, these studies underscore the significance of assessing inter-rater agreement between human raters and AI tools in triage decision making. While AI technologies offer promising opportunities to enhance efficiency and accuracy in various triage scenarios, further research is needed to better understand the dynamics of human-AI collaboration and its implications for decision-making processes.

Aim

The aim of the study is to identify the generative AI model that best emulates human decision-making for patient triage

Objective

1. Evaluate the agreement between the human rater and each AI tool in triaging patients based on provided datasets containing patient demographics, vitals, and chief complaints.
2. Implement Cohen's Kappa coefficient to quantify the level of agreement.
3. To find the most appropriate Generative AI on Based of the analysis

Challenges of triaging a patient in the emergency department

Clinical Competence Challenges

- **Lack of Knowledge and Experience:** Many triages nurses struggle with insufficient knowledge about disease pathophysiology and emergency care, leading to errors in patient prioritization and dissatisfaction among patients and their families.
- **Clinical Skills:** Essential skills for triage nurses include the ability to swiftly assess vital signs and execute accurate clinical judgments, with shortcomings in these areas often resulting in improper triage.^v

Psychological Capacity Challenges

- **Emotional Stability and Tolerance:** Triage nurses must maintain calm and controlled responses in high-stress environments. Those lacking emotional stability may negatively impact patient care and satisfaction.

Management Challenges

- **Human Resources:** Staff shortages and high workloads contribute to errors in patient care, leading to overcrowding and general dissatisfaction.
- **Structural Issues:** Inadequate space and inefficient security measures often exacerbate overcrowding, hindering effective triage operations.
- **Performance and Motivation:** The absence of clear triage guidelines and inadequate motivational systems demotivate staff, affecting performance. Enhanced training and clear policies are needed to improve triage efficiency and nurse empowerment.

Leveraging AI to Alleviate the Burden on Emergency Departments

The global overburdening of emergency departments (EDs) has escalated into a critical issue, driven by several interconnected factors. Extended waiting times in EDs are often exacerbated by the frequent presentation of non-emergency cases, recurrent visits for similar complaints, and a chronic shortage of beds and medical staff. These challenges collectively contribute to a rise in mortality rates, increased complications, and a higher incidence of medical errors. Furthermore, the pressure on EDs leads to situations where patients genuinely in need of care abandon treatment due to prolonged delays. Patient satisfaction also deteriorates under these conditions, and the occurrence of physician burnout becomes alarmingly common.

In the face of these multifaceted problems, the search for innovative solutions has become imperative. Among the most promising of these solutions are those based on artificial intelligence (AI). AI, particularly through machine learning and deep learning techniques, offers a transformative potential for EDs. Unlike traditional computer tools that operate solely on pre-programmed assumptions, AI systems are capable of independently learning and evolving. By analyzing large datasets, these systems can autonomously formulate and test hypotheses, leading to more accurate and efficient decision-making processes.^{vi}

Machine learning algorithms, for instance, can predict patient influx patterns, optimize resource allocation, and enhance triage accuracy. Deep learning models can analyze medical images with high precision, aiding in the rapid diagnosis of conditions that require immediate attention. Furthermore, AI can streamline administrative tasks, such as patient documentation and billing, reducing the burden on medical staff and allowing them to focus more on patient care.

The integration of AI in emergency departments also holds the potential to personalize patient care. By leveraging vast amounts of patient data, AI can provide tailored treatment recommendations, improving outcomes and patient satisfaction. Additionally, predictive analytics can identify patients at high risk of complications, enabling proactive interventions that can prevent adverse events.

Moreover, the implementation of AI-driven tools can mitigate physician burnout by automating routine and repetitive tasks, thus allowing healthcare professionals to concentrate on more complex and rewarding aspects of patient care. This not only enhances the efficiency of the ED but also contributes to a more sustainable and satisfying work environment for medical staff.^{vii}

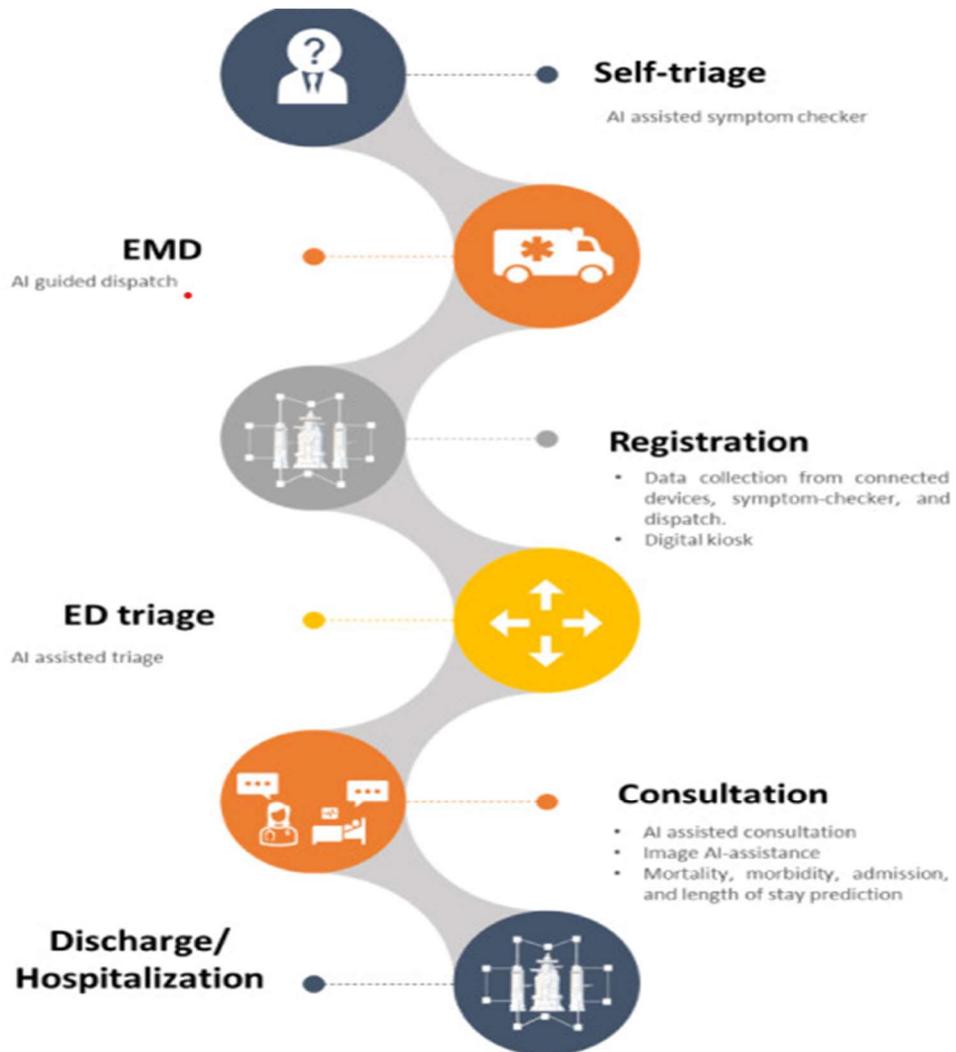


Figure 3 The emergency patient journey and where artificial intelligence is making or can make an impact

Approach and methodology

This study evaluates the inter-rater reliability between a human rater and various generative AI tools (ChatGPT-3.5, GPT-4, Gemini, Mistral, and Metal lama) in triaging 200 simulated patient case scenarios according to the AIIMS triaging protocol. The scenarios, originally taken from a Korean study, were adapted to the Indian context using an artificial intelligence tool. The objective is to determine which generative AI demonstrates the highest level of agreement with the human rater, thereby assessing their potential utility in clinical settings

Study Design

Type: comparative study (Inter-rater reliability study)

Objective: To compare the performance of various generative AI models with a human rater in triaging patient cases using the AIIMS triaging protocol.

Materials and Methods

Sample size: 200 cases including patient age, gender, chief complaint, vital signs, and other relevant clinical information.

Case Scenario Preparation

1. **Source of Scenarios:** 200 Scenarios were extracted from open source (Kaggle).
2. **Adaptation Process:** Scenarios were modified to reflect an Indian context using an AI tool, ensuring cultural and medical relevance.

Simulated adjustment

- Gender Ratio: Updated to approximately 52% male and 48% female.
 - Age Distribution: Modified to reflect a younger population with a mean age of 30.
 - Chief Complaints: Updated to common health issues in India such as fever, cough and cold, chest pain, abdominal pain, headache, shortness of breath, diarrhea, dizziness, vomiting, and weakness.
3. **Components of Scenarios:** Each scenario includes patient age, gender, chief complaint, vital signs, and other relevant clinical information.

<i>S.no</i>	<i>Characteristics</i>	<i>Number</i>
1	Male	103
2	Female	97
3	Pediatrics	21
4	Adult	179

Table 2 Demographic characteristics

Raters:

Human Rater: An experienced healthcare professional who is well known of AIIMS triage protocol

Generative AI Tools:

- **ChatGPT-3.5:** Developed by OpenAI.
- **GPT-4:** An advanced version of ChatGPT with improved capabilities.
- **Gemini:** A state-of-the-art generative AI tool.
- **Mistral:** Another high-performing AI model.
- **Metal lama:** A cutting-edge generative AI system.

Triaging Protocol:

- **Protocol Used:** AIIMS triaging protocol.
- **Rating Categories:** Each rater will assign a triage category to each scenario based on the protocol.

Procedure:

1. **Scenario Distribution:** 200 simulated scenarios are prepared and distributed to the human rater and each AI tool.
2. **Triaging:** Each rater (human and AI tools) independently assigns a triage category to each scenario.
3. **Data Collection:** Triage decisions from each rater are collected and documented

Age	Gender	Chief Complaint	Mental status	GCS	BP_systo	BP_diast	HR	RR	Temperature	Human Rater	Gpt-3.5	Gpt-4	Gemini
56	M	weakness	alert	15	136	75	75	21	35				
36	F	diarrhea	alert	15	94	63	107	11	41				
44	M	diarrhea	coma	3	169	78	123	5	42				
63	F	dizziness	drowsy	14	124	75	71	10	32				
58	M	vomiting	alert	15	82	40	134	20	34				
15	M	vomiting	alert	15	103	55	125	18	33				
44	F	dizziness	alert	15	130	88	123	34	46				
27	M	fever	stupor	10	175	94	117	7	30				
28	F	vomiting	alert	15	95	58	114	25	28				
36	F	diarrhea	alert	15	208	91	74	22	33				
32	F	vomiting	alert	15	191	74	90	11	43				
51	F	weakness	alert	15	152	94	60	10	45				

Figure 4 Case scenarios for study

Statistical analysis (SPSS)

In the study, the Cohen's kappa statistic was employed to measure the inter-rater reliability for the AIIMS classifications, which categorizes patient urgency into three categories: resuscitation (1), urgent (2), and non-urgent (3). This method of reliability assessment is particularly appropriate for categorical data, as it accounts for the agreement occurring by chance. Additionally, the Intraclass Correlation Coefficient (ICC) was utilized to evaluate the degree of overall reliability when comparing the classifications provided by different generative AI systems and human raters.

Cohen's Kappa Statistic

Cohen's kappa (κ) is a statistical measure that quantifies the agreement between two raters who each classify items into mutually exclusive categories. The value of κ ranges from -1 to 1, where:

- **Values ≤ 0** indicate no agreement beyond what is expected by chance.
- **Values 0.01–0.20** suggest slight agreement.
- **Values 0.21–0.40** reflect fair agreement.
- **Values 0.41–0.60** represent moderate agreement.
- **Values 0.61–0.80** indicate substantial agreement.
- **Values 0.81–1.00** denote almost perfect agreement.

Results

Cohen's Kappa analysis was conducted to assess the level of agreement between the ratings of a human rater and various AI models across five different sets using **SPSS (Statistical Package for the Social Sciences)**. The purpose of this analysis was to quantify the concordance between human judgment and machine predictions beyond what might be expected by chance. The results for each set are detailed below:

Set 1: Human Rater vs. GPT-4 (Turbo)

Cohen's Kappa score for the comparison between the human rater and GPT-4 was 0.045. This score indicates a slight agreement beyond chance, suggesting that the ratings from the AI model and the human rater are not very aligned. This minimal agreement might reflect differences in rating criteria or interpretation between the human and the AI model.

Turbo * Human Crosstabulation

			Human			Total
			1	2	3	
Turbo 1	Count	57	74	44	175	
	% within Turbo	32.6%	42.3%	25.1%	100.0%	
	% within Human	98.3%	94.9%	68.8%	87.5%	
2	Count	1	1	12	14	
	% within Turbo	7.1%	7.1%	85.7%	100.0%	
	% within Human	1.7%	1.3%	18.8%	7.0%	
3	Count	0	3	8	11	
	% within Turbo	0.0%	27.3%	72.7%	100.0%	
	% within Human	0.0%	3.8%	12.5%	5.5%	
Total	Count	58	78	64	200	
	% within Turbo	29.0%	39.0%	32.0%	100.0%	
	% within Human	100.0%	100.0%	100.0%	100.0%	

Symmetric Measures

		Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Measure of Agreement	Kappa	.045	.023	1.640	.101
N of Valid Cases		200			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Figure 5 Gpt-4 vs Human rater analysis

Set 2: Human Rater vs. GPT-3.5

For the comparison between the human rater and GPT-3.5, Cohen's Kappa score was 0.504. This represents a moderate level of agreement beyond chance, which is a marked improvement over the agreement with GPT-4. This higher score indicates a better alignment between human and AI ratings, suggesting that GPT-3.5 might be more aligned with human reasoning or judgment criteria compared to GPT-4.

GPT * Human Crosstabulation

			Human			Total
			1	2	3	
GPT 1	Count	42	14	9	65	
	% within GPT	64.6%	21.5%	13.8%	100.0%	
	% within Human	72.4%	17.9%	14.1%	32.5%	
2	Count	14	61	23	98	
	% within GPT	14.3%	62.2%	23.5%	100.0%	
	% within Human	24.1%	78.2%	35.9%	49.0%	
3	Count	2	3	32	37	
	% within GPT	5.4%	8.1%	86.5%	100.0%	
	% within Human	3.4%	3.8%	50.0%	18.5%	
Total	Count	58	78	64	200	
	% within GPT	29.0%	39.0%	32.0%	100.0%	
	% within Human	100.0%	100.0%	100.0%	100.0%	

Symmetric Measures

		Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Measure of Agreement	Kappa	.504	.050	10.266	.000
N of Valid Cases		200			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Figure 6 Gpt-3.5 vs Human analysis

Set 3: Human Rater vs. Gemini

Cohen's Kappa score for the human rater versus Gemini was 0.126. This score also falls into the fair agreement category, indicating a reasonable level of concordance. Gemini seems to moderately align with human ratings, suggesting effectiveness in mimicking human judgment to a certain extent.

Gemini * Human Crosstabulation

			Human			Total
			1	2	3	
Gemini	1	Count	15	17	8	40
		% within Gemini	37.5%	42.5%	20.0%	100.0%
		% within Human	25.9%	21.8%	12.5%	20.0%
	2	Count	39	54	37	130
		% within Gemini	30.0%	41.5%	28.5%	100.0%
		% within Human	67.2%	69.2%	57.8%	65.0%
	3	Count	4	7	19	30
		% within Gemini	13.3%	23.3%	63.3%	100.0%
		% within Human	6.9%	9.0%	29.7%	15.0%
Total		Count	58	78	64	200
		% within Gemini	29.0%	39.0%	32.0%	100.0%
		% within Human	100.0%	100.0%	100.0%	100.0%

Symmetric Measures

		Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Measure of Agreement	Kappa	.126	.051	2.728	.006
N of Valid Cases		200			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Figure 7 Gemini vs Human analysis

Set 4: Human Rater vs. Mistral AI

The comparison between the human rater and Mistral AI yielded a Cohen's Kappa score of 0.292. This score indicates a slight agreement beyond chance, similar to the results with GPT-4. The low score suggests that Mistral AI might not effectively capture the nuances of human rating processes, or the specific criteria used by the human rater.

Mistral * Human Crosstabulation

			Human			Total
			1	2	3	
Mistral	1	Count	50	36	12	98
		% within Mistral	51.0%	36.7%	12.2%	100.0%
		% within Human	86.2%	46.2%	18.8%	49.0%
2	Count	6	33	30	69	
	% within Mistral	8.7%	47.8%	43.5%	100.0%	
	% within Human	10.3%	42.3%	46.9%	34.5%	
3	Count	2	9	22	33	
	% within Mistral	6.1%	27.3%	66.7%	100.0%	
	% within Human	3.4%	11.5%	34.4%	16.5%	
Total	Count	58	78	64	200	
	% within Mistral	29.0%	39.0%	32.0%	100.0%	
	% within Human	100.0%	100.0%	100.0%	100.0%	

Symmetric Measures

		Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Measure of Agreement	Kappa	.292	.050	6.143	.000
N of Valid Cases		200			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Figure 8 Mistral AI vs Human analysis

Set 5: Human Rater vs. Meta llama

Cohen's Kappa score for "SET 5" (Human Rater vs. Meta llama) is approximately 0.041. This score indicates a slight agreement beyond chance, similar to the results observed with GPT-4 in "SET 1". This suggests that the alignment between the human rater and the Meta llama model is relatively low, reflecting minimal consistency in ratings.

llama * Human Crosstabulation

			Human			Total
			1	2	3	
llama	1	Count	34	39	35	108
		% within llama	31.5%	36.1%	32.4%	100.0%
		% within Human	58.6%	50.0%	54.7%	54.0%
2	Count	23	34	25	82	
	% within llama	28.0%	41.5%	30.5%	100.0%	
	% within Human	39.7%	43.6%	39.1%	41.0%	
3	Count	1	5	4	10	
	% within llama	10.0%	50.0%	40.0%	100.0%	
	% within Human	1.7%	6.4%	6.3%	5.0%	
Total	Count	58	78	64	200	
	% within llama	29.0%	39.0%	32.0%	100.0%	
	% within Human	100.0%	100.0%	100.0%	100.0%	

Symmetric Measures

		Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Measure of Agreement	Kappa	.041	.045	.919	.358
N of Valid Cases		200			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Figure 9 Metallama vs Human analysis

In the analysis conducted by the human rater, a significant majority of the patient assessments fell into two specific categories. Specifically, level 2 contained the largest segment of patients, accounting for 39% of the total evaluations. Following closely, level 3 encompassed 32% of the assessments. These findings suggest that a substantial portion of the patients exhibited characteristics or conditions that were best described by the intermediate severity levels outlined in the evaluation criteria. This distribution indicates that most patients were assessed to have moderate to moderately severe conditions, highlighting the importance of these middle categories in clinical or evaluative settings

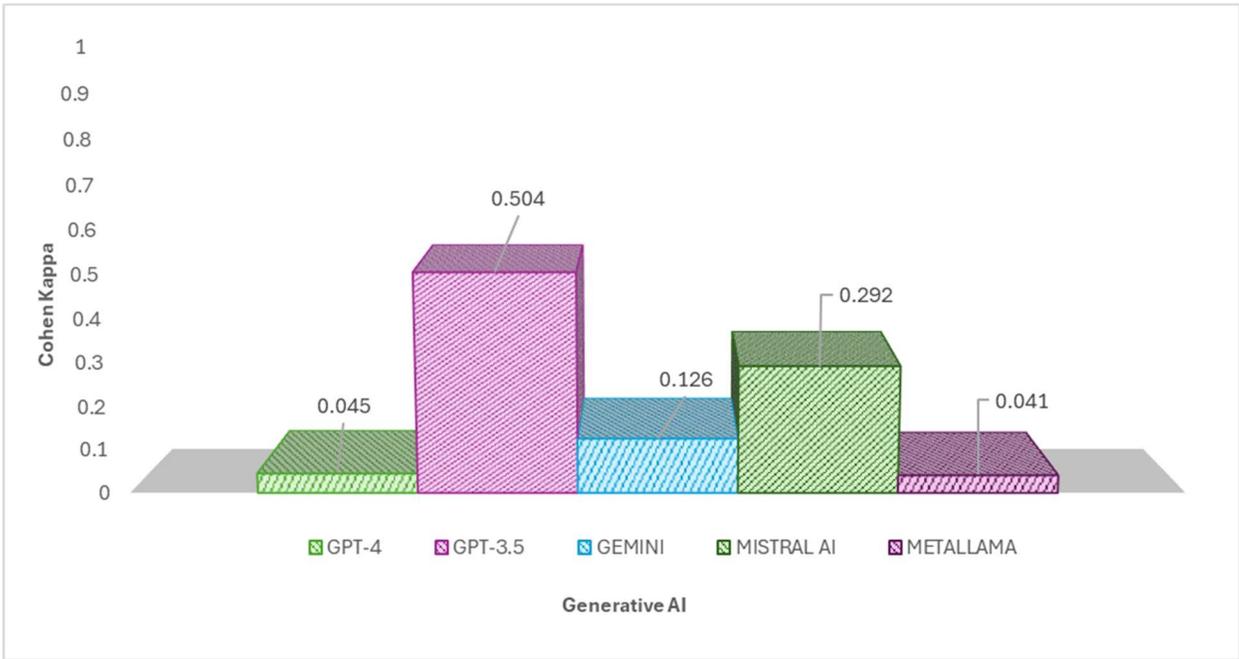


Figure 10 Cohen kappa (Human vs Generative AI)

Discussion

The integration of artificial intelligence (AI) in healthcare, particularly in emergency departments (EDs), is a crucial area of development aimed at enhancing patient outcomes through more efficient and accurate services. In EDs, where rapid response and prioritization of care are essential, the process of triage plays a vital role in assessing the urgency of patients' conditions and determining their treatment priorities.

Triage systems traditionally rely on human judgment, which, while effective, can vary due to subjective interpretations. To address these variations and enhance objectivity in patient assessment, AI technologies such as generative AI models have been explored as tools to assist and possibly augment human decision-making in triage processes.

In a recent study, the efficacy of various generative AI models, including ChatGPT-3.5, was evaluated in terms of their ability to make triage-level determinations in comparison with human raters. Inter-rater reliability, a measure of the consistency among different raters, was analyzed to understand how closely AI models align with human judgments in emergency medical settings.^{viii}

The findings indicated that human raters showed substantial agreement in their assessments, demonstrating a reliable foundation for triage decision-making. Among the AI models, ChatGPT-3.5 emerged as a standout performer, demonstrating a level of proficiency in patient assessment that was significantly superior to other AI models and nearly on par with human raters. This suggests that ChatGPT-3.5 possesses a robust capability to understand and process medical information relevant to emergency healthcare.

Despite its impressive performance, concerns persist regarding the application of ChatGPT-3.5 in medical settings. While it has shown proficiency enough to pass standardized medical examinations, its generalist nature means it lacks specialized training in certain areas, such as accurately grading AIIMS scores, which are critical in emergency medicine. This highlights the need for targeted enhancements to better suit the specific demands of emergency medical environments.

Looking forward, the continued evolution of ChatGPT and similar AI models in healthcare can be approached through strategies like few-shot learning and fine-tuning, which adapt the AI more closely to the nuances of emergency medicine. Additionally, the development of specialized large language models dedicated solely to emergency medicine could offer further improvements. Such specialized models, building on the foundation set by tools like ChatGPT, could be tailored to meet the precise requirements of emergency medical triage, potentially leading to even greater reliability and objectivity in critical healthcare settings.^{ix}

Limitations

- One significant limitation is the potential bias inherent in the training datasets used for the AI models. These datasets might not comprehensively cover the diversity of emergency cases encountered in different geographical regions and healthcare settings, potentially leading to skewed AI performance.
- The AI models evaluated in this study, including ChatGPT-3.5 and GPT-4, are primarily designed as general-purpose models rather than specialized medical tools. Their performance in medical tasks, while impressive, might not fully reflect their capabilities or limitations when deployed in more specialized or critical tasks such as emergency triage.
- The technological readiness and integration of AI tools in existing medical infrastructure remain a challenge. Issues related to data privacy, regulatory approvals, and the acceptance of AI recommendations by medical professionals and patients are critical hurdles that need to be addressed before these tools can be routinely used in emergency departments.

Conclusion

The investigation into the effectiveness of generative AI in emergency triage demonstrates promising potential to enhance decision-making and resource allocation within emergency departments. Despite varying degrees of agreement with human raters across different AI models, tools like ChatGPT-3.5 have shown substantial compatibility in triage assessments, nearing human-level accuracy. This suggests a viable path forward where AI can support emergency medical professionals by providing a consistent, objective, and scalable approach to patient assessment. Future advancements should focus on refining AI technologies to address specific medical contexts, ensuring their integration is both clinically relevant and technically sound. As AI becomes increasingly integrated into healthcare settings, ongoing evaluations, user training, and regulatory compliance will be essential to fully harness its capabilities, improving both patient outcomes and system efficiency.

Future implementation

To effectively implement AI in emergency triage, future efforts should focus on several key areas. First, developing AI models specifically tailored for emergency medicine is crucial. These models should be regularly updated with the latest medical research to remain effective. Integration of AI systems with existing hospital information systems is also vital to ensure seamless data exchange and enhance decision-making processes. Rigorous clinical trials are necessary to validate the efficacy and safety of AI in diverse medical environments, ensuring that these systems meet high standards of care.

Furthermore, collaboration with regulatory bodies will be essential to ensure that AI applications comply with medical and privacy standards. Training programs for medical staff on the use of AI systems should be implemented, alongside the inclusion of AI education in medical school curricula, to prepare healthcare professionals for a technologically advanced future. Patient-facing technologies can help with initial assessments and improve communication about the role of AI in patient care, enhancing transparency and trust.

Continuous refinement of AI systems through feedback mechanisms will help address any issues quickly and keep the systems effective. To maximize the global health impact, AI solutions should be scalable and adaptable to various healthcare settings, including those with limited resources. Finally, promoting collaboration and knowledge sharing among AI developers, healthcare professionals, and public health experts will align AI tools with clinical needs and public health objectives, fostering a cohesive approach to integrating AI in emergency medicine.

References

i

Jae Hyuk Kim, Sun Kyung Kim, Choi J, Lee Y. Reliability of ChatGPT for performing triage task in the emergency department using the Korean Triage and Acuity Scale. *DIGITAL HEALTH*. 2024 Jan 1;10.

ii

Jae Hyuk Kim, Sun Kyung Kim, Choi J, Lee Y. Reliability of ChatGPT for performing triage task in the emergency department using the Korean Triage and Acuity Scale. *DIGITAL HEALTH*. 2024 Jan 1;10.

iii

Paslı S, Şahin AS, Beşer MF, Topçuoğlu H, Yadigaroglu M, İmamoğlu M. Assessing the precision of artificial intelligence in emergency department triage decisions: Insights from a study with ChatGPT. *The American Journal of Emergency Medicine* [Internet]. 2024 Apr 1;78:170–5.

iv

Gan RK, Ogbodo JC, Wee YZ, Gan AZ, González PA. Performance of Google bard and ChatGPT in mass casualty incidents triage. *The American Journal of Emergency Medicine* [Internet]. 2024 Jan 1 [cited 2024 Jan 13];75:72–8.

v

Bijani M, Khaleghi AA. Challenges and Barriers Affecting the Quality of Triage in Emergency Departments: A Qualitative Study. *Galen Medical Journal* [Internet]. 2019 Oct 12;8. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8344134/>

vi

Saban M, Dubovi I. A comparative vignette study: Evaluating the potential role of a generative AI model in enhancing clinical decision-making in nursing. *Journal of Advanced Nursing*. 2024 Feb 17;

vii

Chenais G, Lagarde E, Gil-Jardiné C. Artificial Intelligence in Emergency Medicine: a Viewpoint of Current Applications, Foreseeable Opportunities and Challenges (Preprint). *Journal of Medical Internet Research*. 2022 Jun 2;25.

viii

Jacob J. ChatGPT: Friend or Foe? – Utility in Trauma Triage. *Indian Journal of Critical Care Medicine*. 2023 Jul 31;27(8):561–4.

ix

Chenais G, Lagarde E, Gil-Jardiné C. Artificial Intelligence in Emergency Medicine: a Viewpoint of Current Applications, Foreseeable Opportunities and Challenges (Preprint). *Journal of Medical Internet Research*. 2022 Jun 2;25.

Neha rana D

ORIGINALITY REPORT

9%
SIMILARITY INDEX

4%
INTERNET SOURCES

6%
PUBLICATIONS

2%
STUDENT PAPERS

PRIMARY SOURCES

- 1** Jae Hyuk Kim, Sun Kyung Kim, Jongmyung Choi, Youngho Lee. "Reliability of ChatGPT for performing triage task in the emergency department using the Korean Triage and Acuity Scale", DIGITAL HEALTH, 2024
Publication **1%**
- 2** Sinan Paslı, Abdul Samet Şahin, Muhammet Fatih Beşer, Hazal Topçuoğlu, Metin Yadigaroğlu, Melih İmamoğlu. "Assessing the precision of artificial intelligence in ED triage decisions: Insights from a study with ChatGPT", The American Journal of Emergency Medicine, 2024
Publication **1%**
- 3** Submitted to University College London
Student Paper **1%**
- 4** Rick Kye Gan, Jude Chukwuebuka Ogbodo, Yong Zheng Wee, Ann Zee Gan, Pedro Arcos González. "Performance of Google bard and ChatGPT in mass casualty incidents triage", **1%**

The American Journal of Emergency Medicine, 2023
Publication

- 5** buscador.una.edu.ni
Internet Source **1%**
- 6** Submitted to University of Wales, Bangor
Student Paper **<1%**
- 7** www.mdpi.com
Internet Source **<1%**
- 8** www.researchsquare.com
Internet Source **<1%**