CUCFATE Frameworks for Safe and Effective Large Language Models in Medical Education: Using Qualitative Methods

Majdi Quttainah, Vinaytosh Mishra, Somayya Madakam, Yotam Lurie, Shlomo Mark

Submitted to: JMIR AI
on: August 16, 2023

Table of Contents

CUCFATE Frameworks for Safe and Effective Large Language Models in Medical Education: Using Qualitative Methods

Majdi Quttainah[1] PhD; Vinaytosh Mishra[2] PhD; Somayya Madakam[3] PhD; Yotam Lurie[4] PhD; Shlomo Mark[5] PhD

[1]Kuwait University Kuwait KW
[2]Gulf Medical University Ajman AE
[3]ATLAS Skill Tech University Mumbai IN
[4]Ben-Gurion University Negev IL
[5]Shamoon College of Engineering Be'er Sheva IL

Corresponding Author:
Vinaytosh Mishra PhD
Gulf Medical University
Al Jurf 1
Ajman
AE

Abstract

Background: The world has witnessed increased adoption of Large Language Models (LLMs) in the last year. Although the products developed using LLMs have the potential to solve accessibility and efficiency problems in healthcare, there is a lack of guidelines available for developing LLMs for healthcare and especially medical education.

Objective: The study aims to identify and prioritize the enablers for developing successful LLMs for medical education. The study also discusses the relationship among these identified enablers.

Methods: The study first identifies key enablers for LLM development using the narrative review of extant literature. The next opinion of users of LLMs was taken to determine the relative importance of these enablers using the multi-criteria decision-making method called the Analytical Hierarchy Process. Further, Total Interpretive Structural Modelling (TISM) was used to analyze product developers' perspectives and ascertain the relationship and hierarchy among these enablers. Finally, Cross-impact matrix multiplication was applied to classification (MICMAC) to find these enablers' relative driving and dependence power. The non-probabilistic purposive sampling was used for the study.

Results: The result of AHP concluded that credibility, with a priority weight of 0.37, is the most important enabler, followed by Accountability (0.27642) and Fairness (0.10572). In contrast, usability, with a priority weight of 0.04, has negligible importance. The results of TISM concur with the findings of the AHP. The only striking difference from the user's preference was that product developers gave the least importance to cost. The development of the MICMAC analysis suggests that cost has a strong influence on other enablers. The inputs of the focus group were found reliable, with a consistency ratio (CR=0.084) less than 0.1.

Conclusions: The study is the first to identify, prioritize, and analyze the relationship of enablers for effective LLMs for medical education. The study provides an easy to comprehendible prescriptive framework CUCFATE (Cost, Usability, Credibility, Fairness, Accountability, Transparency, and Explainability) for the same. The study findings are useful for healthcare professionals, health technology experts, medical technology regulators, and policymakers.

(JMIR Preprints 16/08/2023:51834)
DOI: https://doi.org/10.2196/preprints.51834

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
Please make my preprint PDF available to anyone at any time (recommended).
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
✓ **Only make the preprint title and abstract visible.**
No, I do not wish to publish my submitted manuscript as a preprint.
2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

CUCFATE Frameworks for Safe and Effective Large Language Models in Medical Education Using Qualitative Methods

Majdi Quttainah, Vinaytosh Mishra, Somayya Madakam, Yotam Lurie,

Background

The world has witnessed increased adoption of Large Language Models (LLMs) in the last year. Although the products developed using LLMs have the potential to solve accessibility and efficiency problems in healthcare, there is a lack of guidelines available for developing LLMs for healthcare and especially medical education.

Objective

The study aims to identify and prioritize the enablers for developing successful LLMs for medical education. The study also discusses the relationship among these identified enablers.

Methods

The study first identifies key enablers for LLM development using the narrative review of extant literature. The next opinion of users of LLMs was taken to determine the relative importance of these enablers using the multi-criteria decision-making method called the Analytical Hierarchy Process. Further, Total Interpretive Structural Modelling (TISM) was used to analyze product developers' perspectives and ascertain the relationship and hierarchy among these enablers. Finally, Cross-impact matrix multiplication was applied to classification (MICMAC) to find these enablers' relative driving and dependence power. The non-probabilistic purposive sampling was used for the study.

Results

The result of AHP concluded that credibility, with a priority weight of 0.37, is the most important enabler, followed by Accountability (0.27642) and Fairness (0.10572). In contrast, usability, with a priority weight of 0.04, has negligible importance. The results of TISM concur with the findings of the AHP. The only striking difference from the user's preference was that product developers gave the least importance to cost. The development of the MICMAC analysis suggests that cost has a strong influence on other enablers. The inputs of the focus group were found reliable, with a consistency ratio (CR=0.084) less than 0.1.

Conclusion

The study is the first to identify, prioritize, and analyze the relationship of enablers for effective LLMs for medical education. The study provides an easy to comprehendible prescriptive framework CUCFATE (Cost, Usability, Credibility, Fairness, Accountability, Transparency, and Explainability) for the same. The study findings are useful for healthcare professionals, health technology experts, medical technology regulators and policymakers.
Keywords: Large Language Models, ChatGPT, CUCFATE Framework, AHP, TISM

Introduction

Natural Language Programming solutions have been available for the last fifteen years. However, the avalanche breakdown phenomena recently hit the use of these models with the launch of ChatGPT by a company named OpenAI. The company received investment from Elon Musk and others when it was established a few years ago. With 1.6 billion monthly users, this freemium is the fastest-growing application in the history of the internet. Released on November 30, 2022, OpenAI released ChatGPT (Chat Generative Pre-Trained Transformer), a generative language model tool that enables users to converse with machines about various subjects. Since its debut, ChatGPT has sparked much discussion and enthusiasm in multiple industries, including medicine. ChatGPT and related technologies have been identified as disruptive innovations with the potential to revolutionize academia and scholarly publishing [1]. Additionally, preliminary research suggests that ChatGPT has practical applications throughout the clinical workflow [2]

The introduction of ChatGPT and the subsequent release of several extended products and functional plug-ins have profoundly impacted scientific researchers. They have also influenced the ideas and methodologies used in traditional research, including recommendation, emotion recognition, and information generation. ChatGPT's assistance has improved some of the work, particularly in data generation. ChatGPT can offer helpful supplementary information to raise the calibre of data generation. With the integration of machine learning and artificial intelligence (AI) technologies, medical imaging has advanced quickly. Among these developments, using cutting-edge language models like Large Language Model (LLM), ChatGPT, and GPT-4 has shown significant promise in elevating several elements of medical imaging and revolutionizing radiology. These models can produce and comprehend human-like text thanks to access to various textbooks, journals, and research materials. This could provide the necessary context and prior knowledge to support a variety of tasks involving medical imaging, such as synthesis, reconstruction, analysis, segmentation, interpretation, automated reporting, and more. It has been improved using supervised and

reinforcement learning methods and is based on OpenAI's GPT big language models. These models have performed excellently in various NLP tasks, including language translation, text summarization, and question-answering. They have been pre-trained on enormous amounts of text data. Users can ask questions, get responses, and engage in genuine conversation with the bot thanks to ChatGPT's human-like conversational experience. ChatGPT and other big models remain a research hotspot in multimedia analysis and application. However, several crucial difficulties must be resolved: 1) How to interact with ChatGPT to collect more useful auxiliary information; 2) How to combine with traditional inquiries to fully exploit ChatGPT's better benefits; and 3) How to analyze the data obtained from ChatGPT and incorporate it with the intended usage. Effectively using past information from huge models and investigating consistency and complementary features across many modalities to improve multi-modal generation performance is a significant challenge, particularly in AI Generated Content (AIGC). The finest use cases for ChatGPT, a well-liked chatbot built on a potent AI language model, are still being worked out. This article offers some suggestions on how to make the tool work for you when writing academic papers. The following steps in writing an essay, thesis, or dissertation can be helped with via ChatGPT: creating a research question, developing a plan, developing literary concepts, rewriting text and getting feedback.

Moreover, Natural language processing and automated data analysis capabilities offered by ChatGPT enable researchers, marketers, and organizations to analyze papers quickly and accurately. Its AI-powered skills can assist you in spotting significant trends and insights in your data that might otherwise be challenging to find. Additionally, ChatGPT can assist you in creating top-notch prompts for paper analysis.

LLM Functionality

ChatGPT is a prediction system that anticipates what it should write based on previously processed texts. This sort of artificial intelligence is known as a language model. What makes it more promising than its predecessors is that it is trained on enormous amounts of data, much of which originates from the abundant supply of data available on the internet. According to OpenAI, ChatGPT was also trained on examples of back-and-forth human interaction, which makes it sound much more human in its discourse, thus advancing the capability of natural language processing (NLP) solutions.

NLP is a field of artificial intelligence employing linguistics, statistics, and machine learning to enable computers to comprehend spoken language. NLP systems can infer meaning from spoken or

written words, including all the subtleties and complexity of an accurate narrative text. This makes it possible for machines to get value from even unstructured data. NLP has witnessed significant advancements in recent years. The Large Language Model (LLM) is a Deep Learning algorithm that can be used to perform NLP tasks, including, among other abilities, summarizing and generating text. One of the applications of LLM-based chatbot. These are computer programs that can simulate conversations with human users. NLP techniques can be used to enable chatbots to understand and respond to user input. LLM uses deep learning techniques to understand and generate human language. It requires training on vast amounts of text data and using statistical algorithms to learn patterns and relationships within language. They can perform various tasks, including language translation, question answering, sentiment analysis, and summarization. Users can learn, compare, and validate answers for different academic subjects, including physics, math, and chemistry, as well as abstract topics like philosophy and religion, using ChatGPT [3]. They can also generate human-like text, such as news articles, chatbot conversations, and even literary works like essays and romantic poems. What makes GPTs different from other LLMs is their architecture and training methodology. GPTs are based on a deep learning architecture called the "Transformer". Transformers are designed to process sequential data, such as language, more efficiently than other architectures. Large language models are currently at the forefront of intertwining AI systems with human communication and everyday life [4]. Large pre-trained language models have significantly advanced natural language processing research on various applications [5,6]. Although these more complicated language models can produce complex and coherent natural language, several recent studies have shown that they can also pick up unfavourable social biases that can feed negative stereotypes [7].

NLP in Healthcare

Healthcare consumers may turn to the research literature for information not provided in patient-friendly documents. However, reading medical literature can be difficult. A study looked at four elements made possible by natural language processing to increase access to medical papers: explanations of foreign terminology, plain language section summaries, a list of crucial questions that direct readers to the portions that provide the answers, and simple language summaries of those passages. Significant advancements in smart healthcare have been made in recent years [8]. New AI technologies enable a range of intelligent applications in various healthcare contexts. Natural language processing (NLP), a fundamental AI-powered technology that can analyze and comprehend human language, is crucial for smart healthcare [9]. Natural language processing methods have been utilized to organize data in healthcare systems by sifting out pertinent information from narrative

texts to offer information for decision-making. Thus, NLP approaches help lower healthcare costs and are essential for streamlining healthcare procedures [10]. Advancements in NLP will make robotic process automation possible in healthcare, which can further drive efficiency in healthcare. Healthcare data is complex, and this should be given due consideration at the time of designing healthcare applications. Deep Learning (DL) approaches, such as Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) models, have become prominent in healthcare applications, and their accuracy is promising. Still, much must be done to enable their usage without human supervision. Deep Learning techniques offer an effective and efficient model for data analysis by revealing hidden patterns and extracting valuable information from a large volume of health data, which standard analytics cannot perform within a given time frame [11].

ChatGPT in Medical Education

ChatGPT has many potential applications in healthcare education, research, and practice [12]. It can enhance medical education by helping students develop subjective learning and expression skills [13]. The number of ChatGPT users has shown exponential growth and is being increasingly utilized by students, residents, and attending physicians to direct learning and answer clinical questions [14]. However, authors using ChatGPT professionally for academic work should exercise caution as it is unclear how ChatGPT handles hazardous content, false information, or plagiarism [15]. While ChatGPT can make radiological reporting simpler, there is still a chance of inaccurate statements and missing medical information [15]. Therefore, it needs refinement before being used frequently in medicine [16]. A recent review explores ChatGPT's applications and reports challenges such as ethical concerns, data biases, and safety issues [17]. Thus, it is imperative to balance AI-assisted innovation and human expertise [18]. ChatGPT has quickly gained significant attention from academia, research, and industries despite these shortcomings. This study attempts to determine the requirements for a successful LLM application in medical education using a narrative review of existing literature.

Enablers of LLM for Medical Education

Enablers in this research refer to factors, resources, or conditions that facilitate or support achieving a good LLM application for medical education. Medical education prepares would-be physicians and other healthcare professionals with the knowledge, skills, and attitudes necessary for competent and compassionate patient care. Enablers make it easier for something to happen or someone to accomplish a particular task. Enablers of LLM for Medical Education can be tangible or intangible and should play a crucial role in achieving outcomes expected from the application.

As LLMs are trained on massive data, they are resource demanding. The cost of training for an LLM for medical education may be prohibitive [19]. Thus, it is imperative to use efficient computing to address this issue [20]. Usability is one of the key criteria making an application useful in medical education, and LLMs are no exception [21]. Extant literature cites usability as an important criterion for a successful technology in education [22]. Similar credibility of application becomes very important in medical education [23]. Another recently published article cites the credibility of technological interventions in medical education [24]. Although ChatGPT puts disclaimers about the source of information, it doesn't disclose it categorically. Worse is it hallucinates about the source sometimes and may be misleading. Large language models also have issues with fairness, computation, and privacy. By perpetuating social prejudices and stereotypes, they risk causing unfair discrimination, physical harm, and harm to their reputation [25]. In their study, Ma et al. provide an overview of fairness in multilingual and non-English situations, emphasizing the limitations of recent studies and the challenges faced by English-only methodologies [26].

Next comes the issue of accountability with LLMs such as ChatGPT. It is taking responsibility for one's obligation to treat others honestly and morally. Who will be held accountable and responsible if the LLM model provides incorrect recommendations or forecasts for a particular downstream activity? Overall, employing big language models has considerable dangers; therefore, precautions must be taken to minimize these risks and ensure their ethical and responsible use. To foster a cross-disciplinary global inclusive consensus on the ethical use, disclosure, and proper reporting of GAI/GPT/LLM technologies in academia, Cacciamani et al. presented the ChatGPT, Generative Artificial Intelligence, and Natural Large Language Models for Accountable Reporting and Use Guidelines initiative in 2023. The underlying model of GPT3.5 deviates from the ethical guidelines proposed by Cacciamani et al. [23]. Another important criterion reported for medical applications is transparency. It is an ethic across science, engineering, business, and the humanities. It refers to functioning in a way that makes it simple for others to observe what actions are taken [28]. Transparency is a sign of responsibility, honesty, and openness. LLMs are opaque to users. Recently suggested explainability techniques aim to make language models more transparent. Although they are not a cure-all, they might act as the basis for models with fewer flaws or, at the very least, can explain their logic. In their systematic experiments with synthetic data, Wu, Z. et al. demonstrate that autoregressive and masked language models can successfully learn to emulate semantic relations between expressions in strong transparency, where all expressions have context-independent denotations [29].

Finally, the LLMs used in medical education must be explainable, and the best freely available

options lag here. Most Large Language Models (LLMs) are complex models built using Deep Learning [30]. They can produce better predictions with more information or network parameters but at the sacrifice of explainability. Some models fail to describe how they came to their conclusion. Recently suggested explainability techniques aim to make language models more transparent. Even though they are not solutions, they can act as the basis for less problematic models or, at the very least, models that can explain their logic. However, the authors Du, M. et al. identified false patterns detected by LLMs using explainability in their study [31]. The enablers identified using the secondary research are listed in Table 1.

Table 1: Summary of Enablers of LLM for Medical Education

	Enabler	Description	References
E1	Cost	Cost of computation, including hardware, software, and energy requirement.	[19,20]
E2	Usability	User-centric design, ease of use, and positive user experiences	[21,22]
E3	Credibility	Level of trust and reliability that users place in the application.	[23,24]
E4	Fairness	Absence of unfair discrimination, physical harm, and harm to user reputation.	[25,26]
E5	Accountability	Taking responsibility for the obligation to treat users with honesty and morality.	[27,28]
E6	Transparency	Functioning in a way that makes it simple for others to observe what actions are taken.	[28,31]
E7	Explainability	Ability to describe how the models came to their conclusion	[30,31]

Source: Author's Compilation

Need of the Study

The need for this study arises from the rapid integration of Large Language Models (LLMs) like ChatGPT in various fields, including medical education. LLMs offer promising benefits for healthcare, but their effective integration in medical education is still a developing area. This study aims to identify and prioritize the key enablers for successful LLM implementation in medical education. It addresses the lack of comprehensive frameworks guiding the development and use of LLMs in this field. By exploring the dynamics of various enablers such as credibility, accountability, fairness, cost, usability, transparency, and explainability, the study provides a structured approach to

enhance the quality and effectiveness of LLMs in educating healthcare professionals. Thus, this study attempts to answer three major research questions: (1) What are the enablers of a suitable LLM application for medical education? (2) What is the relative importance of these enablers in achieving the goals of medical education? and (3) What is an approach to developing an LLM to achieve medical education goals? With this background, the following are the research objectives (RO) of the study:

RO1: Identify the enabler of a suitable LLM for medical education.
RO2: Prioritize the identified enablers in achieving the goals of medical education.
RO3: Propose a framework for developing an LLM to achieve the medical education goals.

Methodology

To achieve the first research objective, this study uses a narrative review of extant literature published on technology solutions in medical education. A narrative review is a scholarly article synthesizing existing research on a particular topic in a narrative or story-like manner. Unlike systematic reviews or meta-analyses, which use rigorous methodologies to analyze and summarize research findings quantitatively, narrative reviews provide a qualitative, comprehensive overview of a subject. They often involve critical analysis and discussion, integrating the author's expertise and interpretation. Narrative reviews are useful for obtaining a broad understanding of a topic and identifying trends, gaps, and controversies within a field. The authors SM and VM searched Scopus, Web of Science, and Google Scholar databases to identify suitable literature for the review. The articles selected for the study are literature in the English language and have been published in the last five years. The second stage eliminated duplicates and articles for which full text was unavailable. One article published in 2010 was added on the recommendation of the focus group as it was found useful in explaining competing interests in medical education. The seven identified enablers (E1 to E7) help us answer Research Question 1. These enablers were presented in front of a focus group comprising seven experts working in universities and institutions delivering medical education in India and the United Arab Emirates to validate the selection of barriers Table 2. The focus group endorsed the choice of their seven enablers for further research. The researcher VM facilitated the focus group discussion to finalize the enablers.

Table 2: Characteristics of the Focus Group Used for AHP

Expert	Qualification	Experienc	Age	Nationality

		e		
Cardiologist	Masters in Medicine	12	42	India
Endocrinologist	Masters in Medicine	20	45	India
Technology Expert	Doctor of Philosophy	15	50	USA
Dentistry Educator	Masters in Dentistry	10	40	UAE
Podiatrists Educator	Doctor of Philosophy	10	35	UAE
Diabetes Educator	Doctor of Philosophy	18	43	India
Nursing Educator	Doctor of Philosophy	15	41	UAE
Radiologist	Doctor of Philosophy	12	41	India

Source: Author's Compilation

AHP Modelling

The Analytical Hierarchy Process (AHP) was utilized to achieve the second objective. AHP is a popular method for calculating the relative importance of the criteria in the Multi-Criteria Decision Analysis (MCDA) method. It has been extensively used in the management and social science literature [32]. The advantage of the process is that it incorporates the mechanisms to assure reliability in the decision-making case of ambiguity; researchers have suggested using a Fuzzy version of AHP [33]. Further, some researchers have suggested the entropy weight method to reduce the negative effect of individual subjective evaluation bias on the accuracy of comprehensive evaluation [34]. Since the ranking obtained by the AHP method is further validated by Total Interpretive Structural Modeling (TISM), Fuzzy Logic or Entropy Weight was avoided. The five steps used for AHP are (1) Defining the Decision Problem, (2) Creating a Hierarchy, (3) Pairwise Comparison, (4) Derive Weighted Priority, and (5) Consistency Check for Decision. The method used for pairwise comparison was the Delphi method. A Cut-off of 75% was used to accept the value for the pairwise comparison. The standard scale Saaty suggested was used for the pairwise comparison [35]. Researcher VM facilitated the data collection for the AHP model.

TISM Modeling

Finally, for the third objective, this study investigates the relationships among key enablers for building a suitable medical education LLM. Qualitative research design is useful to understand the phenomenon under study instead of assessing the strength and direction of causal relationships in the conceptual model [36]. This objective utilized a focus group with five information technology experts with product development and research experience. The details of this expert group are listed in Table 3.

Table 3: Characteristics of the Focus Group Used for TISM

Expert	Qualification	Experience	Age	Country
Product Development	Masters in management	21	42	Singapore
Product Development	Bachelors in engineering	21	42	UAE
Technology Expert	Bachelors in engineering	19	40	India
Technology Expert	Masters in engineering	10	33	India
Decision Science Expert	Doctor of Philosophy	10	38	India

Source: Author's Compilation

This study has used TISM to model the enabler for medical education LLM application. In his seminal paper, Sushil provides a detailed account of the interpretation of ISM and TISM and the latter's advantage over the former [37]. For the sake of brevity, the authors have not included the details of the TISM method in this paper. It is a process that converts poorly articulated mental models of systems into visible and well-defined models useful for better understanding and decision-making [38]. The presence and absence of a relationship between enablers were ascertained based on an unstructured interview of the focus group conducted by the researcher SM. If more than fifty per cent of the focus group members think there is a relationship between two enablers, the response was taken as 'Y'. The summary of the method used in the research is described in Figure 1.

Ethical Considerations

This study, involving a qualitative focus group discussion, did not require approval from an ethical review board as it did not involve human subjects in a manner necessitating such review. No informed consent was required for the same reason. However, to maintain ethical standards, we ensured that all data collected was either anonymized or de-identified. This means any information that could potentially identify individual participants was removed or altered to protect their privacy. No compensation was provided to participants, as is common in studies of this nature. This decision was made considering the study's design and the ethical imperative to avoid undue influence on participants' responses. The absence of compensation was communicated to all participants. Throughout the study, we adhered to strict data protection protocols to safeguard the confidentiality of the information shared during the focus group discussions. These measures included secure data storage, restricted access to authorized personnel, and adherence to data protection laws and regulations. This approach ensured that the privacy and integrity of participant information were

always maintained.

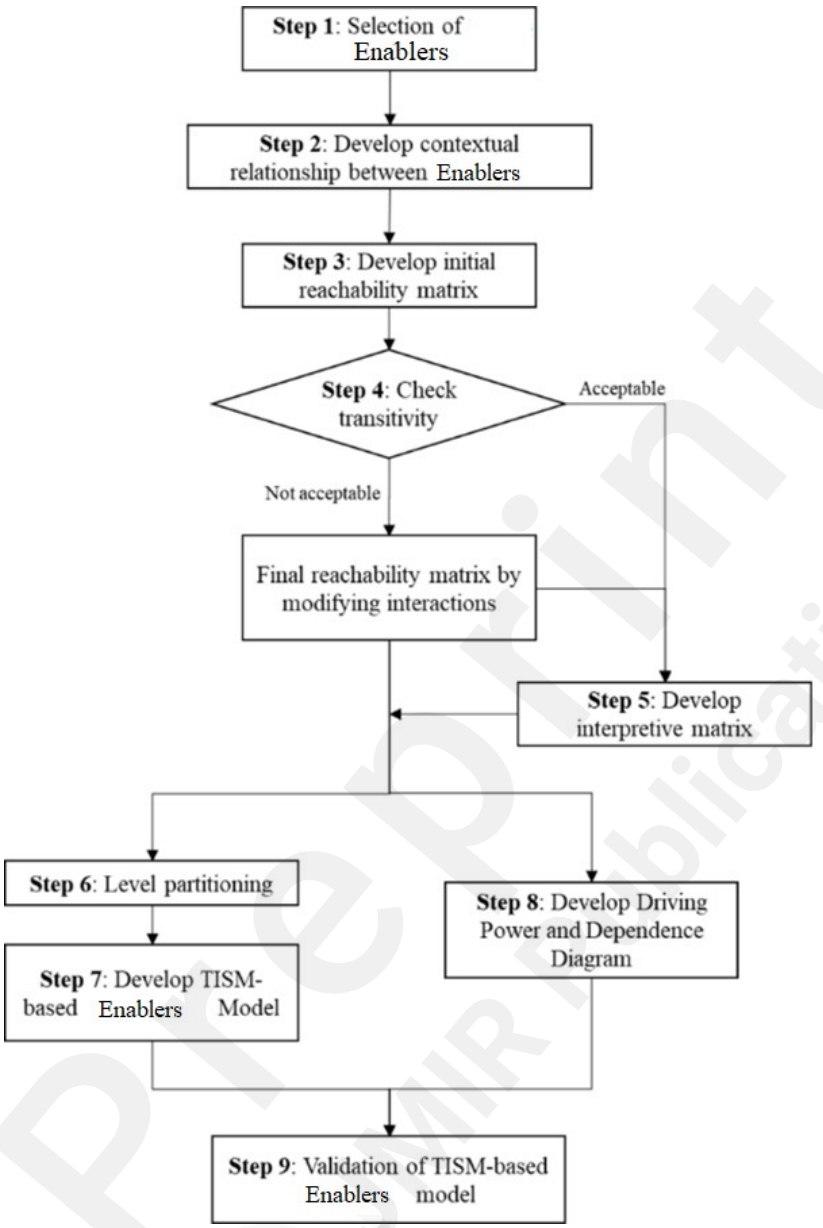


Figure 1: Summary of the TISM Approach Used in the Study

Source: Adapted by Authors [34]

Results

The results of the narrative review are summarised in Table 1 in the earlier section. This section presents the results related to the second and third research objectives, respectively.

AHP Modelling

The list of the selected enablers for developing a suitable LLM medical education application is depicted in Table 1. The focus group was asked to provide their input for pairwise comparison, and the resultant matrix [A] is described in Table 4.

Table 4: Initial Pairwise Comparison Matrix for the AHP

		E1	E2	E3	E4	E5	E6	E7
Cost (E1)	E1	1	3	0.2	1	0.2	3	3
Usability (E2)	E2	0.33	1	0.11	0.33	0.11	1	1
Credibility (E3)	E3	5	9	1	5	5	3	3
Fairness (E4)	E4	1	3	0.2	1	0.2	3	3
Accountability (E5)	E5	5	9	0.2	5	1	5	5
Transparency (E6)	E6	0.33	1	0.33	0.33	0.2	1	1
Explainability (E7)	E7	0.33	1	0.33	0.33	0.2	0.2	1

Source: Author's Compilation

Once the Initial Comparison Matrix was determined, it was normalized, and an average of each row was taken to calculate the priority weight [X]. The normalized matrix, priority weight (PW) and rank of the enablers is given in Table 5. The priority weight, the eigenvector, is used to calculate the consistency ratio further.

Table 5: Normalized Matrix and Priority Weight of Enablers

	E1	E2	E3	E4	E5	E6	E7	PW	Rank
E1	0.077	0.1111	0.0844	0.077	0.0289	0.1852	0.1765	0.10572	3
E2	0.0254	0.037	0.0464	0.026	0.0159	0.0617	0.0588	0.03871	7
E3	0.3849	0.3333	0.4219	0.385	0.7236	0.1852	0.1765	0.37289	1
E4	0.077	0.1111	0.0844	0.077	0.0289	0.1852	0.1765	0.10572	3
E5	0.3849	0.3333	0.0844	0.385	0.1447	0.3086	0.2941	0.27642	2
E6	0.0254	0.037	0.1392	0.025	0.0289	0.0617	0.0588	0.0538	5
E7	0.0254	0.037	0.1392	0.025	0.0289	0.0123	0.0588	0.04674	6

Source: Author's Compilation

Finally, the consistency Ratio of judgement was calculated by dividing the consistency index (CI) by the Random Index (RI). The next stage is to calculate λ_{max} Leading to the calculation of CI. For the calculation of eigenvector X, the following equation is applicable:

[A] X= λ_{max} X – (1)

Using Table 3-4 and Equation 1:

$[A]\,X=¿\ 0.28,3.46,0.76,2.26,0.39.0.34]$ – (2)

$$\lambda_{max} = Average\left\{\frac{0.76}{0.11}, \frac{0.24}{0.04}, \frac{3.46}{0.37}, \frac{0.76}{0.11}, \frac{2.26}{0.28}, \frac{0.39}{0.05}, \frac{0.34}{0.05}\right\} \text{ - (3)}$$

$\lambda_{max}=7.66$ – (4)

CI= (7.66-7)/6=0.11 – (5)

The random index (RI) value for a 7X7 Matrix is 1.32 from the Random Index table. Thus, the Consistency Ratio (CR) becomes 0.084, less than 0.1, hence acceptable.

TISM Modelling

Now, we move to the result of TISM for ascertaining the relationship between these seven enablers. Table 6 shows a matrix indicating the interrelationships between enablers listed in Table 1. The existence of a relationship between enablers is depicted by the letter 'Y', while absence is represented by the letter 'N'. The resultant matrix is called the Structural Self-Interaction Matrix (SSIM).

Table 6: Structural Self-Interaction Matrix for the Study

SN	Enablers for LLMs	E1	E2	E3	E4	E5	E6	E 7
E1	Cost (E1)	Y	Y	N	N	N	Y	N
E2	Usability (E2)	Y	Y	N	N	N	Y	Y
E3	Credibility (E3)	N	N	Y	Y	Y	N	N
E4	Fairness (E4)	N	N	Y	Y	N	N	N
E5	Accountability (E5)	N	N	Y	N	Y	N	N
E6	Transparency (E6)	Y	Y	N	N	N	Y	Y
E7	Explainability (E7)	N	Y	N	N	N	Y	Y

Source: Authors' Compilation

The next step is to replace all 'Y'' with '1' and all 'N' with 0 and incorporate the transitivity rule to get the Final Reachability Matrix (FRM) listed in Table 7.

Table 7: Final Reachability Matrix for the Study

SN	Enablers for LLMs	E1	E2	E3	E4	E5	E6	E7	Driving Power
E1	Cost (E1)	1	1	0	0	0	1	1	4
E2	Usability (E2)	1	1	0	0	0	1	1	4
E3	Credibility (E3)	0	0	1	1	1	0	0	3
E4	Fairness (E4)	0	0	1	1	0	0	0	2
E5	Accountability (E5)	0	0	1	0	1	0	0	2
E6	Transparency (E6)	1	1	0	0	0	1	1	4
E7	Explainability (E7)	0	1	0	0	0	1	1	3

	Dependence Power	3	4	3	2	2	4	4	

Source: Authors' Compilation

The next step in developing models for medical education LLMs is to list Reachability and Antecedent sets for each enabler and perform Level Partitioning (LP). LP is an iterative process of assigning barriers at different levels. Enablers with similar intersection sets as reachability sets are placed at the top level. The process is repeated until levels for all the enablers are established. In this study, all enablers were assigned after three iterations; hence, there are three levels in the hierarchy. The summary of level partitioning for this study is listed in Table 8. The level of an enabler indicates driving and dependence power, as indicated in Table 7. The higher the level of the enabler, the more dependent it is. At the same time, the driving ability improves as we go down to the lower levels.

Table 8: Summary of Label Partitioning (LP) Iterations (1 to 6)

Enablers (Mi)	**Reachability Set R(Mi)**	**Antecedent Set A(Ni)**	**Intersection Set R(Mi)∩A(Ni)**	**Level**
1	1	1	1	III
2	1,2,6,7	1,2,6,7	1,2,6,7	I
3	3,4,5	3,4,5	3,4,5	I
4	3,4	3,4	3,4	I
5	3,5	3,5	3,5	I
6	1,2,6,7	1,2,6,7	1,2,6,7	I
7	7	1,7	7	II

Source: Authors' Compilation

TISM Model

Once the level partitioning was done, the TISM model was developed and presented to the focus group for validation. Only significant transitive links were included in the model to make it easy to interpret. The final digraph for the TISM model developed in the study is depicted in Figure 2.

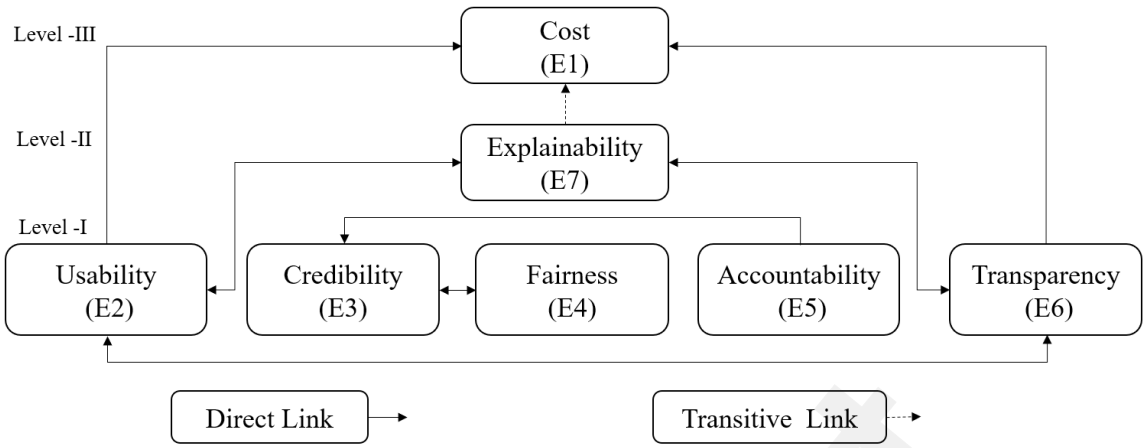


Figure 2: Diagraph of TISM for LLM in Medical Education

Source: Authors Compilation

MICMAC Analysis

The study further uses Cross-impact matrix multiplication applied to classification (MICMAC) analysis to validate the study's findings and derive conclusions. It involved the development of a graph that classifies enablers based on their driving and dependence power Figure 3. The first quadrant contains autonomous enablers E3, E4, and E6 (Credibility, Fairness and Accountability). The variables falling in this quadrant have low driving and dependence power. The two enablers falling in the grey region between the third (linkage) and fourth (independent) quadrants are E2 and E6 (Usability and Transparency) and have medium driving and dependence power. Similarly, E7 (Explainabilty) falls in the grey region between the first (autonomous) and second (dependent) variables. Finally, E1 (Cost) falls under the fourth (independent) quadrant.

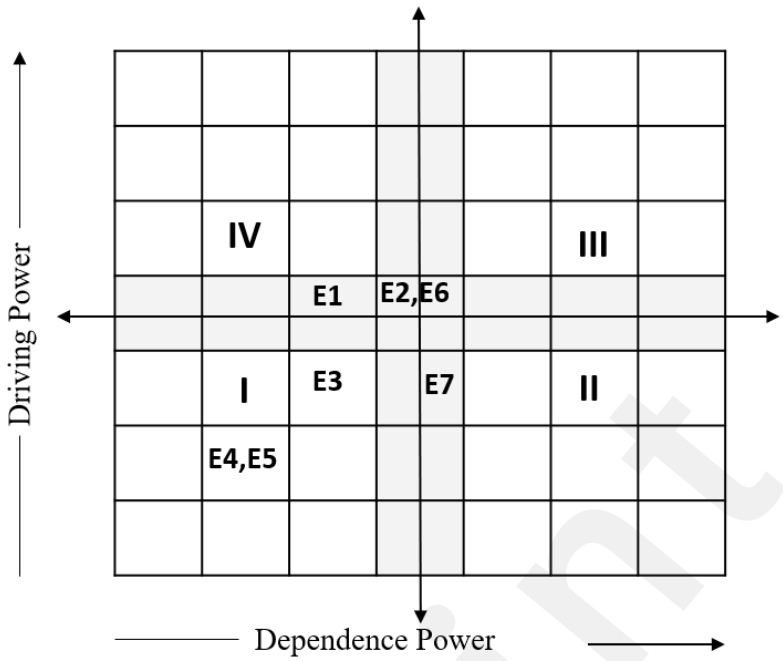


Figure 3: MICMAC Analysis for Enablers of LLM in Medical Education

Source: Authors Compilation

Discussion

The results of the AHP suggest that credibility, followed by accountability, are the foremost enablers for effective LLMs in medical education. Extant literature supports this finding, as literature mentions the source of information based on which the response was generated [39]. Similarly, the importance of defining accountability is reported in recent literature. Similar to an existing study, the researchers also advocate accountability as an important factor in increasing the adoption of LLMs in medical education, training, and practice [40]. Next in importance are ethical issues such as fairness and cost. LLMs have been criticized for bias against gender or ethnic groups [41]. These problems need to be addressed to make LLMs effective in medical education. Secondly, training LLMs on billions of parameters is demanding; thus, technology giants will launch these LLMs [42]. The governments should ensure that the cost of using these LLMs doesn't become prohibitive for end users, and they resort to half-baked solutions, eventually affecting the safety of patients.

Contrary to existing studies, transparency and explainability come in fifth and sixth place in importance [40]. Many best practices related to health technology suggest that models should use explainable artificial intelligence in medical devices [41]. The low priority of these enablers indicates that the end user is unaware of the criticality of these factors, and healthcare professionals need to be educated about them as they are not technology savvy [43]. Governments should make guidelines for the approval of Software as Medical Devices so that these enablers are taken care of at the product

development stage. Finally, the focus group thinks that usability is the least important factor out of the enablers discussed in the study. Although the general-purpose LLMs, such as ChatGPT, are less cluttered, their performance is input-dependent. Improving the prompt use of the recommendation system can enhance the usability and accuracy of the LLMs in medical education [44]. The expert group advised that the LLMs will improve on these factors with time.

Next, we will discuss product development and technology expert's input for the TISM model. The model results suggest a slight difference in the perspective of product developers and end users. They give equal importance to Credibility, Fairness, Accountability, Transparency, and Explainability. These results are consistent with extant literature published in peer-reviewed journals [40,42]. These are all features related to model development and training.

Contrary to earlier studies, this group also gives usability less significance and puts it on a medium level [44]. Thus, the finding of the TISM validates the results of AHP. The only difference is the cost is the least important for product developers. Recently published studies opine that economic and environmental costs are significant in developing general-purpose LLMs [45].

A successful LLM development involves a complex interplay between technical innovation, regulatory compliance, production costs, and end-user needs. The aim should be to develop products that excel in functionality and positively impact the lives of those who rely on them without causing financial hardship. Thus, this study calls for collaboration between product developers, original equipment manufacturers, regulators, and other stakeholders to find solutions that align with technological advancements and societal expectations for affordability and accessibility.

Finally, the study validates its findings using MICMAC analysis, creating a graph that categorizes enablers based on driving and dependence power. Here, enablers, namely, Credibility, Fairness, and Accountability, are in the first quadrant (Autonomous) with low power. These variables are relatively independent and have limited influence on other variables. While Usability and Transparency are in the grey region between the third quadrant (Linkage) and fourth (independent) quadrants with medium power, they have a medium influence on other variables and are similarly influenced by them. Explainability falls in the grey region between the first (Autonomous) and second (Dependent) quadrants, while cost is in the fourth quadrant. Again, made us conclude that their enablers have a medium influence on other variables and a similar influence on them. Finally, the cost falls under the fourth quadrant (Independent) and makes us believe it strongly influences other enablers without being significantly influenced by them. MICMAC analysis comprehensively explains the relationships and dynamics among variables within a complex system. It helps decision-makers

identify key drivers, dependencies, and interactions, enabling them to make informed strategic decisions and allocate resources effectively.

Conclusion

The study emphasizes key factors for effective Language Models (LLMs) in medical education: Credibility and Accountability are vital, while addressing bias and cost is crucial for LLM potential. Though important, Transparency and Explainability rank lower among health professionals, suggesting a need for education. Usability is the least important, but enhancing prompt use improves LLM accuracy. The study highlights a slight difference between product developers and end users. Both prioritize Credibility, Fairness, Accountability, Transparency, and Explainability. Usability ranks lower for developers. Successful LLM development balances innovation, compliance, costs, and user needs. Collaboration among stakeholders is crucial for aligning with technology and societal expectations.

The study has one implication each for theory as well as practice. For theory, this study extends the FATE framework. It proposes a more comprehensive CUC-FATE (Cost, Usability, Credibility, Fairness, Accountability, Transparency, and Explainability) for developing LLMs for healthcare professionals. For practice, this study is the first of its kind and provides a prescriptive framework for developing LLMs in healthcare, especially medical education. The study's findings are useful for policymakers, medical device regulators, education policymakers, healthcare professionals, and product developers at the helm of making software as a medical device (SaMD).

One of the limitations of the study is that it uses experts from India and UAE only. Although technology and healthcare practices are standardized globally, the findings should only be generalized to the population from these geographies. The study provides the relationship between different enablers but does not discuss the strength of these associations. Graph Theory or Structured Equation Modeling can be used to address these gaps in future studies.

Acknowledgment

The authors are highly indebted to all focus group participants for their time and effort. Authors are also obliged to their respective institutions for infrastructural support provided. The authors also disclose using Artificial Intelligence tools Grammarly and Quillbot for manuscript language editing.

Funding Statement

The Article Processing Charges (APC) for the publication of the manuscript are funded by the College of Business Administration, Kuwait University.

Data Availability

The necessary data and calculations for the Analytic Hierarchy Process (AHP) model and the Self Interaction Matrix for the Total Interpretive Structural Modeling (TISM) Model are available on a GitHub repository. To access this information, please refer to the link to our document provided in reference [46].

Conflict of Interest

The authors state that there is no conflict of interest to report for this study.

Authors Contribution

Conceptualization: Vinaytosh Mishra, Majdi Quttainah, Somayya Madakam, Yotam Lurie, and Shlomo Mark

Data curation: Vinaytosh Mishra, Somayya Madakam

Formal Analysis: Vinaytosh Mishra, Yotam Lurie, and Shlomo Mark

Funding acquisition: Majdi Quttainah

Methodology: Vinaytosh Mishra, Majdi Quttainah

Project administration: Majdi Quttainah

Supervision: Yotam Lurie and Shlomo Mark

Validation: Yotam Lurie and Shlomo Mark

Visualization: Vinaytosh Mishra

Writing – original draft: Vinaytosh Mishra, Majdi Quttainah

Writing – review & editing: Yotam Lurie and Shlomo Mark

References

1. Haque, M. U., Dharmadasa, I., Sworna, Z. T., Rajapakse, R. N., & Ahmad, H. (2022). " I think this is the most disruptive technology": Exploring Sentiments of ChatGPT Early Adopters using Twitter Data. *arXiv preprint arXiv:2212.05856*.
2. Nastasi, A. J., Courtright, K. R., Halpern, S. D., & Weissman, G. E. (2023). Does ChatGPT provide appropriate and equitable medical advice? A vignette-based, clinical evaluation across care contexts. *medRxiv*, 2023-02.
3. Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., ... & Ge, B. (2023). Summary of ChatGPT/gpt-4 research and perspective towards the future of large language models. *arXiv preprint arXiv:2304.01852*.
4. Hagendorff, T. (2023). Machine psychology: Investigating emergent capabilities and behaviour in large language models using psychological methods. *arXiv preprint arXiv:2303.13988*.
5. Májovský, M., Černý, M., Kasal, M., Komarc, M., & Netuka, D. (2023). Artificial

Intelligence Can Generate Fraudulent but Authentic-Looking Scientific Medical Articles: Pandora's Box Has Been Opened. Journal of Medical Internet Research, 25, e46924.
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.
7. May, C., Wang, A., Bordia, S., Bowman, S. R., & Rudinger, R. (2019). On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.
8. August, T., Wang, L. L., Bragg, J., Hearst, M. A., Head, A., & Lo, K. (2022). Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing. *ACM Transactions on Computer-Human Interaction*.
9. Kaelin, V. C., Valizadeh, M., Salgado, Z., Parde, N., & Khetani, M. A. (2021). Artificial intelligence in rehabilitation targeting the participation of children and youth with disabilities: Scoping review. Journal of Medical Internet Research, 23(11), e25745.
10. Iroju, O. G., & Olaleke, J. O. (2015). A systematic review of natural language processing in healthcare. International Journal of Information Technology and Computer Science, 8, 44-50.
11. Lavanya, P. M., & Sasikala, E. (2021, May). Deep learning techniques on text classification using Natural language processing (NLP) in social healthcare network: A comprehensive survey. In 2021, the 3rd International Conference on Signal Processing and Communication (ICPSC) (pp. 603-609). IEEE.
12. Sallam, M. (2023), ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare,11(6),887, https://doi.org/10.3390/healthcare11060887
13. Seetharaman, R. (2023). Revolutionizing Medical Education: Can ChatGPT Boost Subjective Learning and Expression? Journal of Medical Systems, 47(1), 1-4.
14. Grabb, D. (2023). ChatGPT in Medical Education: a Paradigm Shift or a Dangerous Tool? Academic Psychiatry, 1-2. https://doi.org/10.1007/s40596-023-01791-9
15. Kleebayoon, A., & Wiwanitkit, V. (2023). ChatGPT in medical practice, education, and research: malpractice and plagiarism. Clinical Medicine, 23(3), 280-280.
16. Liu, J., Wang, C., & Liu, S. (2023). Utility of ChatGPT in clinical practice. Journal of Medical Internet Research, 25, e48568.
17. Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations, and future scope. Internet of Things and Cyber-Physical Systems. https://doi.org/10.1016/j.iotcps.2023.04.003
18. Milano, S., McGrane, J. A., & Leonelli, S. (2023). Large language models challenge the future of higher education. Nature Machine Intelligence, 5(4), 333-334.
19. Chen, J., & Ran, X. (2019). Deep learning with edge computing: A review. Proceedings of the IEEE, 107(8), 1655-1674.
20. Bharany, S., Sharma, S., Khalaf, O. I., Abdulsahib, G. M., Al Humaimeedy, A. S., Aldhyani, T. H., ... & Alkahtani, H. (2022). A systematic survey on energy-efficient techniques in sustainable cloud computing. Sustainability, 14(10), 6256.
21. Johnson, S. G., Potrebny, T., Larun, L., Ciliska, D., & Olsen, N. R. (2022). Usability methods and attributes reported in usability studies of mobile apps for health care education: a scoping review. JMIR Medical Education, 8(2), e38259.
22. Lu, J., Schmidt, M., Lee, M., & Huang, R. (2022). Usability research in educational technology: A state-of-the-art systematic review. Educational technology research and development, 70(6), 1951-1992.
23. Hein, H. J., Glombiewski, J. A., Rief, W., & Riecke, J. (2022). Effects of a video intervention on physicians' acceptance of pain apps: a randomized controlled trial. BMJ open, 12(4), e060020.
24. Skalidis, I., Muller, O., & Fournier, S. (2022). CardioVerse: The cardiovascular medicine in

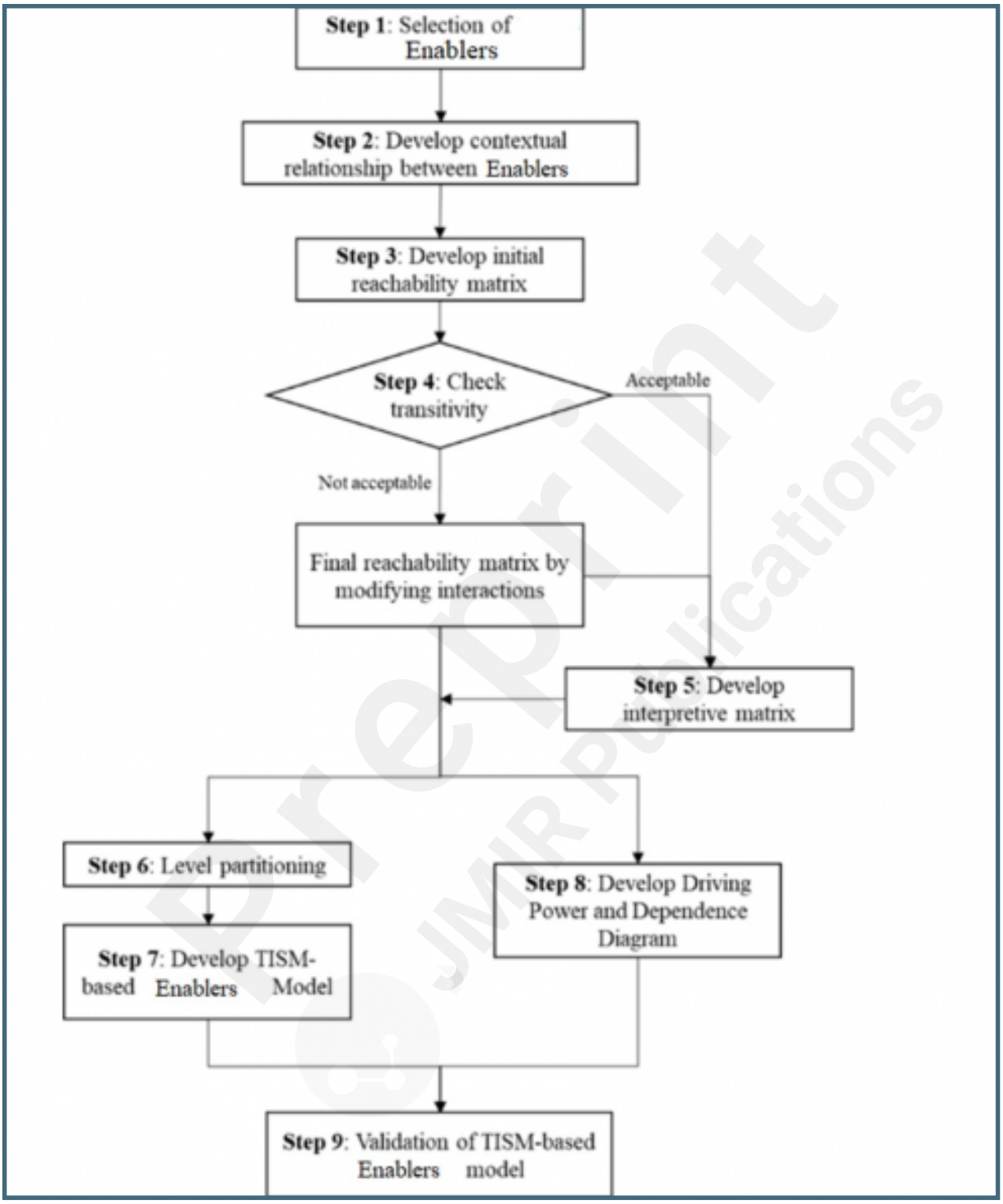
the era of Metaverse. Trends in Cardiovascular Medicine. https://doi.org/10.1016/j.tcm.2022.05.004
25. Lund, B. D., & Wang, T. (2023). Chatting about ChatGPT: how may AI and GPT impact academia and libraries? Library Hi Tech News, 40(3), 26-29.
26. Ma, H., Zhang, C., Bian, Y., Liu, L., Zhang, Z., Zhao, P., ... & Wu, B. (2023). Fairness-guided Few-shot Prompting for Large Language Models. arXiv preprint arXiv:2303.13217.
27. Cacciamani, G. E., Eppler, M. B., Ganjavi, C., Pekan, A., Biedermann, B., Collins, G. S., & Gill, I. S. (2023). Development of the ChatGPT, Generative Artificial Intelligence and Natural Large Language Models for Accountable Reporting and Use (CANGARU) Guidelines. arXiv preprint arXiv:2307.08974.
28. Hébert, P. C., MacDonald, N., Flegel, K., & Stanbrook, M. B. (2010). Competing interests and undergraduate medical education: time for transparency. CMAJ, 182(12), 1279-1279.
29. Wu, Z., Merrill, W., Peng, H., Beltagy, I., & Smith, N. A. (2023). Transparency Helps Reveal When Language Models Learn Meaning. Transactions of the Association for Computational Linguistics, 11, 617-634.
30. Susnjak, T. (2023). Beyond Predictive Learning Analytics Modelling and onto Explainable Artificial Intelligence with Prescriptive Analytics and ChatGPT. International Journal of Artificial Intelligence in Education, 1-31. https://doi.org/10.1007/s40593-023-00336-3
31. Du, M., He, F., Zou, N., Tao, D., & Hu, X. (2022). Shortcut learning of large language models in natural language understanding: A survey. arXiv preprint arXiv:2208.11857.
32. Mishra, V., & Singh, J. (2022). Health technology assessment of telemedicine interventions in diabetes management: Evidence from UAE. FIIB Business Review, 23197145221130651.
33. Dua, S., Sharma, M. G., Mishra, V., & Kulkarni, S. D. (2022). Modelling perceived risk in a blockchain-enabled supply chain utilizing fuzzy-AHP. Journal of Global Operations and Strategic Sourcing, 16(1), 161-177.
34. Mishra, V., & Rana, S. (2023). Understanding barriers to inbound medical tourism in the United Arab Emirates from a provider's perspective. *Worldwide Hospitality and Tourism Themes*, *15*(2), 131-142.
35. Ahmed, F., & Mishra, V. (2020). Estimating relative immediacy of water-related challenges in Small Island Developing States (SIDS) of the Pacific Ocean using AHP modelling. Modelling Earth Systems and Environment, 6(1), 201-214.
36. Groenland, E., & Dana, L. P. (2020). Qualitative methodologies and data collection methods: Toward increased rigor in management research. World Scientific
37. Sushil (2012), "Interpreting the interpretive structural model. Global Journal of Flexible Systems Management, 13(2). 87-106.
38. Prasad, U.C. & Suri, R.K. (2011). Modelling of continuity and change forces in private higher technical education using total interpretive structural modelling (TISM). Global Journal of Flexible Systems Management, 12 (3), 31-40.
39. Jamal, A., Solaiman, M., Alhasan, K., Temsah, M. H., & Sayed, G. (2023). Integrating ChatGPT in medical education: adapting curricula to cultivate competent physicians for the AI era. Cureus, 15(8), DOI: 10.7759/cureus.43036
40. Tan, L. F., Heng, J. J. Y., & Teo, D. B. (2023). Response to:" The next paradigm shift? ChatGPT, artificial intelligence, and medical education". Medical teacher, 1-1. https://doi.org/10.1080/0142159X.2023.2256961
41. Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*. https://doi.org/10.1016/j.iotcps.2023.04.003
42. Rudolph, J., Tan, S., & Tan, S. (2023). War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. *Journal of Applied Learning and Teaching*, 6(1). DOI: https://doi.org/10.37074/jalt.2023.6.1.23

43. Baslom, M. M. M., & Tong, S. (2019). Strategic management of organizational knowledge and employee awareness about artificial intelligence with mediating effect of learning climate. International Journal of Computational Intelligence Systems, 12(2), 1585.
44. Rao, A., Pang, M., Kim, J., Kamineni, M., Lie, W., Prasad, A. K., ... & Succi, M. D. (2023). Assessing the utility of ChatGPT throughout the entire clinical workflow: Development and usability study. Journal of Medical Internet Research, 25, e48659.
45. Zhang, J., Krishna, R., Awadallah, A. H., & Wang, C. (2023). EcoAssistant: Using LLM Assistant More Affordably and Accurately. arXiv preprint arXiv:2310.03046.
46. Mishra, V. (n.d.). Data for AHP and TISM Model for the CUCFATE Framework. Retrieved December 20, 2023, from https://github.com/vinaytosh/datasharing/blob/master/Data_CUCFATE.xlsx

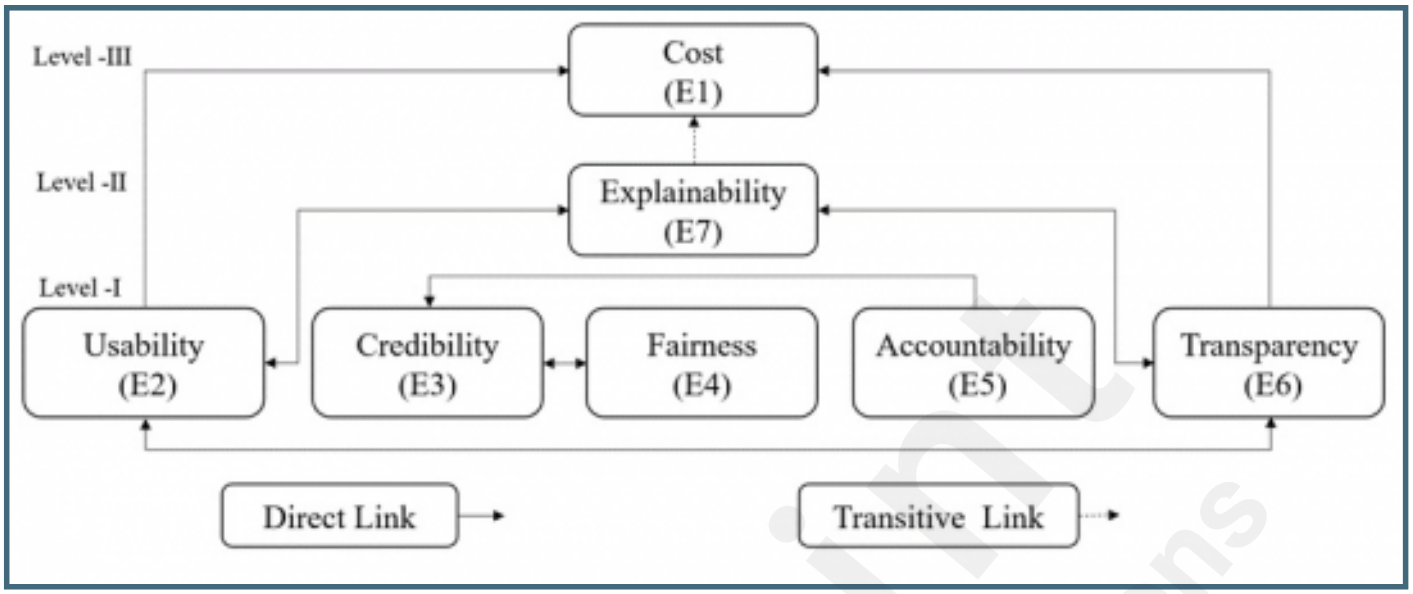
Supplementary Files

Figures

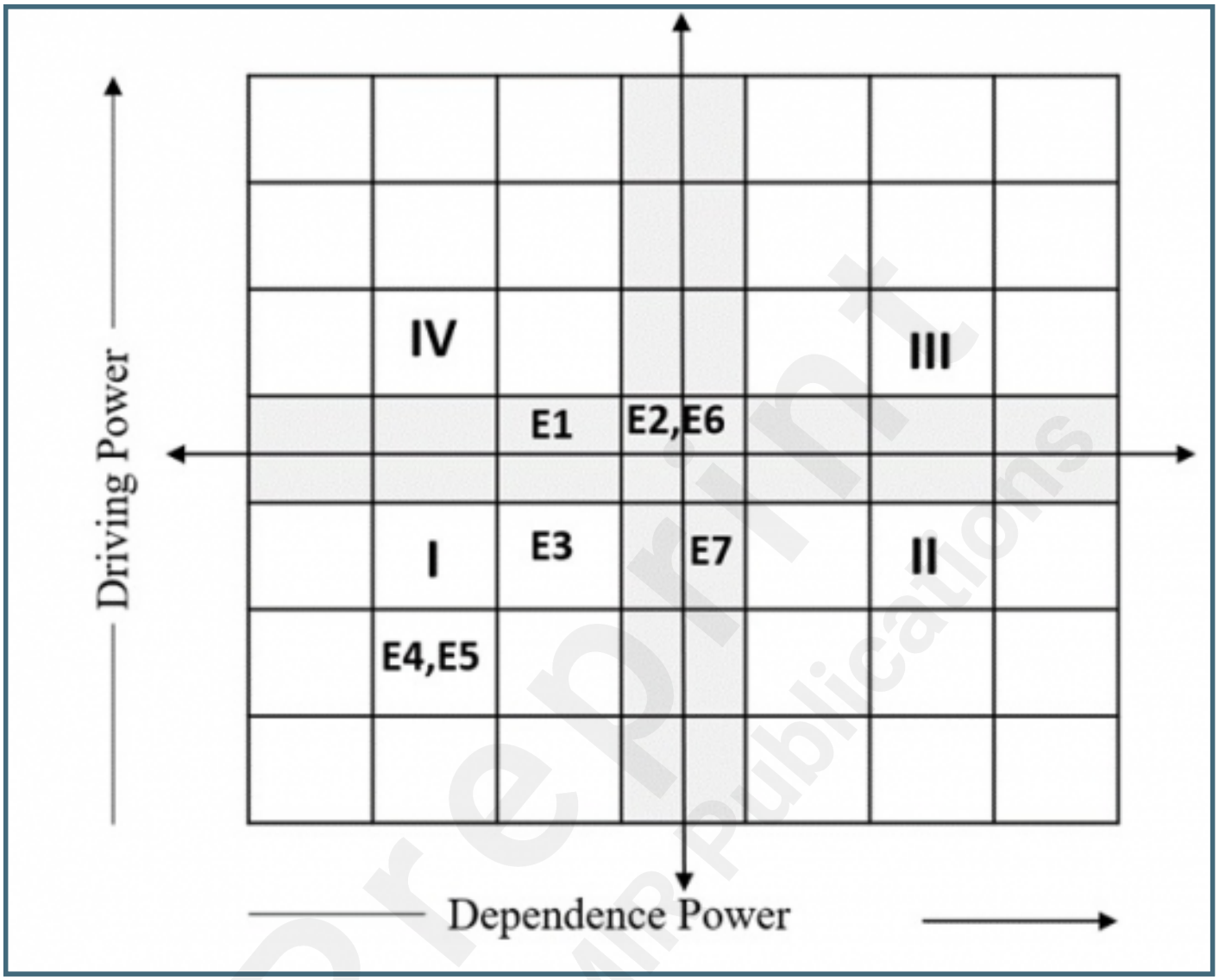
Summary of the TISM Approach Used in the Study.



Diagraph of TISM for LLM in Medical Education.



MICMAC Analysis for Enablers of LLM in Medical Education.



Multimedia Appendixes

Data Source for AHP and TISM Modeling.
URL: http://asset.jmir.pub/assets/15957931c196ef84b1dd285a5a68788c.xlsx

CONSORT (or other) checklists

COREQ Check List.
URL: http://asset.jmir.pub/assets/7a731e773d26e165110d4df188d99dc3.pdf