

Viewpoint

# From Innovation to Impact: The CREATE Framework as a Blueprint for Large Language Model Adoption in Opioid Treatment Programs

Marianthi Markatou<sup>1</sup>, PhD; Raktim Mukhopadhyay<sup>1</sup>, PhD; Jeff Good<sup>2,3</sup>, PhD; Yihao Tan<sup>1</sup>, MA; Arpan Dharia<sup>4</sup>, MD; Lawrence S Brown<sup>5</sup>, MD; Andrew H Talal<sup>4</sup>, MD, MPH

<sup>1</sup>Department of Biostatistics, SPHHP, University at Buffalo, State University of New York, Buffalo, NY, United States

<sup>2</sup>Department of Linguistics, CAS, University at Buffalo, State University of New York, Buffalo, NY, United States

<sup>3</sup>Department of AI and Society, CAS, University at Buffalo, State University of New York, Buffalo, NY, United States

<sup>4</sup>Division of Gastroenterology, Hepatology and Nutrition, JSMB, University at Buffalo, State University of New York, Buffalo, NY, United States

<sup>5</sup>Department of Public Health, Weill Cornell Medicine, New York, NY, United States

**Corresponding Author:**

Marianthi Markatou, PhD  
Department of Biostatistics  
SPHHP, University at Buffalo, State University of New York  
Kimball Tower  
Buffalo, NY 14214  
United States  
Phone: 1 7168292894  
Email: [markatou@buffalo.edu](mailto:markatou@buffalo.edu)

## Abstract

Large language model (LLM)-based systems have tremendous potential to improve patient-centered health care, especially for medically underserved populations. However, realizing this potential requires careful consideration of the socio-technical contexts in which LLMs are used. The design of such systems should consider the vulnerabilities of any medically underserved group that the system intends to support, and provide trustworthy evidence for its use. This viewpoint reports the lessons learned from our experience and research with the CREATE (Culture, Respect, Education, Advancement, Trust, and Expertise) framework for engaging multiple stakeholders to guide the integration of LLM-based systems for improved health care delivery. We propose an extension of the traditional hierarchy of evidence model and identify key junction points in clinical workflows at which LLM-based systems can potentially be used to improve patient care. The key messages can be summarized as follows: (1) users of AI technologies, such as LLM-based tools, must be aware of the strengths, limitations, and impact of these technologies on health care delivery workflows and on the quality of generated evidence on which health care recommendations are based; (2) users must also be aware of the impact of data quality on AI outputs and on the evidence generated from the use of these systems; (3) the evidence pyramid provides a framework that facilitates the evaluation of the output generated by AI systems; (4) the CREATE framework facilitates the engagement of a variety of stakeholders. We propose concrete approaches for its validation and implementation; (5) any system designed to support people with opioid use disorder (OUD) needs to consider the overall lack of trust and stigmatizing experiences of this population with the health care system, and we discuss various aspects of the evaluation process necessary to build trust; (6) to discuss research directions that need to be addressed before LLM-based systems can be integrated usefully in patient health care.

*J Med Internet Res* 2026;28:e97131; doi: [10.2196/97131](https://doi.org/10.2196/97131)

**Keywords:** Large language models; medically underserved populations; opioid use disorder; evidence generation; evaluation; risk; engagement; CREATE framework

## Introduction

Human history has been marked by technological innovation. Generative AI (GenAI), including large language models

(LLMs), has resulted in a technological revolution [1]. While LLMs have tremendous technological potential, we, as a society, must be cognizant of the enormous potential social costs of LLMs without deliberate inclusiveness and

safeguards. Failure to include the entire population, including medically underserved populations, in the AI technological revolution risks increased wage disparities, reduced productivity, and lower government revenue. Medically underserved populations include geriatric individuals, those with learning disabilities, low-income individuals, homeless individuals, and those with substance use disorder. The effects of the digital divide, conventionally referred to as persistent health care disparities in digital health care access [2], are most acutely felt in these populations. The disparities can lead to exclusion from the use of new technologies, such as AI, especially among underserved populations.

For technological innovations to be acceptable to and accessible to the public, we must consider both social and technical aspects. Examples of sociotechnical innovations in health care include electronic health records (EHRs), communication applications, clinical decision support, and telehealth models. The social aspect of these health care tools involves user workflows and organizational policies surrounding their use, which are influenced by individuals' level of digital literacy.

To address the social aspects of the deployment of digital technologies, we adapted the CREATE framework proposed by Talal et al [3]. Our adaptation reflects CREATE as (C=culture, R=respect, E=education, A=advancement, T=trust, E=expertise). The CREATE framework outlines the processes to engage stakeholders in the acceptability, feasibility, and usability of digital technologies in various communities. Implementation of new technologies must consider how individuals will engage with the technology. An engagement framework, such as CREATE, is necessary when considering people with opioid use disorder (OUD), a patient population that is typically stigmatized, medically underserved, and difficult to reach [4]. People with OUD enroll in opioid treatment programs (OTPs) for medical and behavioral treatment of addiction recovery. The recovery journey can be difficult, lengthy, and requires dedicated OTP staff involvement to comprehensively manage patients' progress. New York State was a pioneer in the treatment of OUD [5]. As a result, its procedures for treatment of OUD have been honed over decades. Although in this paper we refer to New York State-based OUD treatment procedures as a use case, substance use disorder treatment requires the collection of extensive clinical information. The lessons learned are transferable to the treatment of substance use disorder and health care more broadly.

In the treatment of OUD, one of the fundamental instruments that inform the creation of treatment plans in OTPs is the psychosocial evaluation form completed within 30 days of admission. The psychosocial form is a comprehensive and detailed history of critical life areas and includes, among other aspects, past and present drug and alcohol use history, prior treatment history, family, legal, and trauma history, and physical and mental health information. The collection of these data and the construction of a patient-centered treatment plan are one of the most important and time-consuming processes of a counselor's responsibilities. Enhancing the value of the psychosocial instrument by

implementing new technologies in the workflows of OTPs has the potential to streamline data collection and mitigate inefficiencies. It might ensure that patient information is collected in a systematic, reproducible, and standardized manner, resulting in high-quality data for use in the development of treatment plans.

The pursuit of high-quality, data-driven evidence enhanced by information generated from complementary sources, such as high-quality qualitative studies, allows individuals to make better choices about health and health care [6]. In the OTP setting, evidence is translated into trust. Before LLM-based systems can be trusted with personal health information, an important objective is to ensure that all stakeholders agree that such systems are error-free, rigorous, and reproducibly generate data-driven evidence. These systems can also be of assistance when collecting data necessary to generate evidence that ultimately leads to improved health outcomes and/or better health care delivery.

The evaluation of technology incorporated into OTP workflows consists of assessing the safety and process compliance of the tools as well as evaluating their effectiveness. Therefore, before embedding LLM-based systems into OTP workflows, a safety evaluation is necessary to avoid and/or correct errors and ensure compliance with institutional metrics and standards (process compliance) [7]. An additional level of evaluation that is required is an assessment of effectiveness. This means that unambiguous and reproducible evidence is obtained that demonstrates improved outcomes. The goal of the evaluation process is a clear demonstration that the embedded technology has a positive impact on the health of individuals and/or the health care delivery process.

A risk-benefit assessment is important to compare potential risks against beneficial outcomes. Identified risks associated with the use of LLM-based systems in OTP workflows include potential biases, hallucinations (ie, confabulation by the LLM-based system), and loss of privacy, security, and confidentiality. Benefits include the potential to streamline administrative tasks, resulting in considerable time savings, and improve patient assessment and education. Some of these tasks tolerate the incurrence of small errors that can be corrected by the human interfacing with these systems, while others, such as patient assessment, must be error-free. The development of quantitative risk-benefit measures ensures informed decision-making. Regulatory processes need to ensure the initial safety of the AI tools with continuous monitoring to guarantee quality and adaptability to current circumstances.

The paper explores important social considerations and evidence generation needs when attempting to integrate LLMs into clinical spaces for health care delivery. Recent investigation within the health care arena has focused on LLMs promoting increased patient-centered health care [7-9], partially through the expression of empathy [9]. Therefore, the primary aims of this viewpoint are: (1) to elucidate the strengths and limitations of using AI technologies, such as LLM-based tools, in health care delivery workflows (this aim is articulated in the remainder of the paper). To

provide recommendations for potential adoption in clinical scenarios with an emphasis on underserved populations to facilitate patient-centeredness (see sections “What are Recommendations to Improve Applicability and Trust of LLMs to Vulnerable Populations?” and “Considerations for Deployment of AI Systems for Underserved Populations”). To discuss the adoption of these systems through the lens of Diffusion of Innovations (DOI) theory [8] (see “LLMs as Socio-Technical Systems: Diffusion of Innovation”). (2) To discuss the evaluation of LLM-based systems from multiple aspects (see “Evaluation of LLM-based Systems”). (3) To discuss the impact of data quality on AI outputs and evidence generated from these systems; and to discuss the impact of using the generated evidence on health care decision-making (see sections “Use of LLM-based Systems for High-Quality Data Acquisition and Improving Clinical Outcomes,” and “Evidence Generation and Risk Assessment in OTPs”). (4) To present and discuss an extension of the traditional hierarchy of evidence model to include the risks associated with deploying AI-based systems at various stages of data collection, exemplified in the case of OTPs (“Evidence Generation and Risk Assessment in OTPs”). (5) To propose the CREATE framework as a facilitator of the participation of all relevant stakeholders in the development and deployment of AI systems in health care workflows and present several considerations to gain acceptance of LLM-based systems by necessary stakeholders to maximize the likelihood of deployment and use (“Engagement Approaches: CREATE Framework”).

## LLMs: Definition, Advantages, and Limitations

According to the US Congress [9], “Large language models (LLMs) are AI systems that aim to model language, sometimes using millions or billions of parameters.” On a more technical level, LLMs are complex neural networks using the transformer architecture and the attention mechanism proposed in [10]. The word “large” usually refers to the large number of parameters (often billions of parameters) used in training the neural network model to process the language. LLMs require considerable computational resources to be trained. LLMs form the core of LLM-based systems designed to provide various kinds of functionality such as machine translation, summarization, and conversational systems. [Multimedia Appendix 1](#) presents a brief history of natural language processing (NLP) along with a timeline of evolution of the field of NLP (Figure S1 in [Multimedia Appendix 1](#)), and [Multimedia Appendix 2](#) briefly discusses details associated with LLMs and applications in health care [11-14].

With newer applications of LLMs emerging every day, it is imperative to be informed of the limitations of these systems as well. Several technical and social challenges are associated with deploying any LLM-based system and their use context. The technical challenges of such systems are discussed in detail in this article from the perspective of accuracy of the generated output and usability

of such systems, which are evaluated primarily on output quality. Unfortunately, their usability is often overlooked in outcome assessment. We emphasize that usability of such systems, including their safety and reliability, is an important consideration because in clinical settings such systems do not exist in isolation but rather as integrated socio-technical systems (STSs) to facilitate patient care (see section “Evaluation of LLM-based Systems”).

Deploying LLM-based tools in health care might also lead to potential social challenges. The authors in [15] describe various risks associated with using LLM-based tools in public health. These risks are even more prominent when serving an underserved population. Given that OTPs function as destigmatizing safe spaces, the implications of introducing LLM-based tools in this environment must be carefully evaluated. The key risks include erecting additional barriers to help-seeking, degradation of patient-provider trust and support systems, missed opportunities to proactively introduce help, and dehumanization and impersonality in care, among others. We provide detailed descriptions of various activities in which LLM-based systems can be used in OTPs while keeping in mind these risks and ensuring that patient-centeredness is not lost.

## LLMs as STSs: DOI

STS is defined as “a complex system that includes both technical and social elements, such as people, technologies, rules, and regulations, that work together to create work processes and products” [16]. The definition of STS immediately points to the fact that LLM-based tools in the context of health care are essentially STSs. The conceptual model for health information technology (HIT) introduced by Sittig and Singh [17] is appropriate for this context. The 8 dimensions of this model are computing infrastructure, clinical content (clinical “data-information-knowledge continuum”), human-computer interface, people, workflow and communication, internal organizational policies, procedures and culture, and external rules and regulations. This conceptual model is appropriate for an LLM-based system as well. In the context of an LLM-based system, the technical component is represented by the computing infrastructure, the model architecture, datasets, and the associated infrastructure required to build such systems. The social component includes system developers, patients, the institution where such a system is deployed, the clinical staff, as well as ethics and any associated regulations. Failure to consider the interdependencies between the different components has the potential to result in the pitfalls described below [18].

1. Framing trap: this trap occurs when a system is deployed without completely understanding the associated social setting and details of the problem the tool aims to solve. The problem formulation needs to incorporate the social context such as people, institutional environments, decision-making processes, and existing regulatory systems.

2. Portability trap: this arises when technologies designed for different social contexts and purposes are transferred to new settings without considering characteristics of the new social setting. This can often lead to harmful consequences.
3. Formalism trap: the evaluation of an STS often requires considering the context in which such systems are used, without which the evaluation of a system is merely mathematical in nature.
4. Ripple effect trap: ripple effects occur due to unintended consequences of introducing new technology to existing social systems (such as an OTP in this case). This can lead to unintended changes in behaviors and values of the existing system.
5. Solutionism trap: the failure to recognize that the solution to a problem might not involve introducing new technology.

The effects of getting stuck in any of these traps in a health care setting are consequential for treatment outcomes and patient health. The successful integration of LLM-based systems into existing settings and workflows depends on understanding the abilities and limitations of the components of the STS, the LLM-based system itself, as well as the social context in which it is deployed. The CREATE framework described later in the paper (Section “Engagement Approaches: CREATE Framework”) emphasizes the social considerations needed to increase the likelihood of full integration of LLM-based systems into OTPs.

To avoid the pitfalls discussed previously, understanding the technical aspects of LLM-based systems alone is insufficient; one must also consider the social system context. We offer insights on the adoption of LLM-based tools in OTPs through the lens of DOI theory. DOI theory has been widely applied to describe the important considerations when introducing any new technology. The adoption of new technology within the health care context is always a multistakeholder approach, which extends beyond the potential technical advantages, considering multiple stakeholders and defining their role in the adoption process. While DOI describes how an intervention spreads within a setting, CREATE serves as a framework for consideration of the social aspects of intervention adoption. Of particular interest in the OTP setting is its resource-scarce nature and the underserved population it serves, leading to additional trade-off considerations between purported advantages and potential loss of patient-centeredness and straining limited resources. We have described in detail the elements and characteristics of innovation (LLM-based tools) in the context of OTPs in [Tables 1](#) and [2](#). The DOI theory highlights that the adoption of LLM-based tools is inherently an interplay between the social and the technical components. The aspects of trust and risk associated with such STS are discussed in a later section in this paper.

**Table 1.** The elements of innovation adapted to large language model (LLM)-based tools in the context of opioid treatment programs.

| Attribute     | Entity                                                                                                                                                            | Relevance                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
|---------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Innovation    | <ul style="list-style-type: none"> <li>• LLM-based tools to improve and aid in patient outcomes.</li> </ul>                                                       | <ul style="list-style-type: none"> <li>• Focus on patient safety in a medically underserved population.</li> <li>• Efficiency and decision-making may improve patient outcomes.</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| Adopters      | <ul style="list-style-type: none"> <li>• Providers</li> <li>• Government &amp; third-party payers</li> <li>• Patients and patient advocates.</li> </ul>           | <ul style="list-style-type: none"> <li>• Clinical staff are crucial for technological integration into clinical workflows. Regulatory authorities can standardize adoption, maintain care quality, and ensure appropriate system use.</li> <li>• Patients and stakeholders provide feedback and ensure effective and ethical use.</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| Time          | <ul style="list-style-type: none"> <li>• AI is in the early stages of adoption in the OTP<sup>a</sup> setting.</li> </ul>                                         | <ul style="list-style-type: none"> <li>• Patient safety is crucial for medically underserved populations.</li> <li>• Patients face stigma and other barriers.</li> <li>• Ensure patient safety through maintaining privacy &amp; data confidentiality, without harm.</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| Communication | <ul style="list-style-type: none"> <li>• System Development</li> <li>• Financial Engagement</li> <li>• Care Delivery</li> <li>• Shaping Policy</li> </ul>         | <p>AI system developers with input from regulatory authorities &amp; clinical staff:</p> <ul style="list-style-type: none"> <li>• Effective communication is important during product development.</li> <li>• Regulatory and health care providers provide necessary inputs to integrate AI systems into existing workflows.</li> </ul> <p>Staff↔staff, staff↔patients, patient↔patient:</p> <ul style="list-style-type: none"> <li>• Communication between staff and patients is essential to build trust in the system.</li> <li>• Effectiveness of the systems depends on integration into existing workflows.</li> </ul> <p>Payers↔treatment providers:</p> <ul style="list-style-type: none"> <li>• Reimbursement policies and support are needed for care provided using LLM-based tools.</li> </ul> |
| Social System | <ul style="list-style-type: none"> <li>• Government and Regulatory Authorities</li> <li>• Financial Stakeholders</li> <li>• Care Delivery Stakeholders</li> </ul> | <ul style="list-style-type: none"> <li>• Successful development, deployment, and use of socio-technical systems in real-world settings depend on the social system.</li> <li>• Engagement with the ecosystem facilitates continuous refinement through feedback.</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |

| Attribute | Entity                                                                                                                | Relevance                                                                                                                                                                                         |
|-----------|-----------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|           | <ul style="list-style-type: none"> <li>Professional Organizations and Advocacy Groups</li> <li>Researchers</li> </ul> | <ul style="list-style-type: none"> <li>The CREATE<sup>b</sup> framework (Figure 1) addresses the social aspects when integrating new technology into substance-use treatment programs.</li> </ul> |

<sup>a</sup>OTP: Opioid Treatment Program.

<sup>b</sup>CREATE: Culture, Respect, Education, Advancement, Trust, Expertise.

**Table 2.** Characteristics of innovation for large language model (LLM)–based tools in opioid treatment programs. Adapted from [8].

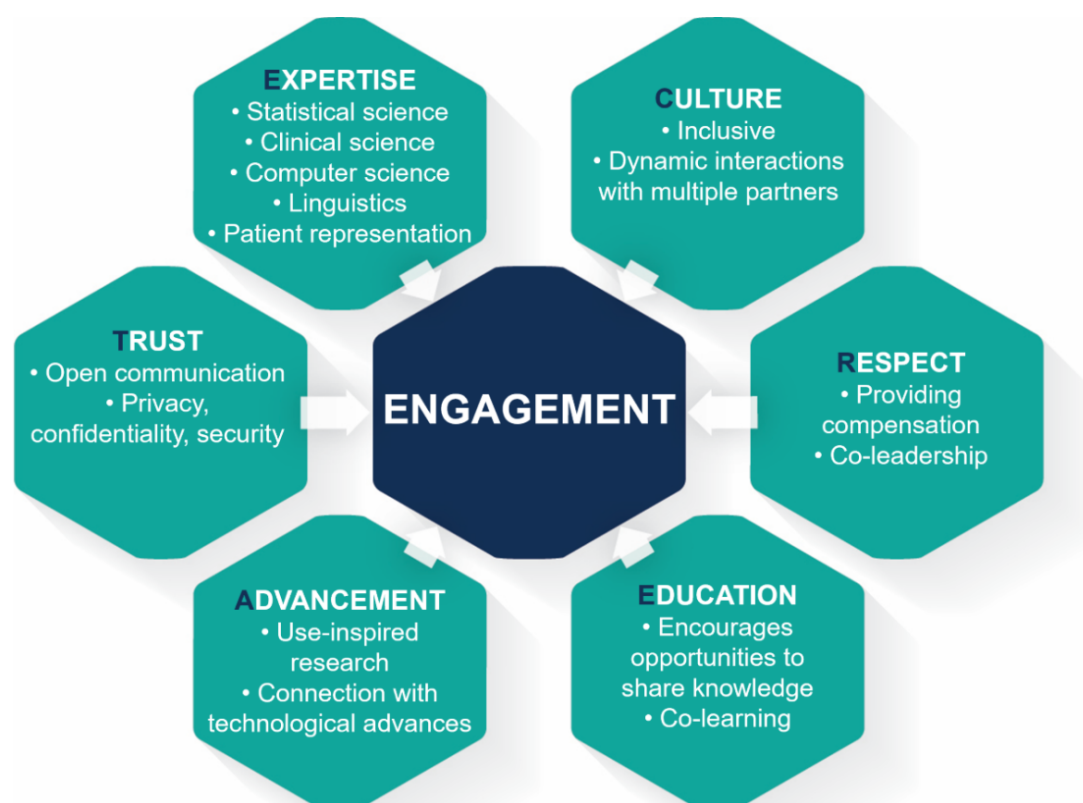
| Attributes         | Relevance                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
|--------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Relative advantage | <ul style="list-style-type: none"> <li>Improved efficiency: LLM-based tools reduce repetitive administrative tasks, freeing providers to focus on clinical tasks.</li> <li>Enhanced decision-making: LLMs can process vast medical texts to aid diagnostic and clinical decision-making. <ul style="list-style-type: none"> <li>LLM-based systems assisting health care providers can potentially improve care quality and reduce errors.</li> </ul> </li> <li>An important consideration is whether these LLM-based systems should be patient-facing or used solely for internal workflows. <ul style="list-style-type: none"> <li>Potential applications: aid in filling out intake forms.</li> <li>Risks include digital literacy in the OUD<sup>a</sup> population.</li> <li>Furthermore, a benefit is the destigmatization that can occur. Patients may share sensitive information more comfortably without discrimination and shame occurring during human interactions.</li> </ul> </li> </ul> |
| Compatibility      | <ul style="list-style-type: none"> <li>New systems must be compatible with existing workflows.</li> <li>Emphasis on usability. Usability focuses on learnability, efficiency, memorability, error prevention &amp; recovery, and user satisfaction.</li> <li>Companies are integrating LLM-based solutions into EHR<sup>b</sup> systems for various uses [19].</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| Complexity         | <p>Applications can be grouped into one of the three groups:</p> <p>Low complexity</p> <ul style="list-style-type: none"> <li>Simple tasks for patient communication are low complexity and have minimal associated risk.</li> </ul> <p>Medium Complexity</p> <ul style="list-style-type: none"> <li>LLM-based systems used for patient education, such as answering queries, providing instructions, and offering health education.</li> <li>Associated risk can be mitigated potentially through grounding the responses on a knowledge base.</li> <li>The system should not be detrimental to human health.</li> </ul> <p>High Complexity</p> <ul style="list-style-type: none"> <li>Examples include diagnostic and decision-making support.</li> </ul> <p>LLM-based systems must ensure safety, build trust, ensure usability, risk education, and effective training for staff and patients for seamless integration.</p>                                                                        |
| Trialability       | <ul style="list-style-type: none"> <li>Most LLM-based applications are not presently implemented in OTPs<sup>c</sup>, indicating an opportunity for adoption in this domain.</li> <li>Predeployment trials are essential before deploying any AI system, especially in health care, where errors are unacceptable. <ul style="list-style-type: none"> <li>Community partnerships (eg, between OTPs and research universities) can facilitate these trials.</li> <li>Active stakeholder engagement (eg, site visits and stakeholder engagement) is needed for extending, addressing, and facilitating large-scale implementation.</li> </ul> </li> </ul>                                                                                                                                                                                                                                                                                                                                                |
| Observability      | <ul style="list-style-type: none"> <li>Current implementations of LLM-based systems directed toward the OUD population or OTPs do not exist.</li> <li>Evidence from other areas in medicine highlights the potential of LLM-based tools in improved patient-doctor communication [20], generating patient-friendly and accurate notes [20], enhanced diagnostic accuracy in pulmonology and endocrinology [21], pathology report explanation [22], medical exam recommendations and diagnosis [23], clinical decision support systems [24].</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                                 |

<sup>a</sup>OUD: opioid use disorder.

<sup>b</sup>EHR, electronic health record.

<sup>c</sup>OTP: Opioid Treatment Program.

**Figure 1.** CREATE (C=culture, R=respect, E=education, A=advancement, T=trust, E=expertise) framework for engaging stakeholders in the integration of AI within health care.



## Engagement Approaches: CREATE Framework

To guide engagement of stakeholders with LLM-based systems, we extended the CREATE framework developed in [3] (Figure 1). The CREATE framework outlines the processes to address the social aspects of technological integration into a community. Elucidating clinical workflows for patient care facilitates understanding of the OTP culture. Respect for OTP staff and patients is manifested by incentives, avoiding stigmatizing language, and through co-leadership decisions. Education about technology, addressing anxiety through knowledge, and co-learning between patients and staff promote a respectful attitude toward technology. Education also promotes technological advancement through invention and the application of new tools. Promoting trust among patients and staff in LLM-based tools is required

for deployment. LLM use is enhanced by ensuring privacy, confidentiality, and security of the technology. Successful LLM deployment and use require expertise in multiple disciplines, and the CREATE framework can address the social aspects of technological deployment in underserved communities. Considerations for evaluation matrices of each of the 6 CREATE framework domains are illustrated in Table S1 in [Multimedia Appendix 3](#). Consideration of the CREATE framework domains, as judged against conventional implementation frameworks, is listed in Table S2 in [Multimedia Appendix 3](#).

As a major funder of comparative effectiveness research, the Patient-Centered Outcomes Research Institute (PCORI) has developed 6 foundational aspects of research partnerships. [Table 3](#) illustrates how the CREATE framework aligns with PCORI's 6 foundational expectations for partnerships in research [25].

**Table 3.** Foundational aspects of research partnership.

| Category                         | Definition                                                                                         | Relevance to CREATE <sup>a</sup> framework                                                                        |
|----------------------------------|----------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------|
| Representative involvement       | Inclusion of partners and organizations that reflect a range of affected patients and communities. | Multidomain expertise is required for engagement success and promoting advancement.                               |
| Build capacity to work as a team | Identify strengths and obstacles to engagement; provide education to address obstacles.            | Understanding culture and education can identify engagement opportunities; obstacles are overcome with education. |
| Early and ongoing engagement     | Responsible party engaged throughout the technology lifecycle                                      | Understanding culture illustrates how and where to engage.                                                        |

| Category                                                | Definition                                                              | Relevance to CREATE <sup>a</sup> framework                                   |
|---------------------------------------------------------|-------------------------------------------------------------------------|------------------------------------------------------------------------------|
| Ongoing review and assessment of engagement             | Perform continuous assessment and evaluation of successes and failures. | CREATE provides an overview of how engagement might be bolstered or modified |
| Meaningful inclusion of partners in decision-making     | Use approaches to include multiple partners in decision-making          | Principle promotes co-learning and multidomain expertise.                    |
| Dedicated funds for engagement and partner compensation | Resources to compensate for time and effort.                            | Funds can promote respect and trust as CREATE components.                    |

<sup>a</sup>CREATE: C=culture, R=respect, E=education, A=advancement, T=trust, E=expertise, explains social aspects of technology deployment in a community.

## What Are Some of the Functions of LLM-Based Systems in OTPs?

LLM-based systems have shown good performance in several use cases. Some examples include responding to prevention questions from patients in cardiology [26], hip replacement [27], and radiology report findings [28]. Novel experiments integrating LLMs as clinical decision support tools are needed to evaluate their effect on outcomes, productivity, and patient satisfaction [29]. Existing literature has evaluated LLMs' abilities to respond to portal patient messages [30], generate discharge summaries [31], generate structured templates for radiology [28], and support the management of breast cancer tumor boards [32]. Furthermore, LLM-based systems have been successfully implemented in diagnostic image processing and clinical decision support [33]. They play a role in social media monitoring and track trends in OUD that correlate with real-time morbidity and mortality reporting [34]. AI can also be useful in the development of medical guidelines [35] and in the construction of systematic reviews [36].

Part of the challenge of integrating LLM-based systems into clinical workflows is that minor changes in the inputs produce unpredictable outputs [29]. Furthermore, the technology is evolving rapidly, and as it becomes more widespread, a potential issue is overreliance on its output while negating its limitations and biases. As an alternative to AI in the application to health care, the American College of Physicians recommends the term "augmented intelligence" when referring to the role of AI in clinical decision-making. The objective is to promote the concept that human intelligence continues to be central to the clinical decision-making process even when integrating AI, and that AI is a tool to assist clinicians [37,38].

LLM-based systems can be used to address the digital divide. We can use knowledge generation and educational functionalities to enhance digital literacy among patients. LLM-based systems can also be used to combine data on a particular patient within an electronic medical record and to support prior authorizations for medications. In these situations, it would be helpful to have the human in the loop [39] to be able to verify information. Human-in-the-loop (HITL) refers to the integration of an expert into a machine learning or AI workflow to improve the quality of the system's intended outcomes. In the health care context, expert knowledge and experience are essential, requiring the augmentation of skilled experts, rather than subordinating

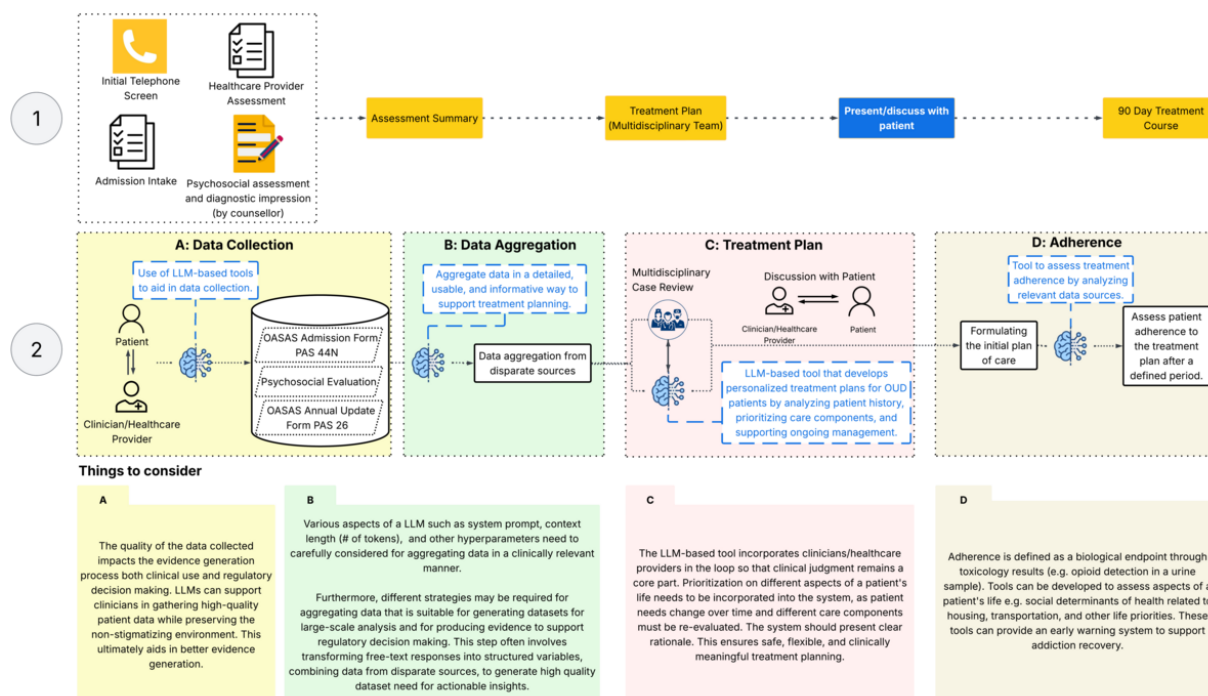
their skills. Similar to HITL, the terminology "Doctor-in-the-loop" has also been coined [40]. Several works [41] have explored the incorporation of expert knowledge to improve the trustworthiness of AI system-based workflows. In health care, transparency is one of the key requirements for LLM-based systems, which can then support the HITL framework to increase trust in the system [42].

Appropriate LLM-based systems might promote patient-centeredness. LLMs can de-stigmatize language and can suggest alternative language choices. The goal is to promote more empathetic language among health care providers and in public health messaging [43]. LLMs can potentially play an important role in patient and staff education. They can develop culturally and linguistically relevant educational materials, potentially improving patient understanding, medication adherence, and trust in providers [44]. LLMs, in combination with structured EHR data, can develop models that predict treatment attrition, enabling targeted interventions [45]. They may have a potential role in OTP staff training and simulation. LLM-driven virtual patient agents can simulate realistic patient-provider interactions, offering a controlled and ethical environment for training OTP staff in therapeutic dialogue strategies for addiction recovery [46]. The goal is to promote LLM-directed patient-centered messaging and education.

## Potential Use of LLM-Based System in Treatment Plan Development and Implementation

As presently performed, the generation of a client's treatment plan is a relatively lengthy and laborious process as described in Figure 2.1. When a patient is admitted to an OTP in New York State, the first point of contact is an initial phone screen that gathers basic demographic information and assesses patient eligibility for treatment of OUD. The next step is an admission eligibility assessment by a physician or advanced practice provider. Once the patient is admitted to the OTP, the clinician completes an admission form that includes more extensive information on drug use, residential situation, family history, and information on employment and education. The patient is then assigned to a primary counselor within 24 hours who schedules an initial session to obtain the psychosocial form, which must be completed within 30 days of admission (email, personal communication, Office of Addiction Services and Supports [OASAS], New York, March 31, 2026).

**Figure 2.** (1) Inputs to and Development of Treatment Plan: The steps for the development of the treatment plan are outlined in the text. Yellow highlight indicates opportunities for task coordination with large language models. (2) Extension of Figure 2.1, which highlights the key junctures where AI-based systems can be leveraged in the day-to-day Opioid Treatment Program workflow. Additionally, we have emphasized certain aspects of AI-based systems that need to be considered before such systems are implemented in real-world settings. For each distinct phase, we illustrate the role of the AI-based system in the blue box. The interaction of various components of AI-enhanced Opioid Treatment Programs is depicted using arrows. Each stage of the workflow can have different extents of patient centeredness embedded within it depending on the patient needs. LLM: large language model; OASAS: Office of Addiction Services and Supports.



Once the psychosocial form has been completed, data are aggregated from the 4 sources illustrated in Figure 2.1 to develop an assessment summary. Upon completion of the assessment summary, the multidisciplinary team meets to develop a treatment plan. The next step is to present and discuss the treatment plan with the patient. Following the discussion with the patient, an initial 90-day treatment course is implemented.

There are several steps at which LLM-based systems can be involved during the creation and implementation of the treatment plan, as illustrated in Figure 2.2. LLM-based systems can be integrated into the data collection stage to facilitate the collection of higher-quality data to inform patient care. Additionally, LLM-based systems can be used to aggregate the collected data from different sources with varied structures to develop treatment plans that consider the unique circumstances of a patient, prioritizing specific aspects of the treatment and supporting ongoing addiction management. Finally, such systems can be used to aid treatment adherence, which is known to be a difficult challenge in the case of addiction recovery. Important considerations for integration of LLM-based systems in different areas of the workflow are also presented.

Furthermore, an empathetic LLM-based system could potentially administer the psychosocial assessment and develop its clinical summary. LLMs could also develop the assessment summary through the aggregation of data from multiple sources. In terms of developing a treatment plan, LLMs could prioritize the most important aspects and assess

their severity. The multidisciplinary team could ultimately assess and verify the output of the LLM-based system, similar to the HITL approach described previously. Finally, the LLM-based system can support patient adherence to the 90-day treatment plan.

## Use of LLM-Based Systems for High-Quality Data Acquisition and Improving Clinical Outcomes

In the previous section, we have illustrated how LLM-based systems can be used for developing individualized treatment plans for patients with OUD. One of the major components of developing these treatment plans is collecting data using high-quality instruments. The goal of using LLM-based systems is to generate individualized treatment plans dedicated to specific patient types and reduce associated errors.

From a higher level of abstraction, an important research interest is how LLM-based systems or LLMs can be used for processing vast amounts of data to inform better patient care. OTPs collect massive amounts of textual data consisting of clinical narratives, patient-reported information, and other records. LLMs can be used for converting free text into structured datasets, performing data extraction, and other data processing tasks. These procedures broadly aid in dataset generation that informs care about specific populations.

Table 4 lists the various proposed applications of LLM-based systems along with associated evidence of their implementation in health care situations and briefly indicates whether HITL might be required for oversight.

**Table 4.** Various proposed applications of large language model (LLM)–based systems in health care along with associated evidence and need for human-in-the-loop oversight.

| Proposed application using LLM-based system                    | Evidence | HITL <sup>a</sup> required (Yes or no)                                                                                                                          |
|----------------------------------------------------------------|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Patient portal messaging                                       | [30]     | <ul style="list-style-type: none"><li>Scheduling does not require HITL.</li><li>Health care inquiries need HITL validation.</li></ul>                           |
| Clinical decision support systems                              | [47,48]  | <ul style="list-style-type: none"><li>Yes</li></ul>                                                                                                             |
| Destigmatizing language and suggesting empathetic alternatives | [43,49]  | <ul style="list-style-type: none"><li>Yes</li></ul>                                                                                                             |
| Patient and staff education using relevant materials           | [11]     | <ul style="list-style-type: none"><li>Partial. HITL is needed for verification of information accuracy and updating of knowledge base.</li></ul>                |
| Predicting treatment, attrition & targeted interventions       | [45]     | <ul style="list-style-type: none"><li>Yes</li></ul>                                                                                                             |
| Patient simulation                                             | [46]     | <ul style="list-style-type: none"><li>Partial. HITL is needed for updating and verification.</li></ul>                                                          |
| Administering intake questionnaires                            | [50]     | <ul style="list-style-type: none"><li>Yes</li></ul>                                                                                                             |
| Generating clinical assessment summaries                       | [51]     | <ul style="list-style-type: none"><li>Yes</li></ul>                                                                                                             |
| Free text to structured report generation                      | [52]     | <ul style="list-style-type: none"><li>Yes</li></ul>                                                                                                             |
| Developing treatment plans                                     | [53,54]  | <ul style="list-style-type: none"><li>Yes</li></ul>                                                                                                             |
| Supporting patient adherence to treatment plans                | [55]     | <ul style="list-style-type: none"><li>Dependent on the magnitude of the task. For simple and routine verification of adherence, HITL is not required.</li></ul> |
| Enhancing digital literacy                                     | Proposed | <ul style="list-style-type: none"><li>Partial</li></ul>                                                                                                         |
| Data aggregation from various sources to produce a new dataset | Proposed | <ul style="list-style-type: none"><li>Yes</li></ul>                                                                                                             |

<sup>a</sup>HITL: human-in-the-loop.

## Evaluation of LLM-Based Systems

Prior to the integration of any LLM-based systems, or rather any system in the health care context, understanding associated challenges and evaluating risks vs benefits is of utmost importance. These challenges impact the reliability of generated output, and consequently the quality of evidence

and hence need to be addressed to ensure safety and trust. A summary of challenges in the health care context is given in Table 5.

The challenges discussed in this section necessitate a robust evaluation strategy considering multiple aspects described in the next section.

**Table 5.** Key challenges, descriptions, and potential mitigation strategies for use of large language model (LLM)–based systems in health care.

| Challenges       | Description                                                                                                                                                                                                                                                                                                                                                                                                                                   | Mitigation and solution                                                                                                                                                                                                         |
|------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Hallucinations   | Producing outputs that are factually incorrect and misleading [56]. Divided into factual (discrepancy between generated output and real-world information) or faithfulness (discrepancy between generated output and provided instructions). These issues arise due to training data with outdated or false information [57]. Validation fatigue and optimization for answers that look correct make detecting hallucinations difficult [58]. | Retrieval augmented generation (RAG), where responses are based on external knowledge bases [59]; several techniques described in [56].                                                                                         |
| Sycophancy       | The tendency of LLMs to excessively agree with users [60].                                                                                                                                                                                                                                                                                                                                                                                    | Implementing guardrails, ie, controlling the output of an LLM to respect some human-imposed constraints [61].                                                                                                                   |
| Black-box nature | Poses a significant challenge to explainability of the model outputs.                                                                                                                                                                                                                                                                                                                                                                         | Potential techniques for improving explainability for LLMs can be found in [62].                                                                                                                                                |
| Data protection  | Systems must strictly adhere to global data protection laws to ensure patient privacy and confidentiality are not compromised.                                                                                                                                                                                                                                                                                                                | Strict adherence to laws like Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR).                                                                                        |
| Nondeterminism   | Poses a serious threat to reproducibility. Arises primarily due to a combination of floating-point non-associativity and parallel execution [63].                                                                                                                                                                                                                                                                                             | Quantifying the uncertainty associated with LLM-based tools is essential for building trust and consequently their large-scale adoption and use in health care. Potential research in these directions is discussed in [64-66]. |

# Multidimensional Evaluation of LLM-Based Systems

## Overview

The American Medical Informatics Association (AMIA) recommends evaluation of AI systems in health care through multiple lenses, namely technical performance, health impact (effectiveness), and usability and workflows (operational efficiency) [67]. This emphasizes that evaluation of LLM-based systems must not be based on technical aspects alone; rather, it must account for the STS in which it operates.

We describe these evaluation aspects below for clarity. We would like to note that these aspects are not isolated; rather, they are interdependent. For the purposes of this article, we consider usability under technical evaluation, but it is an overlapping concept that links the technical implementation of the system with its adoption and hence effectiveness.

## Technical Evaluation

The technical evaluation should consider the following factors affecting LLM outputs:

1. Parameters: LLM outputs depend on several user-defined parameters such as temperature, top-p, frequency penalty, presence penalty, and thinking or reasoning levels, which in turn affect the output. For instance, how does a set of parameters, for example,  $\theta$ =[temperature, thinking level, frequency penalty], impact the output  $O$ ? Understanding the impact of these parameters is an important consideration for our work. Prompting also plays an important role in the output of LLM-based systems. Various prompting techniques, such as described in [68], need to be investigated depending on the task.
2. Repeatability and faithfulness: few works [69] have explored concepts such as:
  - a. Semantic repeatability: the consistency of LLM responses across repeated runs under identical conditions.
  - b. Semantic reproducibility: the consistency across runs under different conditions.
3. HITL validation approaches are needed to ensure that evidence generated by LLM-based systems is accurate.
4. Usability evaluation: the definition of usability as given in ISO 9241-11 is “The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use” [70]. Usability evaluation of LLM-based systems is needed to understand whether the tool is practical, intuitive, and readily adoptable by clinical staff without introducing significant workflow hurdles or cognitive burdens. Several usability assessment techniques are available such as the Nielsen usability checklist, questionnaires, interviews, and observational studies, the System Usability Scale, usability testing, cognitive walkthrough, etc. Dehghani et al [71] present a review of methods used in usability evaluation of hospital information systems.
5. Computational reproducibility for provenance: computational reproducibility is defined [72] as ensuring that the same results using the exact same raw materials, computational steps, codes, and conditions as the original analysis can be obtained, and is the foundation for system provenance. In the context of LLM-based health care systems, provenance requires recording and tracking the entire lifecycle of an input (eg, patient data) and its corresponding output (eg, a treatment recommendation). Provenance in case of health care systems is extremely important for legal and regulatory accountability as well as for ensuring that the system behaves as intended. LLM-based systems should be able to trace back how a certain output was generated. LLM-based systems should be developed while keeping this in mind, and in turn, must be evaluated based on it as well.
6. Postdeployment monitoring: an analogue of postdeployment monitoring is postmarketing surveillance of medical products, which emphasizes the need to monitor medical products once they have been placed on the market following rigorous clinical trials. Similarly, to understand the performance of an LLM-based system and address potential shortcomings in real-world settings, postdeployment monitoring is of crucial importance to ensure that the system supports its intended users for its intended purpose. When LLM-based systems are deployed as a component of clinical workflows, several additional mechanisms are required simultaneously to ensure their holistic integration. Several frameworks have been proposed emphasizing various aspects of AI governance [73–75]. Regarding postdeployment monitoring, Keyes et al [73] propose 3 complementary principles: system integrity, performance, and impact. System integrity addresses information technology components, such as maximizing system uptime, detecting run-time errors, and mitigating unintended consequences. Activities essential for performance monitoring include logging, auditing, evaluating bias, accuracy, predictability, transparency, and version control. Impact monitoring focuses on evaluating the value of the system and may lead to recalibration, rollback, suspension, or retirement when deployed systems no longer perform adequately. Additional administrative and institutional policies are needed for incident management and response. El Arab et al [75] found that trustworthy AI implementation in health care settings is increasingly associated with continuous “lifecycle governance” as compared to predeployment validation only. Lifecycle governance is also a cornerstone of continuous quality improvement, which represents iterative improvement of “processes, safety, and patient care” [76]. However, there is a lack of formal consensus, and practical challenges remain regarding metric selection, thresholds for action, review frequency, corrective response, and institutional accountability. These activities align with guidance from World Health Organization (WHO), which recommends that

governments should introduce mandatory postdeployment auditing and impact assessments [77].

Additional aspects of technical evaluation should also be considered, keeping in mind the context in which the system is deployed. For example, whether the output adheres to specific guardrails enforced in the system.

### ***Operational Efficiency (Workflows)***

The objective for implementing any LLM-based system in a health care setting is to improve operational efficiency without compromising patient care. Given the administrative workload in OTPs, the system must not introduce additional cognitive overload as discussed in the usability evaluation. The LLM-based system should be evaluated on its ability to streamline workflows. Efficiency in this context is measured by how well the tool reduces the time and cognitive resources of clinical staff in accomplishing a task as compared to currently used solutions (baseline).

### ***Effectiveness***

The implementation of any new system must prioritize and enhance patient care and/or health care delivery. In an OTP setting where LLM-based systems could be used to assist or potentially develop treatment plans, an error-free, highly accurate, and efficient performance is a requirement.

Evaluating health care delivery workflows involves analyzing clinical and administrative processes with the goal of identifying bottlenecks, reducing errors, and ultimately enhancing clinical care. The goal is to evaluate the impact of LLM-based workflows on health care outcomes. Several methods that evaluate technology-enhanced workflows, for example, the use of EHRs, exist in the literature. In the OTP context, if we desire to compare the effectiveness of LLM-based versus standard workflows, we could propose a noninferiority clinical trial using a biological outcome, for example, quantity of opioids measured in urine (ie, toxicology results). In designing trials of clinical effectiveness, one needs to understand the clinical workflows where LLM-based

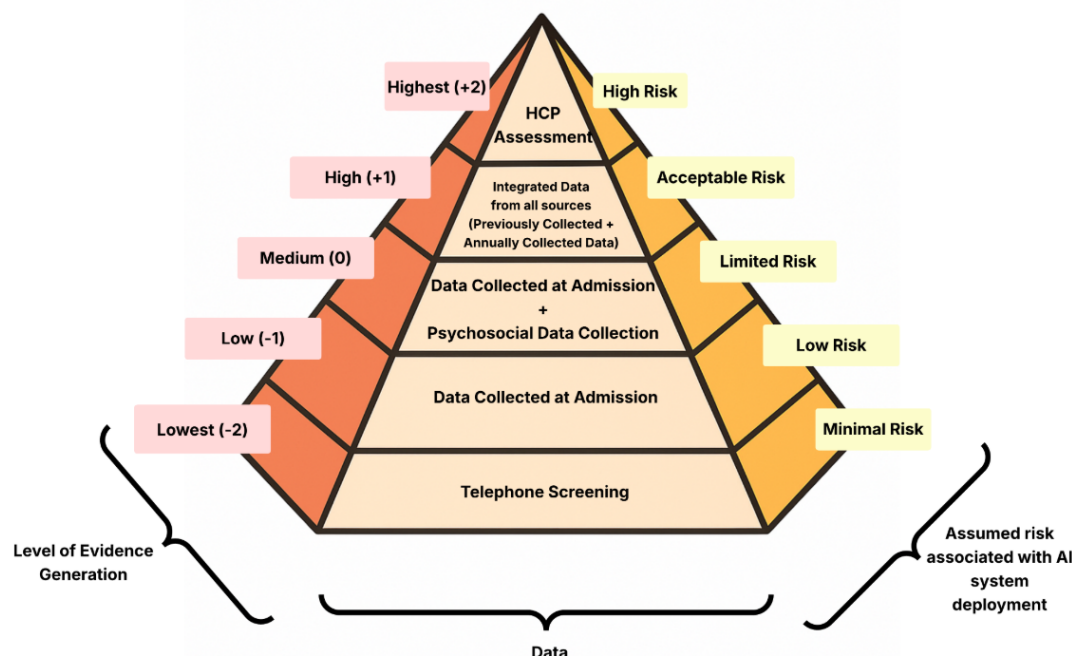
systems will be implemented and the factors beyond the technical capabilities of LLMs to bridge the gaps between implementation and adoption by OTP staff.

## **Evidence Generation and Risk Assessment in OTPs**

Given the potential benefits and risks of implementing LLM-based systems in health care programs, decisions about their adoption must be guided by a balanced assessment of both dimensions. Beyond the general concerns discussed in prior sections, levels of risk of LLM integration in a health care program should be assessed critically based on the function of the model and the type of input data. For a better visualization of the benefits and risks aspects of LLM-based systems, the pyramid model presented in [Figure 3](#) illustrates the central rationale for LLM integration in OTPs. The deeper the integration of an LLM model with more clinically dedicated tasks and the richer the data sources, the higher the level of evidence obtained, and consequently, the higher the risk associated with LLM-based systems ([Figure 3](#)).

Telephone screening is the foundation of the pyramid, which typically occurs before formal OTP admission. This interaction involves acquiring basic demographic information and assessing patient eligibility for OUD treatment. The LLM-based system functions (eg, summarizing call transcripts, extracting key variables, and generating reminders for follow-up) work with the lowest level of input. As a result, errors from LLM-based systems at this stage rarely affect immediate clinical decisions, so the risk is minimal. Thus, we obtain the lowest level of evidence generation (–2) in the pyramid model. During OTP admission, programs collect patient information on psychosocial forms. The pyramid illustrates that, at each subsequent level of risk, the complexity of clinical information increases. Simultaneously, the risks and consequences such as a breach become increasingly costly.

**Figure 3.** A multifaceted pyramid of data, evidence generation, and risk associated with AI systems in an Opioid Treatment Program setting. The front face, representing data, highlights the hierarchy of data available at Opioid Treatment Programs for opioid use disorder treatment. When an opioid use disorder patient is first admitted, a screening phone call is conducted, acquiring preliminary data (lowest quality). Gradually, as the patient is incorporated into the Opioid Treatment Program, collection and aggregation of additional sources of data take place, which leads to a higher quality of data. Corresponding to each level of data quality, there is an associated level of evidence generation (left face). The right face of the pyramid represents the risk associated with deploying AI-based systems in each stage of data collection or aggregation, depicted in the front face. HCP: health care provider.



## Regulatory Aspects

Policy frameworks for AI are currently in a state of rapid and continuous change. Regulations and laws written to govern the use of AI neither address the challenges nor the opportunities associated with such systems [77]. In July 2024, the European Union adopted Regulation (EU) 2024/1689, commonly known as the EU Artificial Intelligence Act, thereby establishing the first AI regulation as a comprehensive framework to regulate AI systems [78]. The act includes a risk-based regulatory framework that differentiates obligations for AI developers and deployers according to the level of risk posed by specific systems [78-80], categorizing the risk into unacceptable risk, high risk, transparency risk, and minimal risk. For further information see [78].

In the United States, there are no federal laws dedicated solely to AI. The existing framework relies on a patchwork of device, data privacy (such as the Health Insurance Portability and Accountability Act [HIPAA]), and liability laws. However, the regulatory landscape for AI in the United States is currently undergoing a period of rapid and significant transition [81, 82]. In December 2025, an executive order was issued to foster a more unified national environment for AI by aligning federal and state policies to ensure a supportive regulatory framework for AI innovation [83]. Over a dozen states have enacted laws designed to ensure safeguards and guardrails for the use of AI. For example, New York State's RAISE Act outlines key AI risks including cybersecurity threats, system failures, criminal contact, weaponization, and critical harm resulting in death or serious injury. For more

information, please see the RAISE Act (S.53/A.6453) [84]. The act also has reporting requirements for managing risks including safety protocols, third-party reporting requirements, and auditing procedures. California has also enacted a similar law "Transparency in Frontier Artificial Intelligence Act (SB53)" that is designed to enhance online safety and install guardrails on the development of frontier AI models. Other governments are rapidly developing new laws or regulations or adopting targeted restrictions [77].

Companies are expected to release successively more powerful and capable LLM-based systems in the coming months, which may introduce new benefits but also new regulatory challenges. While these laws and regulations provide general guidelines for the use of AI systems, one must keep them in mind when developing LLM-based systems targeted at underserved populations.

## What Are Recommendations to Improve Applicability and Trust of LLMs to Vulnerable Populations?

While there are many reporting guidelines for the use of AI tools in medicine [85] and many relevant consensus statements [86], it is imperative that the adoption of AI systems is based on robust, reproducible evidence obtained via scientifically sound evaluation processes.

When introducing new technologies to people with OUD, recommendations to build patient-doctor trust include stigma

minimization, promotion of empathy in the patient-provider relationship, and engaging community organizations to build bridges with health care providers and institutions. With regard to LLMs, to guard against bias and hallucinations, rigorous benchmarks must evaluate LLM-based systems using medical data and tasks, so that reproducibility can be evaluated and developmental progress tracked [29]. These systems must also be evaluated against, for example, frameworks that have emphasized ethical aspects, such as respecting human values and being inclusive [87] and promoting transparency in AI use [88,89]. Humans should have ultimate oversight to ensure trustworthy AI outputs.

The unknown nature of some AI models, as the fundamental “black box” design of contemporary LLMs [90], can promote distrust among providers [33]. In the development of these tools, clinical safety, effectiveness, and health equity must be top priorities. A continuous improvement feedback mechanism should be used to promote improvements in AI tools and to report adverse events resulting from the use of AI. LLM-based systems may be able to assess for potential

biases and inequity in OUD treatment plans across different patient demographics (race, ethnicity, and sex). These systems may also have a potential future role in promoting equitable OUD management decisions [91].

## Considerations for Deployment of AI Systems for Underserved Populations

Policies toward LLM-based system implementation and deployment in health care settings that care for underserved populations have several considerations (Table 6). Consideration of the domains of engagement as well as the different stakeholders is very important. Particularly important are governmental agencies that can encourage adoption of LLMs through financial incentives or reimbursement for their use. Correspondingly, they also have a role in ensuring security, confidentiality, and privacy of LLM-based systems.

**Table 6.** Recommendations for implementation of large language model (LLM)–based systems for underserved populations by stakeholder position.

| Domains of engagement  | Stakeholders                                                                                                                                                                                                                                                                        |                                                                                                                                                          |                                                                                                                                                                                                                        |
|------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                        | Opioid treatment program (patients, leadership, administrators, and staff)                                                                                                                                                                                                          | Academia and industry                                                                                                                                    | Government                                                                                                                                                                                                             |
| Patient care workflows | <ul style="list-style-type: none"><li>• OTP<sup>a</sup> staff explore the effectiveness of LLMs as tools to promote patient engagement.</li></ul>                                                                                                                                   | <ul style="list-style-type: none"><li>• Promoting research on areas where LLMs are effective in improving patient health outcomes.</li></ul>             | <ul style="list-style-type: none"><li>• Evaluating aspects of the OASAS<sup>b</sup> network where LLMs can improve patient health outcomes.</li></ul>                                                                  |
| Organizational design  | <ul style="list-style-type: none"><li>• OTP staff participate in education on the use and dangers of LLMs.</li><li>• OTP staff evaluate and provide feedback on LLM deployment.</li><li>• OTP staff evaluate whether LLMs can safely reduce staff loads in OTP workflows.</li></ul> | <ul style="list-style-type: none"><li>• Understanding OTP culture and community to avoid potential disruptive effects from the uptake of LLMs.</li></ul> | <ul style="list-style-type: none"><li>• Identify areas where LLMs can safely and effectively assist within OTP workflows.</li><li>• Potentially provide incentives or reimbursement to promote safe LLM use.</li></ul> |
| Policy                 | <ul style="list-style-type: none"><li>• OTP leadership establishes clear guidance on the safe use of LLMs in patient care.</li></ul>                                                                                                                                                | <ul style="list-style-type: none"><li>• Collaborate to explore how LLM deployment can enhance patient outcomes.</li></ul>                                | <ul style="list-style-type: none"><li>• Responsible use of LLMs that is, focus on security, confidentiality, and privacy.</li></ul>                                                                                    |

<sup>a</sup>OTP: Opioid Treatment Program.  
<sup>b</sup>OASAS: Office of Addiction Services and Supports.

## Conclusions and Future Research Directions

Technological innovation is a facet of the human condition and has occurred ever since the evolution of humanity. Consideration of how to increase trust and respect in technology is vital, universal, and extremely important, especially when revolutionary technologies, such as AI, are considered.

In this paper, we discuss the adoption of new technology through the lens of DOI theory and extend the CREATE engagement framework to facilitate the implementation of LLM-based systems into OTP workflows. Future research should seek to evaluate the effectiveness of the CREATE

engagement framework as related to LLM deployment. Another area for future investigation is evaluation of implementation outcomes, as would conventionally be assessed using standard implementation frameworks, such as Reach, Effectiveness, Adoption, Implementation, and Maintenance (RE-AIM) [92] or Consolidated Framework for Implementation Research (CFIR) [93]. Evaluation matrices for each of the 6 components of the CREATE framework are listed in the Multimedia Appendix 3 and could serve to highlight specific areas for future research inquiry.

Advances in foundation models, such as multimodal processing, increased planning capabilities, and subsequently the development of autonomous LLM agents with the ability to interact with a wide array of tools, emphasize the need for the CREATE framework even further. “Advancement” drives

the adoption of these newer innovations into various aspects of patient care, while “trust” ensures that it is done under rigorous oversight by giving patient safety and confidentiality the highest priority. At the same time, these advances have led to the rise of new challenges, underscoring that “expertise” continues to incorporate the needed skills as new innovations are introduced in LLM-based systems. These developments highlight the indispensable role of human experts in ensuring that LLM-based systems are safely and effectively integrated into clinical workflows.

Olawade et al [94] present a narrative review of HITL AI in health care and discuss associated implementation challenges and future research directions. Along with HITL, further research is needed on human-computer interaction involving AI tools. These tools must be developed in a manner that facilitates the collaboration between humans and AI to build trust, ensure usability, and promote adoption [95]. Additionally, an important research area is understanding the economic implications of deploying these tools in health care [96].

The multifaceted pyramid of data, evidence generation, and risk associated with LLM-based systems illustrates the

relationships between the different faces and associated levels of evidence generation. The information present in the collected data is fundamental for high-quality evidence generation. The progression among the different levels of the data face of the pyramid can be interpreted as higher levels in the data hierarchy indicate higher information content, which, subsequently, increases the risk associated with the AI system deployment, while simultaneously increasing the quality of evidence associated with the health care recommendations provided to patients. How to formalize these relationships and their interaction with the CREATE framework remains the topic of a different paper.

While the science behind AI is relatively new, issues to resolve remain and must be satisfactorily addressed to ensure the public’s acceptability of the technology. Recently enacted legislation at the state level to provide guardrails for the use of AI is a step in the right direction, although additional topics remain when considering targeting underserved populations. Multidisciplinary perspectives represented by linguistics, computer science, statistical sciences, and medicine are needed to optimize the health care application of LLMs.

---

## Acknowledgments

The authors acknowledge the assistance of Mr. Kenneth Bossert, Former Administrator, Drug Abuse, Research, and Treatment Center (DART) Opioid Treatment Program (OTP), Buffalo, New York. They also acknowledge Dr Ashly Jordan, Office of Addiction Services and Supports (OASAS), and Mr Andrew Heck, MPH, OASAS, for helpful discussions.

---

## Funding

This work was supported by Patient-Centered Outcomes Research Institute (PCORI) Award ME-2024C1-37584 [MM] [97], and partially supported by the Troup Fund of the Kaleida Health Foundation [MM]. The statements in this work are solely the responsibility of the authors and do not necessarily represent the views of PCORI, its Board of Governors or Methodology Committee. Study funders were not involved in data collection, analysis, or manuscript preparation.

---

## Authors’ Contributions

Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing: MM

Investigation, Validation, Visualization, Writing – original draft, Writing – review & editing: RM

Investigation, Writing – original draft, Writing – review & editing: JG, YT, AD

Writing – review & editing: LB

Conceptualization, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing: AHT

---

## Conflicts of Interest

AHT has received research support from Gilead Sciences, Novo Nordisk, AstraZeneca, and Salix paid to his institution. AHT has received consultant and honoraria support from AbbVie, Gilead, Novo Nordisk, and Madrigal. AHT is also President of Empath Medical Innovations. MM is Vice President of Empath Medical Innovations.

---

## Multimedia Appendix 1

A brief history of natural language processing and evolution timeline of the field of natural language processing.  
[\[DOCX File \(Microsoft Word File\), 619 KB-Multimedia Appendix 1\]](#)

---

## Multimedia Appendix 2

Details on large language models and brief literature.  
[\[DOCX File \(Microsoft Word File\), 30 KB-Multimedia Appendix 2\]](#)

---

## Multimedia Appendix 3

Evaluation of CREATE framework.

[\[DOCX File \(Microsoft Word File\), 24 KB-Multimedia Appendix 3\]](#)

## References

1. Angus DC, Khera R, Lieu T, et al. AI, health, and health care today and tomorrow: the JAMA summit report on artificial intelligence. *JAMA*. Nov 11, 2025;334(18):1650-1664. [doi: [10.1001/jama.2025.18490](https://doi.org/10.1001/jama.2025.18490)] [Medline: [41082366](https://pubmed.ncbi.nlm.nih.gov/41082366/)]
2. Papalamprakopoulou Z, Roussos S, Ntagianta E, et al. Considerations for equitable distribution of digital healthcare for people who use drugs. *BMC Health Serv Res*. Apr 10, 2025;25(1):531. [doi: [10.1186/s12913-025-12619-7](https://doi.org/10.1186/s12913-025-12619-7)] [Medline: [40211324](https://pubmed.ncbi.nlm.nih.gov/40211324/)]
3. Talal AH, George SJ, Talal LA, et al. Engaging people who use drugs in clinical research: integrating facilitated telemedicine for HCV into substance use treatment. *Res Involv Engagem*. Aug 2, 2023;9(1):63. [doi: [10.1186/s40900-023-00474-x](https://doi.org/10.1186/s40900-023-00474-x)] [Medline: [37533127](https://pubmed.ncbi.nlm.nih.gov/37533127/)]
4. Talal AH, Jaanimägi U, Davis K, et al. Facilitating engagement of persons with opioid use disorder in treatment for hepatitis C virus infection via telemedicine: stories of onsite case managers. *J Subst Abuse Treat*. Aug 2021;127:108421. [doi: [10.1016/j.jsat.2021.108421](https://doi.org/10.1016/j.jsat.2021.108421)] [Medline: [34134875](https://pubmed.ncbi.nlm.nih.gov/34134875/)]
5. Dole VP, Nyswander ME, Kreek MJ. Narcotic blockade. *Arch Intern Med*. Oct 1966;118(4):304-309. [Medline: [4162686](https://pubmed.ncbi.nlm.nih.gov/4162686/)]
6. Califf RM, Robb MA, Bindman AB, et al. Transforming evidence generation to support health and health care decisions. *N Engl J Med*. Dec 15, 2016;375(24):2395-2400. [doi: [10.1056/NEJMs1610128](https://doi.org/10.1056/NEJMs1610128)] [Medline: [27974039](https://pubmed.ncbi.nlm.nih.gov/27974039/)]
7. Jackson GP, Shortliffe EH. Understanding the evidence for artificial intelligence in healthcare. *BMJ Qual Saf*. Jun 19, 2025;34(7):421-424. [doi: [10.1136/bmjqs-2025-018559](https://doi.org/10.1136/bmjqs-2025-018559)] [Medline: [40246317](https://pubmed.ncbi.nlm.nih.gov/40246317/)]
8. Rogers EM. *Diffusion of Innovations*. 5th ed. Free Press; 2003. ISBN: 9780743258234
9. Generative artificial intelligence: overview, issues, and considerations for congress. United States Congress. 2025. URL: [https://www.congress.gov/crs\\_external\\_products/IF/HTML/IF12426.html](https://www.congress.gov/crs_external_products/IF/HTML/IF12426.html) [Accessed 2026-09-14]
10. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, editors. Presented at: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; Dec 4-9, 2017; Long Beach California USA. 2017. [doi: [10.5555/3295222.3295349](https://doi.org/10.5555/3295222.3295349)]
11. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. *Front Med (Lausanne)*. 2024;11:1477898. [doi: [10.3389/fmed.2024.1477898](https://doi.org/10.3389/fmed.2024.1477898)] [Medline: [39534227](https://pubmed.ncbi.nlm.nih.gov/39534227/)]
12. Zhang Q, Huang Z, Huang Y, et al. Generative AI in medical education: feasibility and educational value of LLM-generated clinical cases with MCQs. *BMC Med Educ*. Oct 27, 2025;25(1):1502. [doi: [10.1186/s12909-025-08085-8](https://doi.org/10.1186/s12909-025-08085-8)] [Medline: [41146115](https://pubmed.ncbi.nlm.nih.gov/41146115/)]
13. Wilhelm TI, Roos J, Kaczmarczyk R. Large language models for therapy recommendations across 3 clinical specialties: comparative study. *J Med Internet Res*. Oct 30, 2023;25:e49324. [doi: [10.2196/49324](https://doi.org/10.2196/49324)] [Medline: [37902826](https://pubmed.ncbi.nlm.nih.gov/37902826/)]
14. Usuyama N, Wong C, Zhang S, Naumann T, Poon H. Biomedical natural language processing in the era of large language models. *Annu Rev Biomed Data Sci*. Aug 2025;8(1):471-490. [doi: [10.1146/annurev-biodatasci-103123-095406](https://doi.org/10.1146/annurev-biodatasci-103123-095406)] [Medline: [40245359](https://pubmed.ncbi.nlm.nih.gov/40245359/)]
15. Zhou J, Chen AZ, Shah D, Schwab-Reese LM, DE Choudhury M. A risk taxonomy and reflection tool for large language model adoption in public health. *Proc ACM Hum Comput Interact*. Nov 2025;9(7):CSCW363. [doi: [10.1145/3757544](https://doi.org/10.1145/3757544)] [Medline: [41716439](https://pubmed.ncbi.nlm.nih.gov/41716439/)]
16. Sociotechnical systems. Taylor & Francis Knowledge Centers. URL: [https://taylorandfrancis.com/knowledge/Engineering\\_and\\_technology/Computer\\_science/Sociotechnical\\_systems](https://taylorandfrancis.com/knowledge/Engineering_and_technology/Computer_science/Sociotechnical_systems) [Accessed 2026-09-14]
17. Sittig DF, Singh H. A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. *Qual Saf Health Care*. Oct 2010;19 Suppl 3(Suppl 3):i68-74. [doi: [10.1136/qshc.2010.042085](https://doi.org/10.1136/qshc.2010.042085)] [Medline: [20959322](https://pubmed.ncbi.nlm.nih.gov/20959322/)]
18. Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S, Vertesi J. Fairness and abstraction in sociotechnical systems. Presented at: FAT\* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency; Jan 29-31, 2019:59-68; Atlanta GA USA. Jan 29, 2019.[doi: [10.1145/3287560.3287598](https://doi.org/10.1145/3287560.3287598)]
19. Artificial intelligence. Epic systems c. URL: <https://www.epic.com/software/ai> [Accessed 2026-09-14]
20. Kim SI, Park J, Kim T, et al. Enhancing patient participation in emergency department through patient-friendly clinical notes generated by large language models. *Sci Rep*. Dec 5, 2025;16(1):1409. [doi: [10.1038/s41598-025-31113-y](https://doi.org/10.1038/s41598-025-31113-y)] [Medline: [41350380](https://pubmed.ncbi.nlm.nih.gov/41350380/)]
21. Liu X, Liu H, Yang G, et al. A generalist medical language model for disease diagnosis assistance. *Nat Med*. Mar 2025;31(3):932-942. [doi: [10.1038/s41591-024-03416-6](https://doi.org/10.1038/s41591-024-03416-6)] [Medline: [39779927](https://pubmed.ncbi.nlm.nih.gov/39779927/)]

22. Panagoulas DP, Virvou M, Tsihrintzis GA. Evaluating LLM-generated multimodal diagnosis from medical images and symptom analysis. arXiv. Preprint posted online on Jan 28, 2024. [doi: [10.48550/arXiv.2402.01730](https://doi.org/10.48550/arXiv.2402.01730)]
23. Panagoulas DP, Palamidas FA, Virvou M, Tsihrintzis GA. Evaluating the potential of llms and chatgpt on medical diagnosis and treatment. Presented at: 2023 14th International Conference on Information, Intelligence, Systems & Applications (IISA); Jul 10-12, 2023; Volos, Greece. [doi: [10.1109/IISA59645.2023.10345968](https://doi.org/10.1109/IISA59645.2023.10345968)]
24. Levra AG, Gatti M, Mene R, et al. A large language model-based clinical decision support system for syncope recognition in the emergency department: a framework for clinical workflow integration. *Eur J Intern Med*. Jan 2025;131:113-120. [doi: [10.1016/j.ejim.2024.09.017](https://doi.org/10.1016/j.ejim.2024.09.017)] [Medline: [39341748](https://pubmed.ncbi.nlm.nih.gov/39341748/)]
25. Tools and resources to support engagement in research. Patient-Centered Outcomes Research Institute (PCORI). URL: <http://pcori.org/engagement-research/tools-and-resources-support-engagement-research> [Accessed 2026-09-24]
26. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. *JAMA*. Mar 14, 2023;329(10):842-844. [doi: [10.1001/jama.2023.1044](https://doi.org/10.1001/jama.2023.1044)] [Medline: [36735264](https://pubmed.ncbi.nlm.nih.gov/36735264/)]
27. Mika AP, Martin JR, Engstrom SM, Polkowski GG, Wilson JM. Assessing ChatGPT responses to common patient questions regarding total hip arthroplasty. *J Bone Joint Surg Am*. Oct 4, 2023;105(19):1519-1526. [doi: [10.2106/JBJS.23.00209](https://doi.org/10.2106/JBJS.23.00209)] [Medline: [37459402](https://pubmed.ncbi.nlm.nih.gov/37459402/)]
28. Grewal H, Dhillon G, Monga V, et al. Radiology gets chatty: the ChatGPT saga unfolds. *Cureus*. Jun 2023;15(6):e40135. [doi: [10.7759/cureus.40135](https://doi.org/10.7759/cureus.40135)] [Medline: [37425598](https://pubmed.ncbi.nlm.nih.gov/37425598/)]
29. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: the potentials and pitfalls: a narrative review. *Ann Intern Med*. Feb 2024;177(2):210-220. [doi: [10.7326/M23-2772](https://doi.org/10.7326/M23-2772)] [Medline: [38285984](https://pubmed.ncbi.nlm.nih.gov/38285984/)]
30. Liu S, McCoy AB, Wright AP, et al. Leveraging large language models for generating responses to patient messages. *medRxiv*. Jul 16, 2023:2023.07.14.23292669. [doi: [10.1101/2023.07.14.23292669](https://doi.org/10.1101/2023.07.14.23292669)] [Medline: [37503263](https://pubmed.ncbi.nlm.nih.gov/37503263/)]
31. Singh S, Djalilian A, Ali MJ. ChatGPT and ophthalmology: exploring its potential with discharge summaries and operative notes. *Semin Ophthalmol*. Jul 2023;38(5):503-507. [doi: [10.1080/08820538.2023.2209166](https://doi.org/10.1080/08820538.2023.2209166)] [Medline: [37133418](https://pubmed.ncbi.nlm.nih.gov/37133418/)]
32. Sorin V, Klang E, Sklair-Levy M, et al. Large language model (ChatGPT) as a support tool for breast tumor board. *NPJ Breast Cancer*. May 30, 2023;9(1):44. [doi: [10.1038/s41523-023-00557-8](https://doi.org/10.1038/s41523-023-00557-8)] [Medline: [37253791](https://pubmed.ncbi.nlm.nih.gov/37253791/)]
33. Daneshvar N, Pandita D, Erickson S, Snyder Sulmasy L, DeCamp M, Committee A. Artificial Intelligence in the provision of health care: an American College of Physicians policy position paper. *Ann Intern Med*. Jul 2024;177(7):964-967. [doi: [10.7326/M24-0146](https://doi.org/10.7326/M24-0146)] [Medline: [38830215](https://pubmed.ncbi.nlm.nih.gov/38830215/)]
34. Sidorov G, Ahmad M, Basile P, Waqas M, Orji R, Batyrshin I. Monitoring opioid-related social media chatter using natural language processing and large language models: temporal analysis. *JMIR Infodemiology*. Nov 4, 2025;5:e77279. [doi: [10.2196/77279](https://doi.org/10.2196/77279)] [Medline: [41187282](https://pubmed.ncbi.nlm.nih.gov/41187282/)]
35. Sousa-Pinto B, Marques-Cruz M, Neumann I, et al. Guidelines international network: principles for use of artificial intelligence in the health guideline enterprise. *Ann Intern Med*. Mar 2025;178(3):408-415. [doi: [10.7326/ANNALS-24-02338](https://doi.org/10.7326/ANNALS-24-02338)] [Medline: [39869912](https://pubmed.ncbi.nlm.nih.gov/39869912/)]
36. Cao C, Sang J, Arora R, et al. Development of prompt templates for large language model-driven screening in systematic reviews. *Ann Intern Med*. Mar 2025;178(3):389-401. [doi: [10.7326/ANNALS-24-02189](https://doi.org/10.7326/ANNALS-24-02189)] [Medline: [39993313](https://pubmed.ncbi.nlm.nih.gov/39993313/)]
37. McNemar E. How does artificial intelligence compare to augmented intelligence. TechTarget Health IT Analytics. 2021. URL: <https://www.techtarget.com/healthtechanalytics/news/366591056/How-Does-Artificial-Intelligence-Compare-to-Augmented-Intelligence> [Accessed 2026-09-14]
38. Augmented intelligence in medicine. American Medical Association. URL: <https://www.ama-assn.org/practice-management/digital-health/augmented-intelligence-medicine> [Accessed 2026-09-14]
39. Zhang X, Yu J, Yan P, Jiang L, Shen X, Cheng M, et al. Human-in-the-loop interactive report generation for chronic disease adherence. arXiv. Preprint posted online on Jan 10, 2026. [doi: [10.48550/arXiv.2601.06364](https://doi.org/10.48550/arXiv.2601.06364)]
40. Kieseberg P, Weippl E, Holzinger A. Trust for the “doctor in the loop”. *ERCIM News*. Jan 13, 2016. URL: <https://ercim-news.ercim.eu/en104/special/trust-for-the-doctor-in-the-loop> [Accessed 2026-09-24]
41. Caragliano AN, Ruffini F, Greco C, et al. Doctor-in-the-Loop: an explainable, multi-view deep learning framework for predicting pathological response in non-small cell lung cancer. *Image Vis Comput*. Sep 2025;161:105630. [doi: [10.1016/j.imavis.2025.105630](https://doi.org/10.1016/j.imavis.2025.105630)]
42. Liao QV, Wortman Vaughan J. AI transparency in the age of LLMs: a human-centered research roadmap. *Harvard Data Science Review*. 2024;(Special Issue 5). [doi: [10.1162/99608f92.8036d03b](https://doi.org/10.1162/99608f92.8036d03b)]
43. Bouzoubaa L, Aghakhani E, Rezapour R, editors. Words Matter: Reducing Stigma in Online Conversations about Substance Use with Large Language Models. Association for Computational Linguistics; 2024. [doi: [10.18653/v1/2024.emnlp-main.516](https://doi.org/10.18653/v1/2024.emnlp-main.516)]

44. Lo Bianco G, Robinson CL, D'Angelo FP, et al. Effectiveness of generative artificial intelligence-driven responses to patient concerns in long-term opioid therapy: cross-model assessment. *Biomedicines*. Mar 5, 2025;13(3):636. [doi: [10.3390/biomedicines13030636](https://doi.org/10.3390/biomedicines13030636)] [Medline: [40149612](https://pubmed.ncbi.nlm.nih.gov/40149612/)]
45. Nateghi Haredasht F, Lopez I, Tate S, et al. Predicting treatment retention in medication for opioid use disorder: a machine learning approach using NLP and LLM-derived clinical features. *J Am Med Inform Assoc*. Dec 1, 2025;32(12):1865-1876. [doi: [10.1093/jamia/ocaf157](https://doi.org/10.1093/jamia/ocaf157)] [Medline: [40977375](https://pubmed.ncbi.nlm.nih.gov/40977375/)]
46. Voigt H, Sugamiya Y, Lawonn K, Zarriß S, Takanishi A. LLM-powered virtual patient agents for interactive clinical skills training with automated feedback. *arXiv*. Preprint posted online on Aug 19, 2025. [doi: [10.48550/arXiv.2508.13943](https://doi.org/10.48550/arXiv.2508.13943)]
47. Agweyu A, Mwaniki P, Menon V, et al. Generative AI-enabled clinical decision support system in primary care: a pragmatic, cluster-randomized trial. *Nat Med*. Aug 2026;32(8):3032-3039. [doi: [10.1038/s41591-026-04503-6](https://doi.org/10.1038/s41591-026-04503-6)] [Medline: [42362867](https://pubmed.ncbi.nlm.nih.gov/42362867/)]
48. Ong JCL, Jin L, Elangovan K, et al. Large language model as clinical decision support system augments medication safety in 16 clinical specialties. *Cell Rep Med*. Oct 21, 2025;6(10):102323. [doi: [10.1016/j.xcrm.2025.102323](https://doi.org/10.1016/j.xcrm.2025.102323)] [Medline: [40997804](https://pubmed.ncbi.nlm.nih.gov/40997804/)]
49. Sethi R, Caskey J, Gao Y, et al. Detecting stigmatizing language in clinical notes with large language models for addiction care. *medRxiv*. Aug 12, 2025:2025.08.08.25333315. [doi: [10.1101/2025.08.08.25333315](https://doi.org/10.1101/2025.08.08.25333315)] [Medline: [40832420](https://pubmed.ncbi.nlm.nih.gov/40832420/)]
50. Kaiyrbekov K, Dobbins NJ, Mooney SD. Automated survey collection with LLM-based conversational agents. *JAMIA Open*. Oct 2025;8(5):ooaf103. [doi: [10.1093/jamiaopen/ooaf103](https://doi.org/10.1093/jamiaopen/ooaf103)] [Medline: [41180893](https://pubmed.ncbi.nlm.nih.gov/41180893/)]
51. Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat Med*. Apr 2024;30(4):1134-1142. [doi: [10.1038/s41591-024-02855-5](https://doi.org/10.1038/s41591-024-02855-5)] [Medline: [38413730](https://pubmed.ncbi.nlm.nih.gov/38413730/)]
52. Song JW, Jang JY, Kim H, Ko YG, You SC. Transforming free-text coronary angiography reports into structured, analyzable data using large language models. *Sci Rep*. Jan 3, 2026;16(1):2360. [doi: [10.1038/s41598-025-32150-3](https://doi.org/10.1038/s41598-025-32150-3)] [Medline: [41484419](https://pubmed.ncbi.nlm.nih.gov/41484419/)]
53. Mansoor I, Mohammed AM, Blythe S. BPI25-012: developing an artificial intelligence tool for personalized breast cancer treatment plans based on the NCCN guidelines. *J Natl Compr Canc Netw*. 2025;23(3.5):250215698. [doi: [10.6004/jnccn.2024.7135](https://doi.org/10.6004/jnccn.2024.7135)]
54. Wang Q, Wang Z, Li M, et al. A feasibility study of automating radiotherapy planning with large language model agents. *Phys Med Biol*. Mar 21, 2025;70(7). [doi: [10.1088/1361-6560/adbf11](https://doi.org/10.1088/1361-6560/adbf11)] [Medline: [40073507](https://pubmed.ncbi.nlm.nih.gov/40073507/)]
55. He W, Liao J, Pan H, et al. Feasibility of LLM-assisted post-discharge tuberculosis care: a comparative study of medication counseling, patient education, and follow-up planning. *Digit HEALTH*. 2026;12:20552076261472569. [doi: [10.1177/20552076261472569](https://doi.org/10.1177/20552076261472569)] [Medline: [42571185](https://pubmed.ncbi.nlm.nih.gov/42571185/)]
56. Wang H, Fu W, Tang Y, Chen Z, Huang Y, Piao J, et al. A survey on responsible llms: inherent risk, malicious use, and mitigation strategy. *arXiv*. Preprint posted online on Jan 16, 2025. [doi: [10.48550/arXiv.2501.09431](https://doi.org/10.48550/arXiv.2501.09431)]
57. Jiao J, Afroogh S, Xu Y, Phillips C. Navigating LLM ethics: advancements, challenges, and future directions. *AI Ethics*. Dec 2025;5(6):5795-5819. [doi: [10.1007/s43681-025-00814-5](https://doi.org/10.1007/s43681-025-00814-5)]
58. Bender EM, Hanna A. *The AI Con: How to Fight Big Tech's Hype and Create the Future We Want*. HarperCollins; 2025. ISBN: 9780063418554
59. Amugongo LM, Mascheroni P, Brooks S, Doering S, Seidel J. Retrieval augmented generation for large language models in healthcare: a systematic review. *PLOS Digit Health*. Jun 2025;4(6):e0000877. [doi: [10.1371/journal.pdig.0000877](https://doi.org/10.1371/journal.pdig.0000877)] [Medline: [40498738](https://pubmed.ncbi.nlm.nih.gov/40498738/)]
60. Chen S, Gao M, Sasse K, et al. When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior. *NPJ Digit Med*. Oct 17, 2025;8(1):605. [doi: [10.1038/s41746-025-02008-z](https://doi.org/10.1038/s41746-025-02008-z)] [Medline: [41107408](https://pubmed.ncbi.nlm.nih.gov/41107408/)]
61. Rebedea T, Dinu R, Sreedhar MN, Parisien C, Cohen J, editor. *NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails*. Association for Computational Linguistics; 2023. [doi: [10.18653/v1/2023.emnlp-demo.40](https://doi.org/10.18653/v1/2023.emnlp-demo.40)]
62. Zhao H, Chen H, Yang F, et al. Explainability for large language models: a survey. *ACM Trans Intell Syst Technol*. Apr 30, 2024;15(2):1-38. [doi: [10.1145/3639372](https://doi.org/10.1145/3639372)]
63. Yuan J, Li H, Ding X, et al. Understanding and mitigating numerical sources of nondeterminism in LLM inference. Presented at: *Advances in Neural Information Processing Systems 38*; Dec 2-7, 2025; San Diego, California, USA and Mexico City, Mexico. 2025.[doi: [10.52202/085713-5653](https://doi.org/10.52202/085713-5653)]
64. Shorinwa O, Mei Z, Lidard J, Ren AZ, Majumdar A. A survey on uncertainty quantification of large language models: taxonomy, open research challenges, and future directions. *ACM Comput Surv*. Feb 28, 2026;58(3):1-38. [doi: [10.1145/3744238](https://doi.org/10.1145/3744238)]

65. Atf Z, Safavi-Naini SAA, Lewis PR, Mahjoubfar A, Naderi N, Savage TR, et al. The challenge of uncertainty quantification of large language models in medicine. arXiv. Preprint posted online on Apr 7, 2025. [doi: [10.48550/arXiv.2504.05278](https://doi.org/10.48550/arXiv.2504.05278)]
66. Savage T, Wang J, Gallo R, et al. Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment. *J Am Med Inform Assoc*. Jan 1, 2025;32(1):139-149. [doi: [10.1093/jamia/ocae254](https://doi.org/10.1093/jamia/ocae254)] [Medline: [39396184](https://pubmed.ncbi.nlm.nih.gov/39396184/)]
67. AMIA artificial intelligence evaluation showcase. American Medical Informatics Association. URL: <https://web.archive.org/web/20250908082941/https://amia.org/education-events/amia-artificial-intelligence-evaluation-showcase> [Accessed 2026-09-14]
68. Minaee S, Mikolov T, Nikzad N, Chenaghlu M, Socher R, Amatriain X, et al. Large language models: a survey. arXiv. Preprint posted online on Feb 9, 2024. [doi: [10.48550/arXiv.2402.06196](https://doi.org/10.48550/arXiv.2402.06196)]
69. Shyr C, Ren B, Hsu CY, et al. A statistical framework for evaluating the repeatability and reproducibility of large language models. *medRxiv*. Nov 4, 2025:2025.08.06.25333170. [doi: [10.1101/2025.08.06.25333170](https://doi.org/10.1101/2025.08.06.25333170)] [Medline: [41282788](https://pubmed.ncbi.nlm.nih.gov/41282788/)]
70. Usability. Computer Security Resource Center, National Institute of Standards and Technology. URL: <https://csrc.nist.gov/glossary/term/usability> [Accessed 2026-09-14]
71. Dehghani Mahmoodabadi A, Dehghan H, Kimiafar K, Karami M, Mousavi Baigi SF, Farsi Mehrabadi M. Usability evaluation methods for hospital information systems: a systematic review. *BMC Health Serv Res*. Nov 21, 2025;25(1):1540. [doi: [10.1186/s12913-025-13745-y](https://doi.org/10.1186/s12913-025-13745-y)] [Medline: [41272660](https://pubmed.ncbi.nlm.nih.gov/41272660/)]
72. Markatou M, Kennedy O, Brachmann M, Mukhopadhyay R, Dharia A, Talal AH. Social determinants of health derived from people with opioid use disorder: Improving data collection, integration and use with cross-domain collaboration and reproducible, data-centric, notebook-style workflows. *Front Med (Lausanne)*. 2023;10:1076794. [doi: [10.3389/fmed.2023.1076794](https://doi.org/10.3389/fmed.2023.1076794)] [Medline: [36936205](https://pubmed.ncbi.nlm.nih.gov/36936205/)]
73. Keyes T, Callahan A, Pandya AS, et al. Why and how to monitor deployed AI systems in health care. *NEJM Catalyst*. May 20, 2026;7(6). [doi: [10.1056/CAT.25.0372](https://doi.org/10.1056/CAT.25.0372)]
74. Hussein R, Zink A, Ramadan B, et al. Advancing healthcare AI governance through a comprehensive maturity model based on systematic review. *NPJ Digit Med*. Feb 11, 2026;9(1):236. [doi: [10.1038/s41746-026-02418-7](https://doi.org/10.1038/s41746-026-02418-7)] [Medline: [41673321](https://pubmed.ncbi.nlm.nih.gov/41673321/)]
75. El Arab RA, Mustafa MH, Almagharbeh WT, et al. Beyond model development in healthcare AI: post-development robustness, post-deployment monitoring, and lifecycle governance—a scoping review of reviews. *Healthcare (Basel)*. May 25, 2026;14(11):1459. [doi: [10.3390/healthcare14111459](https://doi.org/10.3390/healthcare14111459)] [Medline: [42278712](https://pubmed.ncbi.nlm.nih.gov/42278712/)]
76. O'Donnell B, Gupta V. Continuous Quality Improvement. *StatPearls*; 2026. URL: <https://www.ncbi.nlm.nih.gov/books/NBK559239/> [Accessed 2026-09-24]
77. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. World Health Organization. 2024. URL: <https://www.who.int/publications/i/item/9789240084759> [Accessed 2026-09-14]
78. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending regulations (EC) no 300/2008, (EU) no 167/2013, (EU) no 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (artificial intelligence act). *EUR-Lex*. URL: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj> [Accessed 2026-09-14]
79. EU AI act: first regulation on artificial intelligence. European Parliament. URL: <https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence> [Accessed 2026-09-14]
80. Madiaga T. Artificial intelligence act. European Parliamentary Research Service; 2024. URL: [https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS\\_BRI\(2021\)698792\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf) [Accessed 2026-09-14]
81. California senate bill 53 (2025-2026 regular session): artificial intelligence models: large developers. *Legi Scan*. 2025. URL: <https://legiscan.com/CA/bill/SB53/2025> [Accessed 2026-09-14]
82. Center for drug evaluation and research artificial intelligence for drug development. US Food and Drug Administration. 2025. URL: <https://www.fda.gov/about-fda/center-drug-evaluation-and-research-cder/artificial-intelligence-drug-development> [Accessed 2026-09-14]
83. Ensuring a national policy framework for artificial intelligence. *Fed Regist*. 2025. URL: <https://www.federalregister.gov/documents/2025/12/16/2025-23092/ensuring-a-national-policy-framework-for-artificial-intelligence> [Accessed 2026-09-14]
84. Governor Hochul signs nation-leading legislation to require AI frameworks for AI frontier models. New York State Governor's Office. 2025. URL: <https://www.governor.ny.gov/news/governor-hochul-signs-nation-leading-legislation-require-ai-frameworks-ai-frontier-models> [Accessed 2026-09-14]

85. Kolbinger FR, Veldhuizen GP, Zhu J, Truhn D, Kather JN. Reporting guidelines in medical artificial intelligence: a systematic review and meta-analysis. *Commun Med (Lond)*. Apr 11, 2024;4(1):71. [doi: [10.1038/s43856-024-00492-0](https://doi.org/10.1038/s43856-024-00492-0)] [Medline: [38605106](https://pubmed.ncbi.nlm.nih.gov/38605106/)]
86. Wang Y, Li N, Chen L, et al. Guidelines, consensus statements, and standards for the use of artificial intelligence in medicine: systematic review. *J Med Internet Res*. Nov 22, 2023;25:e46089. [doi: [10.2196/46089](https://doi.org/10.2196/46089)] [Medline: [37991819](https://pubmed.ncbi.nlm.nih.gov/37991819/)]
87. Sim I, Cassel C. The ethics of relational AI — expanding and implementing the Belmont principles. *N Engl J Med*. Jul 18, 2024;391(3):193-196. [doi: [10.1056/NEJMp2314771](https://doi.org/10.1056/NEJMp2314771)] [Medline: [39007542](https://pubmed.ncbi.nlm.nih.gov/39007542/)]
88. Augmented intelligence development, deployment, and use in health care. American Medical Association. URL: <https://www.ama-assn.org/system/files/ama-ai-principles.pdf> [Accessed 2026-09-14]
89. Blueprint for trustworthy AI implementation guidance and assurance for healthcare. Coalition for Health AI. 2023. URL: [https://assets.ctfassets.net/7s4afyr9pmov/4AXIWGllcrjWDaW2ueTaRS/f98e5cb2528187635895cce6ba5ec309/Blueprint\\_for\\_Trustworthy\\_AI.pdf](https://assets.ctfassets.net/7s4afyr9pmov/4AXIWGllcrjWDaW2ueTaRS/f98e5cb2528187635895cce6ba5ec309/Blueprint_for_Trustworthy_AI.pdf) [Accessed 2026-09-14]
90. Xu F, Uszkoreit H, Du Y, Fan W, Zhao D, Zhu J. Explainable AI: a brief survey on history, research areas, approaches and challenges. Presented at: Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019; Oct 9-14, 2019:563-574; Dunhuang, China. [doi: [10.1007/978-3-030-32236-6\\_51](https://doi.org/10.1007/978-3-030-32236-6_51)]
91. Young CC, Enichen E, Rao A, Succi MD. Racial, ethnic, and sex bias in large language model opioid recommendations for pain management. *Pain*. Mar 1, 2025;166(3):511-517. [doi: [10.1097/j.pain.0000000000003388](https://doi.org/10.1097/j.pain.0000000000003388)] [Medline: [39283333](https://pubmed.ncbi.nlm.nih.gov/39283333/)]
92. Glasgow RE, Harden SM, Gaglio B, et al. RE-AIM planning and evaluation framework: adapting to new science and practice with a 20-year review. *Front Public Health*. 2019;7:64. [doi: [10.3389/fpubh.2019.00064](https://doi.org/10.3389/fpubh.2019.00064)] [Medline: [30984733](https://pubmed.ncbi.nlm.nih.gov/30984733/)]
93. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated consolidated framework for implementation research based on user feedback. *Implementation Sci*. 2022;17(1):75. [doi: [10.1186/s13012-022-01245-0](https://doi.org/10.1186/s13012-022-01245-0)]
94. Olawade DB, Plabon SB, Ojo A, Ogunbona MA, Makanjuola BD, Olasilola OR. Human in the loop artificial intelligence in healthcare: applications, outcomes, and implementation challenges. *Int J Med Inform*. Jun 15, 2026;213:106362. [doi: [10.1016/j.ijmedinf.2026.106362](https://doi.org/10.1016/j.ijmedinf.2026.106362)] [Medline: [41740273](https://pubmed.ncbi.nlm.nih.gov/41740273/)]
95. Lazaros K, Vrahatis AG, Kotsiantis S. Human-in-the-loop artificial intelligence: a systematic review of concepts, methods, and applications. *Entropy (Basel)*. Mar 26, 2026;28(4):377. [doi: [10.3390/e28040377](https://doi.org/10.3390/e28040377)] [Medline: [42072503](https://pubmed.ncbi.nlm.nih.gov/42072503/)]
96. López-Úbeda P, Martín-Noguerol T, Luna A. Environmental and economic costs behind LLMs. *Int J CARS*. 2026;21(3):661-663. [doi: [10.1007/s11548-026-03568-5](https://doi.org/10.1007/s11548-026-03568-5)]
97. Developing methods for clustering longitudinal mixed-type data for comparative effectiveness research. University at Buffalo. URL: <https://ubwp.buffalo.edu/clustllm4cer/> [Accessed 2026-09-14]

## Abbreviations

**AMIA:** American Medical Informatics Association  
**CFIR:** Consolidated Framework for Implementation Research  
**CREATE:** Culture, Respect, Education, Advancement, Trust, and Expertise  
**DOI:** diffusion of innovations  
**EHR:** electronic health record  
**EU:** European Union  
**GenAI:** generative AI  
**HIPAA:** Health Insurance Portability and Accountability Act  
**HIT:** health information technology  
**HITL:** human-in-the-loop  
**LLM:** large language model  
**NLP:** natural language processing  
**OASAS:** Office of Addiction Services and Supports  
**OTP:** Opioid Treatment Program  
**ODU:** opioid use disorder  
**PCORI:** Patient-Centered Outcomes Research Institute  
**RE-AIM:** Reach, Effectiveness, Adoption, Implementation, and Maintenance  
**STS:** socio-technical system  
**WHO:** World Health Organization

*Edited by Ivan Steenstra; peer-reviewed by Maria Chatzimina, Mohit Singhal; submitted 03.Apr.2026; final revised version received 02.Sep.2026; accepted 02.Sep.2026; published 07.Oct.2026*

*Please cite as:*

Markatou M, Mukhopadhyay R, Good J, Tan Y, Dharia A, Brown LS, Talal AH  
*From Innovation to Impact: The CREATE Framework as a Blueprint for Large Language Model Adoption in Opioid Treatment Programs*  
*J Med Internet Res* 2026;28:e97131  
URL: <https://www.jmir.org/2026/1/e97131>  
doi: [10.2196/97131](https://doi.org/10.2196/97131)

© Marianthi Markatou, Raktim Mukhopadhyay, Jeff Good, Yihao Tan, Arpan Dharia, Lawrence S Brown, Andrew H Talal. Originally published in the Journal of Medical Internet Research (<https://www.jmir.org>), 07.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on <https://www.jmir.org/>, as well as this copyright and license information must be included.