AI Prompting Best Practices in Evidence-Based Human Medicine: A Comprehensive Guide for Clinical Integration

Dauben, Otte, Mullen

Executive Summary

This report provides a comprehensive examination of AI prompting best practices within the context of evidence-based human medicine. It underscores the transformative potential of Large Language Models (LLMs) in healthcare, highlighting their burgeoning applications across clinical decision support, medical education, diagnostics, and patient care. Despite these advancements, the integration of LLMs presents significant challenges, particularly concerning factual accuracy, bias, privacy, and interpretability. Effective prompt engineering emerges as a critical mechanism to guide LLMs towards generating scientifically grounded, clinically relevant, and ethically sound outputs.

The document details foundational principles for high-quality AI-generated medical information, including accuracy, reliability, transparency, and bias mitigation. It explores advanced prompting frameworks such as Chain-of-Thought (CoT), Self-Consistency, and Retrieval-Augmented Generation (RAG), demonstrating how these techniques enhance reasoning, reliability, and evidence grounding. The report emphasizes the necessity of iterative prompt development, robust version control, and meticulous documentation, treating prompts as living protocols requiring formal governance. Furthermore, it discusses the integration of AI outputs into evidence-based frameworks, stressing the indispensable role of human validation and critical appraisal using established methodologies like GRADE, AMSTAR 2, and PRISMA. Specific medical use cases illustrate the practical application of these principles, while a dedicated section addresses the complex ethical, legal, and security considerations, including data privacy (GDPR, HIPAA), bias mitigation, accountability, regulatory compliance (EU AI Act), and prompt injection attack prevention. The overarching message is that while LLMs offer unprecedented opportunities, their safe, effective, and responsible integration into healthcare demands a rigorous, multi-faceted approach centered on meticulous prompting, continuous oversight, and adherence to evolving regulatory standards.

1 Introduction: The Evolving Landscape of LLMs in Healthcare

1.1 Overview of LLM Capabilities and Their Burgeoning Applications in Clinical Settings

Large Language Models (LLMs) represent a profound advancement in artificial intelligence, showcasing remarkable capabilities in comprehending and generating human-like text.[1] These sophisticated models, built upon deep learning and natural language processing technologies, are rapidly being adopted across various sectors, with a particularly impactful integration into healthcare. Within the medical domain, LLMs are poised to revolutionize numerous facets of practice, from enhancing clinical decision-making to improving patient care, medical education, and research.[1]

The proficiency of LLMs in processing vast amounts of textual data, extracting meaningful insights, and producing high-quality outputs has opened new avenues for innovation. Current implementations demonstrate their promising utility in clinical decision support, medical education, diagnostics, and direct patient care.[1] Specific applications include the ability to answer complex medical questions, generate comprehensive patient information through summarization or translation, and streamline clinical documentation processes.[2] Beyond these, LLMs can extract critical clinical information from Electronic Health Records (EHRs), optimize administrative workflows, and significantly bolster medical research efforts.[2] In clinical decision support, LLMs offer personalized insights for potential diagnoses, recommend appropriate specialists, and assist in assessing urgent care needs. They also support clinicians in refining diagnoses and decision-making, presenting a promising avenue for enhancing patient outcomes and streamlining healthcare delivery.[3] These models are emerging as powerful alternatives to traditional Clinical Decision Support Systems (CDSS), providing unparalleled assistance through active user interaction and direct interpretation of medical information, extending far beyond simple guideline consultation.[4] In the realm of evidence synthesis, LLMs can notably streamline systematic reviews by generating Boolean search queries, aiding in the development of search strategies, and efficiently identifying and abstracting relevant study information, with reported accuracies ranging from 80% to 96%. They can also facilitate the drafting of standardized sections within review manuscripts.[5]

1.2 The Imperative for Robust Prompting in Evidence-Based Medical Practice

Despite the immense promise and diverse applications of LLMs in healthcare, their integration is not without significant challenges. The high-stakes nature of medical decision-making necessitates meticulous validation and responsible deployment.[1] Concerns regarding data privacy, ethical considerations, factual accuracy, and the potential for bias are paramount and require careful resolution for their responsible integration into healthcare systems.[1]

A fundamental challenge stems from the inherent limitations of LLMs, which are often trained on fixed datasets. These datasets can be outdated, lack the specificity required for nuanced biomedical applications, and may include unreliable or untrustworthy sources.[7] This can lead to a phenomenon known as "hallucination," where LLMs generate outputs that are meaningless or inconsistent with factual information, such as fictitious drug recommendations or citations of non-existent clinical studies. Such fabrications pose a critical risk of misdiagnosis, inappropriate treatment, and incorrect medical management.[7] The opaque nature, often referred to as the "black-box" problem, of many LLMs further exacerbates these concerns. Their limitations in contextual understanding and interpretability make it difficult to ascertain how specific decisions or recommendations are reached, which can undermine trust and hinder clinicians' ability to validate AI-generated advice.[6] This lack of transparency, coupled with the potential for biases embedded in training data, poses a risk of perpetuating disparities and inaccuracies in diagnoses.[6] Furthermore, the availability and ease of use of LLMs introduce the risk of over-reliance, potentially diminishing critical thinking and independent decision-making by healthcare professionals.[5]

The confluence of immense promise and profound risks suggests that LLMs are not inherently beneficial or detrimental; rather, they are powerful tools whose impact is critically determined by their careful and responsible deployment. This situation necessitates a proactive and structured approach to their integration, where effective prompt engineering emerges as the central mechanism for mitigating risks and maximizing benefits. Prompt engineering, defined as the art and science of designing and optimizing inputs to guide generative AI solutions, is crucial for eliciting desired, high-quality, and relevant outputs.[10] It acts as a roadmap for the AI, steering it towards specific outputs.[10] Given that LLMs inherently pose risks due to their internal mechanisms (e.g., training data, black-box nature), an external, human-controlled input mechanism, such as prompting, becomes the primary lever to steer their behavior towards safety, accuracy, and ethical compliance. This elevates prompt engineering from a mere technical skill to a critical safety and governance function in healthcare AI, ensuring the development of safe, reliable LLM systems for health technology applications.[12]

2 Foundational Principles of Effective AI Prompting

2.1 Defining Prompt Engineering in the Medical Context

Prompt engineering is precisely the "art and science" of crafting and refining inputs, known as prompts, to guide AI models, particularly LLMs, toward generating specific, desired responses.[10] This process involves carefully constructing these inputs to provide the model with essential context, clear instructions, and illustrative examples, thereby enabling it to accurately interpret user intent and produce meaningful outputs.[10] Essentially, prompt engineering serves as a "roadmap for the AI," directing its computational processes to achieve a particular output.[10]

In the specialized domain of human medicine, this definition takes on heightened significance. Prompt engineering in this context means meticulously tailoring prompts to elicit information that is not only scientifically grounded but also clinically relevant and ethically sound. This transcends the general language tasks LLMs typically perform, moving into highly specialized biomedical applications that demand precision and accuracy.[7] The effectiveness of an LLM in a medical setting is thus profoundly influenced by the quality and specificity of the prompts it receives.

2.2 Core Attributes of High-Quality AI-Generated Medical Outputs

The definition of "high-quality" AI output in medicine must transcend typical AI performance metrics, such as coherence or fluency, to encompass clinical safety, ethical integrity, and alignment with evidence-based practice. This necessitates a tailored set of quality criteria that directly addresses the unique risks and requirements of healthcare, making prompt engineering in medicine fundamentally different from other domains. The very act of crafting a prompt must implicitly or explicitly embed ethical considerations, making it a gatekeeper for responsible AI deployment.

The critical attributes for high-quality AI-generated medical outputs include:

- **Accuracy and Factual Correctness:** Outputs must be rigorously grounded in factual, current medical information and must entirely avoid "hallucinations" or fabricated content.[7] This is paramount in medicine to prevent misdiagnosis, inappropriate treatment, or patient harm.[8] Studies have shown that LLMs can generate convincing yet incorrect medical information, underscoring the need for stringent accuracy.[13]

- **Reliability and Reproducibility:** Outputs should be consistently verifiable and predictable across repeated queries or similar contexts.[5] The opaque nature of LLM data sources and frequent model updates can inherently hinder reproducibility, making explicit efforts to ensure this attribute crucial.[5]
- **Relevance and Contextual Understanding:** Responses must be precisely aligned with specific patient cases, current clinical guidelines, and the nuanced complexities of individual diseases.[6] General LLMs may lack this specific medical contextual understanding, which must be compensated for through careful prompting.[6]
- **Transparency and Interpretability:** The mechanism by which the AI arrives at its conclusions should be understandable and explainable, especially given the "black-box" nature of many LLMs.[6] This interpretability is vital for building trust among clinicians and enabling them to validate AI recommendations effectively.[9]
- **Bias Mitigation and Fairness:** AI outputs must not perpetuate disparities or inaccuracies that stem from biases present in their training data.[6] Fairness mandates consistent performance across diverse patient groups, irrespective of demographic or clinical attributes.[17]
- **Safety and Ethical Compliance:** Adherence to fundamental principles of patient privacy, data security, informed consent, and responsible AI use is non-negotiable.[6] Outputs must never be harmful, unethical, or violate patient trust.[4]
- **Completeness and Comprehensiveness:** Responses should include all necessary elements to fulfill the prompt's expectations, avoiding partial or incomplete information that could lead to misinterpretation.[13]
- **Actionability:** Outputs should provide insights that can be directly and practically applied in clinical practice for informed decision-making or direct patient care.[19]

The shift from general AI performance metrics to clinically-relevant quality criteria is a critical evolution. In a high-stakes domain like healthcare, merely "desired responses" are insufficient. The existence of specific medical concerns such as hallucinations leading to misdiagnosis, biases perpetuating disparities, and the pervasive need for human validation underscores that the definition of "high-quality" AI output must encompass clinical safety, ethical integrity, and alignment with evidence-based practice. This makes prompt engineering in medicine fundamentally different from other domains. Furthermore, the interdependence of technical prompting and ethical compliance is evident. If a prompt leads to a biased output or a hallucination, it is not simply a technical error; it constitutes an ethical violation with potential patient harm. Conversely, designing prompts to mitigate bias or reduce hallucinations directly contributes to ethical AI. This highlights that prompt engineering is not just a technical skill for optimizing model performance but also a crucial ethical responsibility, serving as a gatekeeper

for responsible AI deployment in medicine.

Criterion	Description	Relevance to Prompting	Ref
Accuracy/Factual Correctness	Outputs are grounded in factual, current medical information, free from "hallucinations" or fabrications.	Prompts must explicitly demand factual grounding, specify reliable sources, and instruct the model to verify information.	7
Reliability/ Reproducibility	Consistent and verifiable outputs across similar queries and contexts.	Prompt design should minimize ambiguity, specify output formats, and integrate version control for prompt iterations.	5
Relevance/ Contextual Understanding	Outputs are tailored to specific clinical scenarios, patient data, and medical nuances.	Prompts must provide rich, specific context, define the scope of inquiry, and specify the target audience or clinical setting.	6
Transparency/ Interpretability	The AI's decision-making process is understandable and explainable to human users.	Prompts can instruct the model to show its "thinking process" (e.g., Chain-of-Thought), cite sources, or explain its rationale.	6
Bias Mitigation/ Fairness	Outputs do not perpetuate disparities and perform consistently across diverse patient groups.	Prompts should instruct the model to consider diversity, identify potential biases, and adhere to fairness principles.	6
Safety/ Ethical Compliance	Adherence to patient privacy, data security, informed consent, and responsible AI use.	Prompts must incorporate ethical guidelines, restrict sensitive data handling, and ensure outputs are non-harmful.	6
Completeness	All required elements are included in the response, avoiding partial or missing information.	Prompts should clearly list all necessary components, specify desired depth, and provide examples of comprehensive outputs.	13
Actionability	Outputs provide practical insights that	Prompts should specify the desired format	19

	can be directly applied in clinical practice.	for actionable advice (e.g., recommendations, differential diagnoses, treatment plans).	

table 1 - Key Quality Criteria for Trustworthy AI in Healthcare

3 Advanced Prompt Structuring Frameworks and Techniques

Effective AI prompting in medicine moves beyond simple queries to employ sophisticated frameworks that enhance the model's ability to reason, ensure consistency, and ground responses in verifiable evidence. These advanced techniques are crucial for navigating the complexities and high stakes of clinical applications.

3.1 Chain-of-Thought (CoT) Prompting for Enhanced Clinical Reasoning

Chain-of-Thought (CoT) prompting is a technique that directs an LLM to break down a complex task into a series of smaller, sequential reasoning steps, and then to address each step systematically.[20] This approach significantly improves the reasoning capabilities of LLMs by allowing the model to concentrate on solving one discrete step at a time, rather than attempting to process the entire problem simultaneously.[21]

In medical applications, CoT prompting is particularly valuable. It helps even smaller language models dissect intricate medical queries, thereby enabling more structured reasoning, improving accuracy, and enhancing the interpretability of their outputs.[22] The strength of CoT lies in its ability to mimic human cognitive processes. CoT prompts can be specifically modified to reflect the diagnostic and analytical thought processes utilized by clinicians, such as the formation of differential diagnoses, intuitive reasoning, analytical deduction, and Bayesian inference. This tailored approach has the potential to elicit a deeper and more accurate understanding of LLM performance on complex clinical reasoning tasks.[20] Implementations can range from "Zero-shot CoT," where a simple phrase like "Let's think step by step" is appended to the prompt, encouraging the LLM to generate its own reasoning chain [21], to "Few-shot CoT," which involves providing the model with a few examples of similar problems, complete with their step-by-step rationales, to guide its reasoning strategy.[15] Research has consistently shown that CoT prompting is robust across various linguistic styles, annotators, and language models, consistently outperforming standard baseline prompting methods.[21] The effectiveness of these techniques, particularly CoT, in mirroring human cognitive processes for complex problem-solving suggests

an underlying principle: by structuring prompts to guide AI through "thought processes" analogous to human clinical reasoning, more accurate, reliable, and interpretable AI outputs in medicine can be unlocked. This implies that future advancements in prompting may increasingly draw from cognitive science.

3.2 Self-Consistency Prompting for Improved Reliability and Accuracy

Self-consistency prompting is a technique designed to enhance the reliability and accuracy of LLM outputs, particularly for tasks requiring multi-step reasoning.[21] Instead of generating a single response, this method directs the AI to produce multiple diverse chains of thought for the same problem and then selects the answer that appears most consistently across these generated responses.[21]

This approach significantly increases the probability of finding an accurate answer by cross-checking multiple generated responses, thereby improving overall accuracy and reducing errors in reasoning.[23] It helps mitigate bias by considering various reasoning paths, leading to a more balanced and reliable final output, which is critically important in high-stakes environments like healthcare.[23] Self-consistency enables the model to effectively handle complex or ambiguous tasks by evaluating multiple perspectives, resulting in more accurate and comprehensive answers.[23] A notable advantage is its unsupervised nature; it is compatible with pre-trained LLMs and does not require additional human annotation, training, fine-tuning, or model architectural changes.[21] While highly effective, a consideration for its use is the potential for longer response times due to the generation and aggregation of multiple answers. It may also be less suitable for problems that demand a single, exact answer, such as straightforward mathematical calculations.[23]

3.3 Retrieval-Augmented Generation (RAG) for Evidence Grounding and Hallucination Mitigation

Retrieval-Augmented Generation (RAG) is a powerful technique developed to address inherent limitations of LLMs, such as their knowledge boundaries and the high computational costs associated with continuous retraining.[7] RAG operates by dynamically retrieving external information, such as up-to-date clinical guidelines, peer-reviewed medical literature, or specialized medical databases, and incorporating this information directly into the LLM's generation process.[7]

This method significantly reduces the phenomenon of "hallucination" by grounding LLM

responses in factual, current, and relevant external information, thereby enhancing both the accuracy and reliability of the generated outputs.[7] RAG maintains the original LLM architecture, offering greater flexibility and control compared to extensive fine-tuning. It allows LLMs to adapt to dynamic environments by delivering real-time, up-to-date information.[7] A key advantage in biomedical applications is RAG's ability to integrate external knowledge sources with high interpretability, making the source of information traceable.[7] For instance, an LLM might initially omit a specific drug recommendation, but after integrating retrieved text from current clinical guidelines via RAG, it can then correctly provide the missing information.[7] Prompt engineering plays a significant role in optimizing RAG's effectiveness, including designing prompts that directly aim to reduce hallucinations and employing prompt techniques to automate the construction of resources for hallucination mitigation.[25] However, it is important to acknowledge that limitations within RAG components, such as issues with data sources, query formulation, retriever mechanisms, or retrieval strategies, can still contribute to the generation of confabulations, necessitating ongoing research and optimization of aspects like retrieval granularity and embedding models.[25] The success of RAG highlights the critical need for continuous, real-time grounding of AI outputs in the latest medical evidence, transforming LLMs from mere knowledge recall machines into dynamic information integrators, making them significantly more viable for real-time, evidence-based clinical applications.

3.4 General Best Practices for Crafting Effective Medical Prompts

Beyond specific frameworks, several general best practices are crucial for crafting effective prompts in medical applications:

- **Clarity and Specificity:** Prompts must be precise about the desired output, avoiding vague or ambiguous instructions.[15] For concise responses, explicitly request brevity; for complex, expert-level outputs, specify the required format and depth.[27]
- **Context Setting:** Provide clear and focused context to enable the model to produce accurate and relevant outputs.[15] In healthcare, this often means dynamically formatting prompts to include relevant, hard data pertinent to the query or operation.[12]
- **Examples (Few-Shot Learning):** Including 2-3 well-chosen examples directly within the prompt can effectively demonstrate the desired tone, format, or context, making it easier for the model to align with the user's intent.[15]
- **Structured Output:** Define the exact structure and presentation for the output to ensure consistency and ease of use. Using format constraints guides the model and limits interpretation errors.[15]
- **Break Down Complex Tasks:** Divide broad or multi-goal tasks into smaller, simpler, and

logical steps.[28] This approach improves the depth and structure of the AI's response, effectively guiding it to "think aloud" through the problem.[31]

- **Iterative Refinement and Feedback Loops:** Start with a broad prompt and progressively narrow it based on the AI's responses.[31] If the AI's output misses the mark, reframe or build upon the previous prompts. A powerful technique involves asking the LLM itself to refine its own prompt based on additional context provided by the user.[28]
- **Self-Correction/Self-Evaluation:** Instructing the LLM to review and rate its own answer against predefined scales or specific rules can significantly improve output quality.[15]
- **Manage Certainty:** LLMs often respond with absolute certainty, even when their information is unsupported by facts.[12] In medical systems, this can be misleading and unethical. Techniques like structured prompts, markdown formatting, or the ReAct prompt pattern (which enables reasoning loops) can help manage this overconfidence.[12]
- **Multi-Agent Systems:** For highly complex healthcare tasks, it is often beneficial to split the work across specialized AI agents. For example, a main agent might interact with the clinician, while sub-agents handle image interpretation, tabular data analysis, or EHR data retrieval. This modular approach keeps individual LLMs focused and reliable.[12]
- **Input Filters and Guardrails:** Implementing moderators to filter user inputs is essential. These can block obviously harmful messages, attempts to bypass the system, or off-topic questions.[12] Guardrails can also ensure that the AI maintains a specific tone of voice or prevents unwanted outputs, enhancing safety and compliance.[12]

Technique	Mechanism	Primary Benefit in Medicine	Ideal Medical Use Cases	Limitations/Considerations	Ref
Chain-of-Thought (CoT)	Instructs LLM to break down complex tasks into sequential reasoning steps.	Enhances clinical reasoning, improves interpretability, and supports structured problem-solving.	Diagnostic support, complex case analysis, differential diagnosis formation, medical education.	Requires sufficiently large LLMs; sensitivity to prompt design; coherence of steps is crucial.	20
Self-Consistency	Generates multiple diverse reasoning paths for the same problem and selects the most consistent	Improves reliability and accuracy, reduces bias, and handles ambiguity by evaluating multiple perspectives.	Complex treatment planning, drug discovery, medical imaging analysis, genomic analysis, high-stakes decision	Longer response times; may struggle with tasks requiring one exact answer; higher computational cost.	21

	answer.		support.		
Retrieval-Augmented Generation (RAG)	Retrieves external, up-to-date information (e.g., guidelines, databases) to ground LLM responses.	Mitigates hallucinations, ensures factual accuracy, provides real-time evidence grounding, and enhances interpretability.	Evidence synthesis, drug information queries, clinical guideline adherence, patient-specific information retrieval.	Dependent on quality of external data sources; limitations in RAG components can still lead to issues; requires ongoing optimization.	7

table 2 - Comparison of Advanced Prompting Techniques for Medical AI

4 Iterative Prompt Development and Documentation

The development of effective prompts for AI in medicine is not a one-time event but an ongoing, iterative process. This approach is fundamental to achieving and maintaining the high standards of accuracy, reliability, and safety required in clinical settings.

4.1 Principles of Iterative Prompt Refinement

Iterative prompt refinement is a systematic process of continuously adjusting and optimizing prompts to enhance the relevance, accuracy, and depth of AI outputs.[31] This methodology is built upon two core principles: continuous improvement through feedback loops and structured experimentation.[15] The process typically begins with crafting a clear and specific initial prompt, followed by a thorough review of the AI's generated output. This review assesses accuracy, relevance, format, and completeness.[15] Based on the identified shortcomings in the output, the prompt is then refined. This iterative cycle is crucial for aligning AI results with specific clinical goals, identifying and rectifying issues early in the development process, and ensuring consistency across similar tasks.[15]

Each prompt can be considered a hypothesis, and every AI response serves as feedback that helps sharpen the subsequent question, mirroring the adaptive nature of qualitative research methods.[31] Key adjustments during refinement include adding specific constraints (e.g., word count, desired format), providing more illustrative examples, clarifying ambiguous terms, and precisely specifying the required level of detail or depth.[15] This cyclical approach, akin to a Continuous Quality Improvement (CQI) cycle, is not just a best practice but a necessity in healthcare. Given the high-stakes nature of medicine, where continuous improvement is

paramount for patient safety and outcomes, applying a formal CQI mindset to prompt development is essential. It elevates prompt engineering from a one-off task to an ongoing, systematic process embedded within the quality management framework of healthcare organizations deploying AI.

4.2 Strategies for Version Control and Management of Prompts

For AI prompts to be reliable and trustworthy in clinical applications, they must be managed with the same rigor applied to application code. This entails implementing robust version control, systematic testing, and structured deployment processes.[14] Effective prompt versioning encompasses several critical elements: maintaining a comprehensive version history that clearly documents what changes were made and the rationale behind them; the capability to swiftly roll back to previous versions if issues arise; thorough testing of prompts before deployment; the ability to manage different prompt variations for A/B testing; and tracking which prompt versions are actively running in various environments.[14]

The benefits of such meticulous management are substantial. It ensures transparency by creating an audit trail of how the AI makes decisions, which is vital for understanding or explaining system behavior.[14] It fosters accountability by quickly identifying who made changes and why.[14] It builds reliability and trust by ensuring consistent AI behavior and predictable outputs.[14] Furthermore, versioning supports reproducibility of results, allowing specific outputs to be recreated using exact prompt configurations from any point in time.[14] It also facilitates improved experimentation with new prompt variations without the risk of losing functional versions, and enhances collaboration among team members.[14] Semantic Versioning (X.Y.Z) is a recommended approach for tracking major, minor, and patch updates to prompts, providing a clear and standardized method for managing changes.[32] Tools such as Latitude, Lilypad, and LangSmith are available to simplify prompt management, versioning, and recovery, offering features like automated tracking, dependency tracing, built-in rollback functionality, and performance monitoring.[15]

The analogy between prompts and "living protocols" is particularly apt in healthcare. If a prompt dictates AI behavior that directly influences clinical decisions or patient information, then changes to that prompt are analogous to changes in a clinical protocol. Therefore, prompts must be governed with the same rigor, transparency, and auditability as clinical protocols or software code in a regulated environment. This implies a need for formal change management, review boards, and clear accountability for prompt modifications, particularly in high-risk applications.

4.3 Documenting Prompt Iterations and Performance

Meticulous documentation is paramount for maintaining the quality, reproducibility, and accountability of AI prompts in healthcare. This involves systematically recording the rationale behind changes and the intended objectives of each modification.[32] For every version of a prompt, it is essential to track relevant metadata and the expected outcomes.[14]

Furthermore, continuous monitoring of how prompt changes affect performance metrics, such as user satisfaction and the overall quality of the AI's output, is crucial.[32] Maintaining clear, detailed records of each prompt version and its corresponding output is a fundamental practice.[15] This comprehensive documentation aids significantly in debugging issues that may arise and provides an indispensable audit trail for compliance purposes.[14] Beyond the prompt itself, it is important to define and record the specific, explicit, and justified purposes for which the AI system will utilize private data.[33] This includes documenting the entire model development lifecycle, encompassing training, validation, deployment, and ongoing monitoring, to ensure full transparency and accountability throughout the AI system's operational lifespan.[34]

5 Integrating Evidence and Appraising AI Outputs in Medicine

The integration of AI into evidence-based medicine requires a sophisticated understanding of how LLMs interact with and generate medical information, coupled with robust methods for appraising the quality and reliability of their outputs.

5.1 Role of LLMs in Evidence Synthesis and Clinical Guideline Development

Systematic reviews are cornerstones of evidence-based healthcare, providing comprehensive and unbiased syntheses of research data to inform clinical and public health decisions.[5] LLMs offer significant opportunities to streamline and enhance the production of these labor-intensive and time-consuming reviews.[5] LLMs can assist in generating Boolean search queries and developing search strategies by selecting suitable search terms or translating database syntax.[5] They also have the potential to identify and abstract relevant study information, such as characteristics, methods, and results, from full-text sources, with promising accuracy rates reported between 80% and 96%.[5] Furthermore, LLMs can facilitate the drafting of standardized sections within review manuscripts, improving the quality of academic writing.[5]

However, the use of LLMs for search strategies can introduce issues, such as the creation of

misleading controlled vocabulary, which often requires specialized information retrieval expertise to detect.[5] Crucially, a final human judgment remains indispensable for confirming the semantic and pragmatic fidelity of complex texts generated or processed by LLMs.[5] LLMs can also contribute to evidence integration by gathering scattered evidence from multiple sources and integrating it using advanced knowledge hypergraph-based evidence management models.[35] For complex queries, Importance-Driven Evidence Prioritization (IDEP) algorithms can leverage LLMs to generate multiple evidence features with associated importance scores, which are then used to rank the evidence and produce refined retrieval results.[35]

5.2 Critical Appraisal of AI-Generated Information (e.g., GRADE, AMSTAR 2, PRISMA)

AI-generated medical advice is not yet reliable enough to replace human experts.[13] Studies indicate that LLM responses to medical questions often contain significant inaccuracies or are incomplete.[13] The quality of LLM responses is highly contingent upon the informedness and specificity of the prompt provided.[36] This highlights a fundamental "uncertainty principle" of AI in medicine: LLMs tend to respond with absolute certainty, even when their assertions are unsupported by facts.[12] This creates a dangerous "certainty gap" where AI presents unreliable information with high confidence, underscoring the need for rigorous appraisal frameworks and human oversight, as the AI itself cannot reliably self-assess its certainty in a clinically meaningful way.

To critically appraise AI-generated information and integrate it responsibly into evidence-based practice, established methodologies for evidence appraisal are crucial:

- **GRADE (Grading of Recommendations Assessment, Development and Evaluation):** GRADE provides a transparent and structured framework for assessing the certainty of evidence and formulating recommendations in healthcare.[37] It classifies evidence certainty into four categories: high, moderate, low, or very low.[39] The process begins by categorizing the study design (e.g., Randomized Controlled Trials (RCTs) start at a high level of certainty, while non-randomized studies typically start at a low level).[39] Five domains can downgrade the certainty of evidence: risk of bias, inconsistency, indirectness, imprecision, and publication bias.[37] Conversely, three criteria can upgrade the evidence level of non-randomized studies: strength of association, dose-response gradient, and the effect of opposing plausible residual confounding or bias.[37] GRADE explicitly separates the certainty of evidence from the strength of recommendations, providing clear and comprehensive criteria for rating up or down evidence quality.[39] It acknowledges that expert judgment is involved but enhances transparency by making these judgments explicit and structured.[39]

- **AMSTAR 2 (A MeaSurement Tool to Assess systematic Reviews):** AMSTAR 2 is a critical appraisal tool specifically designed for systematic reviews that include both randomized and non-randomized studies of healthcare interventions.[41] Its purpose is to help users distinguish high-quality reviews, and it is particularly well-suited for systematic reviews that incorporate real-world observational evidence.[41] AMSTAR 2 focuses on identifying weaknesses in critical domains rather than generating an overall numerical score for the review's quality.[42]
- **PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses):** PRISMA represents an evidence-based minimum set of items for reporting systematic reviews and meta-analyses.[43] It comprises a 27-item checklist and a four-phase flow diagram (identification, screening, inclusion, and exclusion) designed to promote transparency and organization throughout the systematic review process.[43] While PRISMA is not a quality assessment instrument itself, it ensures comprehensive and transparent reporting, which is foundational for subsequent critical appraisal.[43]

Integrating AI outputs into evidence-based medicine requires a "hybrid appraisal" approach. Traditional tools like GRADE, AMSTAR 2, and PRISMA are designed for appraising human-generated research evidence. While LLMs can generate "medical advice" and extract study data, these appraisal tools are not explicitly designed for AI-generated *content* as primary evidence. This highlights a critical emerging need: developing or adapting appraisal frameworks that can rigorously evaluate AI-generated content as a form of evidence, or at least as a significant input to traditional evidence. This "hybrid appraisal" would involve not just assessing the AI's accuracy but also its provenance, the prompt engineering applied, and the human oversight involved, creating a new layer of evidence evaluation unique to AI in medicine.

5.3 The Indispensable Role of Human Validation and Oversight

Despite the advancements in AI, these tools should not be viewed as substitutes for medical professionals. Instead, they serve as supplementary resources that, when combined with human expertise, can significantly enhance the overall quality of information provided in healthcare.[13] A significant risk associated with LLM availability and use is over-reliance, which could potentially lead to reduced critical thinking or independent decision-making by healthcare professionals.[5] It is crucial to view these models as tools to augment human expertise rather than replace it entirely.[6]

External validation is essential to ensure that AI tools are accurate and reliable when deployed in real-world clinical settings.[45] Clinical experts, such as data managers, play a crucial role in refining

AI models by providing direct feedback on their predictions and suggesting additional features to enhance performance.[46] The ongoing need for human oversight and verification of AI-generated information is consistently emphasized.[13] Medical educators are encouraged to increasingly incorporate LLMs into their teaching curricula, leveraging the models' limitations to prompt students to consciously provide their own reasoning and validation, thereby fostering a strong sense of accountability.[4] In academic integrity contexts, AI detection tools are employed, and students may be required to explain AI-generated content or provide version histories of their work to confirm originality.[47]

6 Specific Medical Use Cases for AI Prompting

AI prompting is pivotal in unlocking the potential of LLMs across diverse medical applications, enhancing efficiency, accuracy, and patient engagement.

6.1 Clinical Decision Support Systems (CDSS)

LLMs are being actively explored as powerful and innovative tools to assist healthcare practitioners in critical clinical reasoning and decision-making processes.[4] They move beyond the capabilities of simple guideline consultation by engaging in active interactions with users and directly interpreting complex medical information.[4] In this capacity, LLMs can provide personalized insights into likely diagnoses, suggest appropriate specialists, and help assess the urgency of care needs.[3] They are capable of assisting clinicians in refining diagnoses and decision-making, optimizing triage processes—for example, by prioritizing patients based on symptoms and vital signs—and generally improving patient management.[3] Effective prompt engineering is critical in these scenarios, guiding LLMs to predict triage categories, specialty referrals, and diagnoses based on both general user input and specific clinical data.[3]

6.2 Patient Education and Empowerment

LLMs hold significant promise for improving patient education and empowerment by enabling more personalized medical care and broadening access to medical knowledge.[2] These models can assist patients in better understanding their health conditions and various treatment options by providing clear answers to medical questions and translating complex medical information into more accessible language.[2] Furthermore, LLMs have the potential to guide patients in interpreting their symptoms, recommending appropriate specialists, and determining the best course of action, thereby empowering them to actively participate in their healthcare decisions.[3] Beyond efficiency gains for clinicians, a significant implication of AI prompting in medicine is its potential to democratize medical knowledge and foster greater patient autonomy and shared

decision-making, a key trend in modern healthcare. Effective prompting can tailor complex medical information to individual patient literacy and preferences, bridging knowledge gaps.

6.3 Medical Documentation and Information Management

LLMs offer substantial capabilities for streamlining administrative tasks within clinical practice. This includes efficiently extracting clinical information from electronic health records, summarizing lengthy medical texts, structuring unstructured data, and explaining complex medical concepts.[2] They can also assist with general clinical paperwork and the generation of patient information, such as medical text summarization or translation services.[2] Prompt engineering is instrumental in these applications, enabling LLMs to restructure, refine, and enhance technical content, including medical documentation, by specifying desired style and tone or by integrating reference material through Retrieval-Augmented Generation (RAG).[28]

6.4 Other Emerging Applications

The versatility of LLMs, guided by effective prompting, extends to several other emerging applications in medicine:

- **Drug Discovery:** Self-consistent models can rapidly examine numerous chemical compounds, assessing their potential as therapeutic agents and simultaneously checking for potential side effects, a process significantly faster than human-led efforts. This capability could accelerate the discovery of new drugs and even facilitate the design of gene-tailored medications.[24]
- **Medical Imaging Analysis:** AI provides a more accurate and efficient means of detecting health problems in various medical images, including X-rays, CT scans, and MRIs. Self-consistent models are particularly adept at examining organ shapes, sizes, and structures to identify subtle anomalies, potentially leading to highly personalized care plans and more accurate predictions of treatment efficacy.[24]
- **Genomic Analysis:** LLMs can analyze a patient's genetic information, scrutinize genetic markers, and identify variations in DNA sequences. This capability can help in determining susceptibility to certain diseases, informing personalized treatment options, or revealing inherited conditions.[24]
- **Smart Insulin Pens:** Connected insulin pens and caps enhance precision and adherence in insulin injection therapy, providing actionable insights for both individuals and care teams.[19] These smart pens offer clinical decision support by integrating insulin dosing with glucose and meal data, featuring dose calculators and active insulin tracking to enable safer and more effective insulin management, reducing the risk of hypo- and hyperglycemia.[19] Future iterations of smart insulin pens are projected to leverage AI to

determine and fine-tune insulin therapy settings and titrate insulin doses autonomously.[49]

- **Challenges to Implementation:** Despite their advantages, several barriers impede the widespread adoption of smart insulin pens. These include their high cost (ranging from USD 500 to USD 900, compared to less than USD 100 for traditional pens), limited or no insurance coverage, insufficient awareness and training among healthcare professionals, patient preferences, poor digital competence, and the lack of compatible smartphones.[49]
- **Facilitators to Implementation:** Factors that facilitate their adoption include improvements in diabetes clinic visit quality, their utility as an alternative to insulin pump therapy, a growing inclination towards remote monitoring and digital health solutions, and increasing disposable incomes that enable consumers to afford smart healthcare devices.[49] The challenges in adopting specific AI-powered medical devices are not isolated technical issues but rather reflections of systemic barriers within healthcare, such as fragmented reimbursement, gaps in professional training, and digital literacy disparities. This implies that successful AI integration requires not just robust prompting strategies but also comprehensive health policy changes, educational initiatives, and infrastructure development to address these underlying systemic limitations.

- **Biosimilar Uptake in Hospital Formularies:** Biosimilars significantly improve the affordability and accessibility of biologics, contributing to the long-term sustainability of healthcare systems.[52]
  - **Challenges to Uptake:** Factors limiting biosimilar uptake include a lack of incentives for stakeholders, widespread misinformation and mistrust regarding biosimilars, inadequate reimbursement policies, regulatory hurdles, limited prescriber awareness, issues with interchangeability, the need for more pharmacovigilance data, and variations in manufacturer support programs.[52] Surprisingly, academic hospitals, despite the potential for greater savings, have sometimes shown slower uptake or lower shares for certain biosimilars.[54]
  - **Facilitators to Uptake:** Factors that positively influence biosimilar adoption include a significant price difference between the biosimilar and the originator product, strong relationships between commissioners and providers, proactive leadership, the sharing of best practices from early adopters, multi-stakeholder approaches, and comprehensive communication and education initiatives for key stakeholders.[52] Drugs and Therapeutics Committees (DTCs) play a critical role in evaluating biosimilars and improving their uptake within acute care settings.[53]

7 Ethical, Legal, and Security Considerations

The integration of AI into healthcare necessitates a robust framework of ethical, legal, and security considerations to ensure patient safety, data integrity, and responsible deployment.

7.1 Data Privacy and Security (GDPR, HIPAA, Consent)

The healthcare industry operates with an immense volume of Protected Health Information (PHI) and Sensitive Personal Information (SPI), making data privacy and security paramount.[18] Regulations like the

General Data Protection Regulation (GDPR), originating from the European Union but impacting global healthcare organizations, mandate explicit consent from individuals before their data can be processed by AI systems.[56] Key principles under GDPR include:

- **Data Minimization:** Organizations are required to collect only the data strictly necessary for specific purposes, preventing unnecessary data gathering and protecting patient privacy.[33]
- **Explicit Consent:** Patients must provide clear and unambiguous consent for their data to be utilized by AI systems, necessitating transparent consent processes.[9] Traditional consent frameworks may be insufficient for the complexities of AI, where data can be repurposed; thus, innovative approaches like dynamic consent models are being explored to allow individuals to update their preferences as AI systems evolve.[9]
- **Right to Access and Delete:** Individuals are granted the right to request access to their data and to demand its deletion, fostering patient awareness and control.[33]
- **Anonymization and Pseudonymization:** AI mechanisms should employ these techniques to safeguard individual privacy while still enabling insights from large datasets.[33] However, the risk of re-identification persists, even with de-identified data, particularly with advanced analytics or cross-referencing.[9]
- **Protection and Accountability:** AI systems must integrate robust security practices to prevent data breaches and unauthorized access.[9] Both AI developers and users are held accountable for adhering to GDPR, requiring detailed record-keeping of data manipulation activities and the incorporation of data protection by design and by default.[33]
- **Data Protection Impact Assessments (DPIAs):** These are a requirement for AI systems handling high-risk processes, assisting in detecting and mitigating privacy risks.[33]

In the United States, **HIPAA (Health Insurance Portability and Accountability Act)** and the **CCPA (California Consumer Privacy Act)** are relevant legal frameworks that emphasize transparency, accountability, and privacy in healthcare AI.[9] Technical safeguards such as encryption, de-

identification, and secure data storage are crucial for protecting sensitive information.[9] AI systems are also vulnerable to cyberattacks, including data breaches, model hacking, and adversarial inputs designed to manipulate AI predictions. Addressing these threats requires end-to-end encryption, regular security audits, and real-time monitoring.[9]

7.2 Addressing Bias and Ensuring Fairness in AI Outputs

Bias in AI models is a significant ethical concern, as it can lead to discriminatory outcomes that disproportionately affect certain demographic groups.[9] Bias can originate at various stages: during data collection, if training data is unrepresentative or skewed; from algorithmic design, if the algorithm inherently favors certain outcomes; or from societal norms and stereotypes reflected in the data used to train the AI.[16] Common types of bias include selection bias (unrepresentative training data), confirmation bias (over-reliance on pre-existing patterns), measurement bias (systematic differences in collected data), stereotyping bias (reinforcing harmful stereotypes), and out-group homogeneity bias (generalizing individuals from underrepresented groups).[16]

Mitigation Strategies are multifaceted:

- **Diversify Training Datasets:** Ensuring that training datasets include a wide range of perspectives and demographics is fundamental to reducing bias.[16]
- **Bias Detection Techniques:** Implementing fairness audits, adversarial testing, and using specific fairness metrics (e.g., true positive rates, statistical parity, equalized odds) are crucial for identifying and rectifying bias.[16]
- **Transparency:** Encouraging transparency in AI decision-making helps users understand potential biases and their origins.[16]
- **Continuous Monitoring:** Regularly auditing AI systems after deployment is necessary to detect emerging biases that may develop over time.[16]
- **Human Oversight:** Maintaining human involvement in critical decision-making processes is essential, especially where AI biases could have serious ethical or legal implications.[16]
- **Pre-processing and Post-processing:** Modifying data before training or adjusting outputs after generation can help reduce bias.[57]
- **Fairness Constraints:** Incorporating fairness objectives directly into the model training process can guide the algorithm to produce more equitable outcomes.[57]

The FUTURE-AI guideline explicitly emphasizes fairness, stating that AI tools in healthcare should maintain consistent performance across all individuals and groups, actively identifying, reporting, and minimizing potential biases.[17]

7.3 Accountability, Transparency, and Mitigating Over-Reliance

Many AI systems operate as "black boxes," making it challenging to understand how they arrive at specific decisions. This lack of explainability can erode trust among clinicians and hinder their ability to validate AI recommendations in healthcare.[9] Transparency, therefore, demands clear documentation, traceability, and explainability of AI models and their outputs.[9] Maintaining data and model provenance is crucial for identifying and mitigating risks such as bias. This requires detailed data traceability, including information on data ownership, collection methodologies, and curation procedures, alongside thorough documentation of the entire model development lifecycle—from training and validation to deployment and monitoring.[34] The EU AI Act, for instance, explicitly highlights the need for traceability, explainability, and clear disclosure of an AI system's limitations.[34]

Establishing clear accountability mechanisms is paramount to define roles and responsibilities when AI systems are deployed in healthcare.[9] A significant concern is the risk of over-reliance on LLMs, which could potentially lead to reduced critical thinking or independent decision-making among healthcare professionals.[5] AI systems should be viewed as tools to augment, rather than replace, human expertise.[6] This necessitates continuous monitoring and adaptation of AI systems to ensure they remain effective and do not inadvertently foster dependency.[9]

7.4 Regulatory Frameworks (e.g., EU AI Act) and Compliance

The **EU AI Act** stands as the world's first comprehensive AI law, introducing a risk-based classification system for AI systems.[58] This groundbreaking regulatory framework has profound implications for the medical device industry.

- **High-Risk AI Systems:** Medical devices predominantly fall under this category.[58] Systems classified as high-risk are subject to stringent requirements, including:
  - Comprehensive risk management systems.
  - Rigorous conformity assessments before being placed on the market and throughout their operational lifecycle.
  - Detailed documentation requirements.
  - Ongoing monitoring and reporting obligations.[59]
  - A new certification under AI Regulations, in addition to existing CE certification (under MDR/IVDR).[59]
  - Registration in the EU AI database.[59]

- **Generative AI (e.g., ChatGPT):** While not classified as high-risk, these models must still comply with specific transparency requirements. This includes disclosing that content was AI-generated, designing models to prevent the generation of illegal content, and publishing summaries of copyrighted data used for training.[58]
- High-impact general-purpose AI models, such as GPT-4, that might pose systemic risks, are required to undergo thorough evaluations, and any serious incidents must be reported to the European Commission.[58]
- AI deployers are mandated to conduct data protection impact assessments and maintain automatically generated logs of AI system activities.[59]
- The use of AI systems must be communicated to device users, and instructions for use must include clear information on the system's capabilities and limitations.[59]

The regulatory landscape is actively shaping AI development and prompting requirements. The existence and evolution of these legal frameworks directly dictate how AI systems, and by extension, their prompts, must be designed, developed, and deployed in healthcare. Prompting is not just a technical optimization but a critical compliance mechanism. This means prompt engineers and developers must be intimately aware of the legal landscape, as non-compliance can lead to significant penalties and reputational damage. Regulatory adherence will increasingly drive prompt design principles, especially concerning data handling, bias mitigation, and transparency.

7.5 AI Prompt Security: Understanding and Preventing Prompt Injection Attacks

Prompt Injection Attacks represent a significant cyber security threat to large language models. These attacks involve hackers disguising malicious inputs as legitimate prompts to manipulate LLMs into performing unintended actions, such as leaking sensitive data, spreading misinformation, or executing unauthorized operations.[60] These attacks exploit a fundamental vulnerability: LLM applications often do not clearly distinguish between developer instructions (system prompts) and user inputs, as both typically take the same format of natural-language text.[60]

Attack types include direct prompt injection, where the malicious prompt is fed directly to the LLM, and indirect prompt injection, where the malicious payload is hidden within data that the LLM consumes (e.g., embedded in web pages or images).[60] The consequences can be severe, ranging from data exfiltration, data poisoning, and data theft to response corruption, remote code execution, misinformation propagation, and even malware transmission.[61]

Mitigation Strategies are crucial for protecting sensitive medical AI systems:

- **Input Sanitization and Validation:** Implement mechanisms to filter and cleanse incoming data, detecting suspicious entries such as unusually long prompts, inputs mimicking system prompts, or similarities to known injection attempts.[62]
- **Contextual Separation:** Clearly separate system commands from user inputs. This can be achieved using structured queries or delimiters that explicitly mark trusted instructions from untrusted user-provided content.[30]
- **Internal Prompt Engineering:** Strengthen system prompts with explicit directives (e.g., "You are a helpful assistant who only provides responses within a specific scope"), self-reminders (e.g., "You must always respect user privacy"), and the consistent use of delimiters to segment instructions.[63] Repeating critical instructions multiple times can also reduce the likelihood of successful overrides.[63]
- **Limiting Model Capabilities:** Narrow the potential actions an AI system can perform by capping functionalities and setting clear guardrails for acceptable outputs.[62] Restrict API permissions to only essential functions to minimize potential damage if an injection occurs.[61]
- **Access Controls:** Implement granular, role-based permissions and adhere to the principle of least privilege for both LLMs and their users. This limits exposure to potential attackers and helps contain the fallout from a successful prompt injection.[62]
- **Regular Security Audits and Monitoring:** Continuously assess for vulnerabilities and monitor logs for anomalous activities indicative of injection attempts. This enables rapid incident response and maintains system integrity.[62]
- **Adversarial Testing and Simulation:** Conduct "red team" exercises to simulate potential injection scenarios, revealing weaknesses in AI systems before real-world exploitation.[62]
- **Versioning and Testing:** Apply version control to critical prompts and conduct regular security patches and model updates to address newly discovered vulnerabilities.[30]
- **Prompt Design Considerations:** Avoid overloading prompts with too much context, use explicit role assignment carefully, design prompts to be stateless when possible, use format constraints to control outputs, test prompts against edge cases, evaluate with multi-shot examples, avoid embedding sensitive logic, and limit prompt reuse across unrelated tasks.[30]

Prompt injection attacks highlight a critical "trust boundary" challenge in human-AI interaction. The fact that LLMs do not clearly distinguish between developer instructions and user inputs, treating both as natural language text, allows malicious user input to override intended system behavior. This fundamental ambiguity, combined with the AI's inherent "friendliness and trust," creates a significant vulnerability. Therefore, prompt security measures are not just technical

fixes but attempts to re-establish and enforce this critical boundary within a fluid language interface.

Category	Key Principles/Regulations	Core Challenge/Risk	Mitigation/Best Practice	Ref
Data Privacy & Security	GDPR, HIPAA, CCPA, Patient-Centricity	Re-identification of de-identified data, data breaches, cyberattacks, inadequate consent.	Explicit and dynamic consent, data minimization, anonymization/pseudonymization, encryption, regular security audits, DPIAs.	1
Bias & Fairness	FUTURE-AI guideline, Bias Mitigation	Discriminatory outcomes, perpetuation of disparities, unfair treatment of demographic groups.	Diversified training datasets, bias detection techniques (fairness audits), continuous monitoring, human oversight, fairness constraints.	6
Accountability & Transparency	Traceability, Explainability	"Black-box" nature, undermining trust, difficulty in validating AI recommendations, unclear responsibility.	Clear documentation of model lifecycle, data provenance, interpretable AI models, explicit accountability mechanisms, disclosure of limitations.	6
Over-Reliance	Augment, Not Replace Human Expertise	Reduced critical thinking, diminished independent decision-making by healthcare professionals.	View AI as a supplementary tool, foster critical engagement, continuous monitoring, and training on AI limitations.	5
Regulatory Compliance	EU AI Act (High-Risk, Generative AI), Medical Device Regulations	Legal penalties, reputational damage, lack of trust, unapproved deployment.	Comprehensive risk management, rigorous conformity assessments, detailed documentation, ongoing monitoring, specific certifications, transparency requirements.	58

table 3 - Ethical and Legal Considerations for AI in Healthcare

Attack Type	Description	Potential Impact in Healthcare	Mitigation Strategy	Ref
Direct Injection	Malicious	Unauthorized data access,	Input sanitization/validation,	60

	instructions are directly inserted into the user's prompt.	generation of harmful medical advice, system manipulation.	contextual separation (delimiters), explicit internal instructions.	
Indirect Injection	Malicious payload is hidden within data the LLM processes (e.g., web content, images).	Spreading misinformation (e.g., fake health news), data exfiltration from processed documents, malware transmission.	Input sanitization/validation for all data sources, limiting model capabilities, monitoring for anomalous activity.	60
Code Injection	Attacker injects executable code into the prompt to manipulate responses or actions.	Unauthorized access to sensitive patient messages (e.g., via email assistant), system compromise.	Input sanitization, strict access controls (least privilege), limiting model capabilities (e.g., API access).	61
Multimodal Injection	Malicious prompts embedded in non-textual inputs like images or audio.	AI processing medical images could be tricked into disclosing sensitive patient data or misinterpreting scans.	Input validation for all modalities, adversarial testing, contextual separation.	61
Model Data Extraction	Attackers prompt the LLM to reveal its internal instructions or conversation history.	Exposure of proprietary system prompts, revealing vulnerabilities for future attacks, sensitive data leakage.	Strengthening internal prompts (self-reminders, explicit directives), continuous monitoring, access controls.	61
Exploiting LLM Friendliness/Trust	Using persuasive language or social engineering to convince the LLM to execute unauthorized actions.	AI models disclosing protected health information or generating biased recommendations.	Internal prompt engineering (explicit instructions), limiting model capabilities, adversarial testing.	61

table 4 - Prompt Injection Attack Types and Mitigation Strategies

8 Conclusion and Future Outlook

8.1 Summary of Key Best Practices

The integration of Large Language Models into evidence-based human medicine presents a transformative opportunity, yet it is fraught with inherent risks that necessitate a disciplined and scientifically grounded approach to prompting. The core tenets of effective AI prompting in this high-stakes domain encompass:

- **Scientific Rigor:** Ensuring outputs are factually accurate, reliable, and grounded in the latest medical evidence, actively mitigating the risk of hallucinations.
- **Advanced Prompt Structuring:** Employing sophisticated techniques such as Chain-of-Thought (CoT) for enhanced clinical reasoning, Self-Consistency for improved reliability and accuracy, and Retrieval-Augmented Generation (RAG) for robust evidence grounding and hallucination mitigation.
- **Iterative Development and Documentation:** Treating prompts as living protocols that undergo continuous refinement, rigorous version control, and meticulous documentation to ensure transparency, reproducibility, and accountability.
- **Rigorous Evidence Integration and Appraisal:** Leveraging LLMs to assist in evidence synthesis while critically appraising their outputs using established methodologies like GRADE, AMSTAR 2, and PRISMA, acknowledging that AI-generated information is not yet a substitute for human expert judgment.
- **Unwavering Adherence to Ethical, Legal, and Security Principles:** Prioritizing data privacy (GDPR, HIPAA), actively addressing bias and ensuring fairness, establishing clear accountability and transparency mechanisms, mitigating over-reliance, and implementing robust prompt security measures against sophisticated attacks.

Crucially, the human-in-the-loop approach remains indispensable. Human validation, oversight, and maintaining accountability are paramount for the safe, effective, and ethical deployment of AI applications in clinical settings.

8.2 Recommendations for Stakeholders in Healthcare AI

The responsible and effective integration of AI into human medicine requires concerted efforts from all stakeholders:

- **For Developers/Researchers:**
  - Prioritize hallucination and bias mitigation through the continuous advancement of prompting techniques and robust validation frameworks.

- Integrate regulatory requirements, such as those outlined in the EU AI Act and GDPR, from the initial design phase of AI systems, ensuring compliance is built-in, not an afterthought.
- Implement comprehensive prompt versioning and advanced security measures to protect against prompt injection attacks and other vulnerabilities.
- Focus research on developing interpretable AI models that clearly explain their decision-making processes, fostering trust and clinical utility.

- **For Clinicians/Users:**
  - Actively engage in continuous education to understand the capabilities and, critically, the limitations of LLMs.
  - Maintain and exercise "critical thinking and contextual understanding" when utilizing AI-generated information, never relying solely on AI outputs for clinical decisions.
  - Actively participate in the validation and feedback processes for AI systems, providing invaluable real-world clinical insights for model refinement.
  - Advocate for greater transparency and accountability in AI systems deployed in their practice.
- **For Healthcare Organizations:**
  - Establish clear, comprehensive policies and protocols for AI use, including detailed guidelines for prompt engineering, output appraisal, and human oversight.
  - Invest significantly in tailored training programs for both AI developers and clinical users, bridging the gap between technical capabilities and clinical application.
  - Foster a culture of continuous quality improvement for AI systems, integrating iterative prompt refinement and performance monitoring into standard operational procedures.
  - Ensure robust data governance frameworks are in place to protect sensitive patient information throughout the AI lifecycle.
- **For Regulators/Policymakers:**
  - Continue to develop agile, comprehensive, and adaptive regulatory frameworks that can keep pace with the rapid advancements in AI technology, particularly in high-risk medical applications.
  - Focus regulatory efforts on ensuring safety, transparency, accountability, and fairness across the entire AI lifecycle, from development to deployment and monitoring.
  - Promote international collaboration to harmonize regulatory standards, facilitating global innovation while maintaining patient safety.

8.3 Areas for Future Research and Development

The field of AI in medicine is rapidly evolving, and several key areas warrant focused future

research and development:

- **Prompt Optimization for Evidence Synthesis:** Further research is needed to understand how different types of prompts impact LLM output in evidence synthesis tasks, aiming to develop standardized guidance for review authors on optimal prompt formulation.[5]
- **Specific Guidelines for RAG in Clinical Settings:** The development of more specific, evidence-based guidelines for Retrieval-Augmented Generation (RAG) applications in diverse clinical settings is crucial to maximize its benefits and mitigate its limitations.[7]
- **Advanced Hallucination and Bias Mitigation:** Continued optimization of techniques like fine-tuning, Reinforcement Learning from Human Feedback (RLHF), and RAG is necessary to further enhance the accuracy and reliability of medical LLMs and comprehensively mitigate hallucinations.[8] Research should also focus on robust methods for detecting and correcting hallucinations even after RAG optimization.[25]
- **Interpretable AI Models:** Further research into developing inherently interpretable AI models that can clearly explain their recommendations, moving beyond post-hoc explanations, is vital for building trust and clinical adoption.[9]
- **Dynamic Consent Models:** Exploration and implementation of innovative dynamic consent models are needed to allow individuals to update their preferences regarding data use as AI systems evolve, addressing the complexities of data repurposing in AI.[9]
- **Addressing Systemic Barriers to AI Adoption:** Comprehensive research and policy initiatives are required to address systemic barriers to AI adoption in healthcare, including high costs, limited insurance coverage, and disparities in digital literacy.[49]
- **Hybrid Appraisal Frameworks:** Development or adaptation of critical appraisal frameworks (e.g., GRADE) specifically designed to rigorously evaluate AI-generated content as a form of evidence, or as a component of traditional evidence synthesis, is an emerging necessity.
- **Advanced Prompt Security:** Ongoing research into more sophisticated prompt security techniques is essential to counter increasingly complex and evolving prompt injection attacks, ensuring the integrity and safety of AI systems in healthcare.

References

Last updated : July 17, 2025,

[1] Saikia, M. (n.d.). Large Language Models in Healthcare and Medical Applications: A Review. *ResearchGate*. Retrieved from https://www.researchgate.net/publication/392568866_Large_Language_Models_in_Healthcare_and_Medical_Applications_A_Review

[2] Bretschneider, S., & Giesler, L. (n.d.). Current applications of LLMs in clinical settings peer-reviewed. *PMC NCBI*. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC11751060/

[8] Li, D., & Wu, J. (n.d.). Limitations of LLMs in medical diagnosis research. *PMC NCBI*. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC12163604/

[6] Al-Adhami, M., & Al-Adhami, M. (n.d.). Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology - a recent scoping review. *ResearchGate*. Retrieved from https://www.researchgate.net/publication/378525063_Challenges_and_barriers_of_using_large_language_models_LLM_such_as_ChatGPT_for_diagnostic_medicine_with_a_focus_on_digital_pathology_-_a_recent_scoping_review

[10] Google Cloud. (n.d.). What is prompt engineering. *Google Cloud*. Retrieved from https://cloud.google.com/discover/what-is-prompt-engineering

[11] Amazon Web Services. (n.d.). What is Prompt Engineering? *AWS*. Retrieved from https://aws.amazon.com/what-is/prompt-engineering/#:~:text=Prompt%20engineering%20is%20the%20process,high%2Dquality%20and%20relevant%20output.

[5] BMJ. (n.d.). Role of LLMs in evidence synthesis clinical guidelines. *BMJ EBM*. Retrieved from https://ebm.bmj.com/content/early/2025/01/09/bmjebm-2024-113320.full.pdf

[7] Kim, J., & Lee, J. (n.d.). Role of LLMs in evidence synthesis clinical guidelines. *Academic OUP*. Retrieved from https://academic.oup.com/jamia/article/32/4/605/7954485

[3] Niu, S., & Wang, Y. (n.d.). LLMs in clinical decision support systems. *PMC NCBI*. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC12064692/

[4] JMIR. (n.d.). LLMs in clinical decision support systems. *JMIR Med Educ*. Retrieved from https://mededu.jmir.org/2025/1/e55709

[17] Holzinger, A., & Kieseberg, P. (n.d.). Quality criteria for AI in medicine transparency traceability reproducibility evidence alignment bias awareness validation. *BMJ*. Retrieved from https://www.bmj.com/content/388/bmj-2024-081554

[34] Kourou, K., & Tsatsos, D. (n.d.). Quality criteria for AI in medicine transparency traceability reproducibility evidence alignment bias awareness validation. *arXiv*. Retrieved from https://arxiv.org/html/2506.22358v1

[58] European Parliament. (n.d.). EU AI Act: first regulation on artificial intelligence. *European Parliament*. Retrieved from https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

[59] PharmaLex. (n.d.). What global AI regulations mean for medical device manufacturers. *PharmaLex*. Retrieved from https://www.pharmalex.com/thought-leadership/blogs/what-global-ai-regulations-mean-for-medical-device-manufacturers/

[56] Simbo.AI. (n.d.). The impact of GDPR on AI systems: Ensuring ethical and secure data processing in healthcare. *Simbo.AI*. Retrieved from https://www.simbo.ai/blog/the-impact-of-gdpr-on-ai-systems-ensuring-ethical-and-secure-data-processing-in-healthcare-4013515/

[33] Exabeam. (n.d.). The intersection of GDPR and AI and 6 compliance best practices. *Exabeam*. Retrieved from https://www.exabeam.com/explainers/gdpr-compliance/the-intersection-of-gdpr-and-ai-and-6-compliance-best-practices/

[18] Frontiers in Artificial Intelligence. (n.d.). Ethical guidelines for AI in clinical settings peer-reviewed. *Frontiers in Artificial Intelligence*. Retrieved from https://www.frontiersin.org/journals/artificial-

intelligence/articles/10.3389/frai.2025.1619463/full

[9] Austin-Gabriel, C., & Monsalve, P. (n.d.). Developing Ethical AI Models in Healthcare: A US Legal and Compliance Perspective on HIPAA and CCPA. *ResearchGate*. Retrieved from https://www.researchgate.net/publication/390208554_Developing_Ethical_AI_Models_in_Healthcare_A_US_Legal_and_Compliance_Perspective_on_HIPAA_and_CCPA

[25] MDPI. (n.d.). Retrieval-Augmented Generation (RAG) in medical LLMs hallucination reduction. *MDPI*. Retrieved from https://www.mdpi.com/2227-7390/13/5/856

[26] Al-Adhami, M., & Al-Adhami, M. (n.d.). Retrieval-Augmented Generation (RAG) in medical LLMs hallucination reduction. *PMC NCBI*. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC12157099/

[22] PubMed. (n.d.). Chain-of-Thought (CoT) prompting medical applications. *PubMed*. Retrieved from https://pubmed.ncbi.nlm.nih.gov/40380574/

[20] Niu, S., & Wang, Y. (n.d.). Chain-of-Thought (CoT) prompting medical applications. *PMC NCBI*. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC10808088/

[24] Digital Adoption. (n.d.). Self-consistency prompting medical AI benefits. *Digital Adoption*. Retrieved from https://www.digital-adoption.com/self-consistency-prompting/

[23] GeeksforGeeks. (n.d.). Self-consistency prompting medical AI benefits. *GeeksforGeeks*. Retrieved from https://www.geeksforgeeks.org/artificial-intelligence/self-consistency-prompting/

[21] Mercity.AI. (n.d.). Advanced prompt engineering techniques for clinical decision support. *Mercity.AI*. Retrieved from https://www.mercity.ai/blog-post/advanced-prompt-engineering-techniques

[27] Acorn.io. (n.d.). Advanced prompt engineering techniques for clinical decision support. *Acorn.io*. Retrieved from https://www.acorn.io/resources/learning-center/prompt-engineering/

[15] Latitude. (n.d.). Iterative prompt development best practices AI. *Latitude Blog*. Retrieved from https://latitude-blog.ghost.io/blog/iterative-prompt-refinement-step-by-step-guide/

[31] Indeemo. (n.d.). Iterative prompt development best practices AI. *Indeemo Blog*. Retrieved from https://indeemo.com/blog/iterative-prompting-generative-ai

[14] LaunchDarkly. (n.d.). Version control for AI prompts and models. *LaunchDarkly Blog*. Retrieved from https://launchdarkly.com/blog/prompt-versioning-and-management/

[32] Latitude. (n.d.). Prompt versioning best practices. *Latitude Blog*. Retrieved from https://latitude-blog.ghost.io/blog/prompt-versioning-best-practices/

[29] Towards Data Science. (n.d.). Documentation of LLM prompt iterations and outputs. *Towards Data Science*. Retrieved from https://towardsdatascience.com/boost-your-llm-outputdesign-smarter-prompts-real-tricks-from-an-ai-engineers-toolbox/

[28] ScoutOS. (n.d.). Top 5 LLM prompts for re-writing your technical documentation. *ScoutOS Blog*. Retrieved from https://www.scoutos.com/blog/top-5-llm-prompts-for-re-writing-your-technical-documentation

[16] Chapman University. (n.d.). Tracking biases in AI model development. *Chapman University*. Retrieved from https://www.chapman.edu/ai/bias-in-ai.aspx

[57] Sapien.io. (n.d.). Bias in AI models and generative systems. *Sapien.io Blog*. Retrieved from https://www.sapien.io/blog/bias-in-ai-models-and-generative-systems

[36] PMC NCBI. (n.d.). Integrating evidence from LLM outputs medical guidelines. *PMC NCBI*. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC12149300/

[35] arXiv. (n.d.). Integrating evidence from LLM outputs medical guidelines. *arXiv*. Retrieved from https://arxiv.org/html/2503.16530v1

[41] BMJ. (n.d.). Appraising evidence from AI outputs GRADE AMSTAR-2. *BMJ*. Retrieved from https://www.bmj.com/content/bmj/358/bmj.j4008.full.pdf

[42] AMSTAR. (n.d.). AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare inter. *AMSTAR*. Retrieved from https://amstar.ca/Amstar-2.php

[45] Simbo.AI. (n.d.). The role of external validation in artificial intelligence deployment: Strategies for enhancing performance in real-world clinical settings. *Simbo.AI*. Retrieved from https://www.simbo.ai/blog/the-role-of-external-validation-in-artificial-intelligence-deployment-strategies-for-enhancing-performance-in-real-world-clinical-settings-2073834/

[46] Applied Clinical Trials. (n.d.). Enhancing clinical data validation with AI: A comprehensive approach to accuracy

and efficiency. *Applied Clinical Trials*. Retrieved from https://www.appliedclinicaltrialsonline.com/view/enhancing-clinical-data-validation-with-ai-a-comprehensive-approach-to-accuracy-and-efficiency

[44] DistillerSR. (n.d.). What is a PRISMA flow diagram. *DistillerSR*. Retrieved from https://www.distillersr.com/resources/systematic-literature-reviews/what-is-a-prisma-flow-diagram

[43] UNC Libraries. (n.d.). PRISMA-style flow for evidence retrieval AI. *UNC Libraries*. Retrieved from https://guides.lib.unc.edu/prisma

[13] News-Medical.net. (n.d.). Study: AI-generated medical advice not yet reliable enough to replace human experts. *News-Medical.net*. Retrieved from https://www.news-medical.net/news/20240402/Study-AI-generated-medical-advice-not-yet-reliable-enough-to-replace-human-experts.aspx

[40] CDC. (n.d.). Chapter 7: GRADE Criteria for Determining Certainty of Evidence. *CDC*. Retrieved from https://www.cdc.gov/acip-grade-handbook/hcp/chapter-7-grade-criteria-determining-certainty-of-evidence/index.html

[37] Medwave. (n.d.). The GRADE (Grading of Recommendations Assessment, Development and Evaluation) approach. *Medwave*. Retrieved from https://www.medwave.cl/revisiones/metodinvestreport/8109.html?lang=en

[38] UNC Libraries. (n.d.). GRADE framework application in clinical guidelines. *UNC Libraries*. Retrieved from https://chs.libguides.com/c.php?g=1149014&p=9263649

[39] GRADEpro. (n.d.). Overview of the GRADE approach. *GRADEpro*. Retrieved from https://book.gradepro.org/guideline/overview-of-the-grade-approach

[47] Reddit. (n.d.). How do grade students with 100% AI generated essays. *Reddit*. Retrieved from https://www.reddit.com/r/Professors/comments/1iknpm3/how_do_grade_students_with_100_ai_generated_essays/

[48] University of Melbourne. (n.d.). Advice for students regarding Turnitin and AI writing detection. *University of Melbourne*. Retrieved from https://academicintegrity.unimelb.edu.au/plagiarism-and-collusion/artificial-intelligence-tools-and-technologies/advice-for-students-regarding-turnitin-and-ai-writing-detection

[50] Ebekozien, O., & Fantasia, K. L. (n.d.). Facilitators and Barriers to Smart Insulin Pen Use: A Mixed-Method Study of Multidisciplinary Stakeholders From Diabetes Teams in the United States. *ResearchGate*. Retrieved from https://www.researchgate.net/publication/363606261_Facilitators_and_Barriers_to_Smart_Insulin_Pen_Use_A_Mixed-Method_Study_of_Multidisciplinary_Stakeholders_From_Diabetes_Teams_in_the_United_States?_tp=eyJjb250ZXh0Ijp7InBhZ2UiOiJzY2llbnRpZmljQ29udHJpYnV0aW9ucyIsInByZXZpb3VzUGFnZSI6bnVsbH19

[19] ADCES. (n.d.). Smart Pens in Diabetes Management. *ADCES*. Retrieved from https://www.adces.org/education/danatech/insulin-medicine-delivery/insulin-medicine-delivery-101/smart-pens-in-diabetes-management

[52] Taylor & Francis Online. (n.d.). Biosimilar uptake hospital formularies OECD regulatory economic clinical factors. *Taylor & Francis Online*. Retrieved from https://www.tandfonline.com/doi/full/10.1080/14712598.2025.2507173?src=

[53] PMC NCBI. (n.d.). Biosimilar uptake hospital formularies OECD regulatory economic clinical factors. *PMC NCBI*. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC6476576/

[51] Bioscientifica. (n.d.). Challenges of smart insulin pens in EU outpatient care. *Bioscientifica*. Retrieved from https://ec.bioscientifica.com/view/journals/ec/12/11/EC-23-0108.xml

[49] MarkNtel Advisors. (n.d.). Challenges of smart insulin pens in EU outpatient care. *MarkNtel Advisors*. Retrieved from https://www.marknteladvisors.com/research-library/europe-smart-insulin-pens-market.html

[55] PMC NCBI. (n.d.). Factors influencing biosimilar adoption France Germany UK hospitals. *PMC NCBI*. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC7803694/

[54] Taylor & Francis Online. (n.d.). Factors influencing biosimilar adoption France Germany UK hospitals. *Taylor & Francis Online*. Retrieved from https://www.tandfonline.com/doi/full/10.1080/14737167.2023.2146579

[60] IBM. (n.d.). Prompt injection attacks LLM security vulnerabilities. *IBM*. Retrieved from https://www.ibm.com/think/topics/prompt-injection

[61] Palo Alto Networks. (n.d.). Prompt injection attacks LLM security vulnerabilities. *Palo Alto Networks*. Retrieved

from https://www.paloaltonetworks.com/cyberpedia/what-is-a-prompt-injection-attack

[62] Sprocket Security. (n.d.). Mitigation strategies for prompt injection attacks. *Sprocket Security Blog*. Retrieved from https://www.sprocketsecurity.com/blog/prompt-injection

[63] Helicone. (n.d.). Preventing prompt injection. *Helicone Blog*. Retrieved from https://www.helicone.ai/blog/preventing-prompt-injection

[30] Palo Alto Networks. (n.d.). AI security in healthcare prompt engineering. *Palo Alto Networks*. Retrieved from https://www.paloaltonetworks.com/cyberpedia/what-is-ai-prompt-security-

[12] The Momentum. (n.d.). Effective AI prompting strategies for healthcare applications. *The Momentum Blog*. Retrieved from https://www.themomentum.ai/blog/effective-ai-prompting-strategies-for-healthcare-applications

Additional References

1. Large Language Models in Healthcare and Medical Applications: A Review - ResearchGate, https://www.researchgate.net/publication/392568866_Large_Language_Models_in_Healthcare_and_Medical_Applications_A_Review
2. Current applications and challenges in large language models for patient care: a systematic review - PMC - PubMed Central, https://pmc.ncbi.nlm.nih.gov/articles/PMC11751060/
3. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis - PMC, https://pmc.ncbi.nlm.nih.gov/articles/PMC12064692/
4. Impact of Clinical Decision Support Systems on Medical Students' Case-Solving Performance: Comparison Study with a Focus Group, https://mededu.jmir.org/2025/1/e55709
5. Opportunities, challenges and risks of using artificial intelligence for evidence synthesis - BMJ Evidence-Based Medicine, https://ebm.bmj.com/content/early/2025/01/09/bmjebm-2024-113320.full.pdf
6. (PDF) Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology – a recent scoping review - ResearchGate, https://www.researchgate.net/publication/378525063_Challenges_and_barriers_of_using_large_language_models_LLM_such_as_ChatGPT_for_diagnostic_medicine_with_a_focus_on_digital_pathology_-_a_recent_scoping_review
7. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines | Journal of the American Medical Informatics Association | Oxford Academic, https://academic.oup.com/jamia/article/32/4/605/7954485
8. Large Language Models in Medicine: Applications, Challenges, and Future Directions, https://pmc.ncbi.nlm.nih.gov/articles/PMC12163604/
9. (PDF) Developing Ethical AI Models in Healthcare: A U.S. Legal and Compliance Perspective on HIPAA and CCPA - ResearchGate, https://www.researchgate.net/publication/390208554_Developing_Ethical_AI_Models_in_Healthcare_A_US_Legal_and_Compliance_Perspective_on_HIPAA_and_CCPA
10. Prompt Engineering for AI Guide | Google Cloud, https://cloud.google.com/discover/what-is-prompt-engineering
11. aws.amazon.com, https://aws.amazon.com/what-is/prompt-engineering/#:~:text=Prompt%20engineering%20is%20the%20process,high%2Dquality%20and%20relevant%20output.
12. Effective AI Prompting Strategies for Healthcare Applications - Momentum, https://www.themomentum.ai/blog/effective-ai-prompting-strategies-for-healthcare-applications
13. Study: AI-generated medical advice not yet reliable enough to replace human experts, https://www.news-medical.net/news/20240402/Study-AI-generated-medical-advice-not-yet-reliable-enough-to-replace-human-experts.aspx
14. Prompt Versioning & Management Guide for Building AI Features - LaunchDarkly, https://launchdarkly.com/blog/prompt-versioning-and-management/

15. Iterative Prompt Refinement: Step-by-Step Guide - Ghost, https://latitude-blog.ghost.io/blog/iterative-prompt-refinement-step-by-step-guide/
16. Bias in AI - Chapman University, https://www.chapman.edu/ai/bias-in-ai.aspx
17. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare | The BMJ, https://www.bmj.com/content/388/bmj-2024-081554
18. Ethical-legal implications of AI-powered healthcare in critical perspective - Frontiers, https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1619463/full
19. Smart Pens in Diabetes Management - adces, https://www.adces.org/education/danatech/insulin-medicine-delivery/insulin-medicine-delivery-101/smart-pens-in-diabetes-management
20. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine - PMC, https://pmc.ncbi.nlm.nih.gov/articles/PMC10808088/
21. Advanced Prompt Engineering Techniques - Mercity AI, https://www.mercity.ai/blog-post/advanced-prompt-engineering-techniques
22. Chain of Thought Strategy for Smaller LLMs for Medical Reasoning - PubMed, https://pubmed.ncbi.nlm.nih.gov/40380574/
23. Self-Consistency Prompting - GeeksforGeeks, https://www.geeksforgeeks.org/artificial-intelligence/self-consistency-prompting/
24. What is Self-Consistency Prompting? - Digital Adoption, https://www.digital-adoption.com/self-consistency-prompting/
25. Hallucination Mitigation for Retrieval-Augmented Large Language Models: A Review - MDPI, https://www.mdpi.com/2227-7390/13/5/856
26. Retrieval augmented generation for large language models in healthcare: A systematic review - PMC, https://pmc.ncbi.nlm.nih.gov/articles/PMC12157099/
27. Prompt Engineering: Techniques, Uses & Advanced Approaches - Acorn Labs, https://www.acorn.io/resources/learning-center/prompt-engineering/
28. Top 5 LLM Prompts for Re-Writing your Technical Documentation - Scout, https://www.scoutos.com/blog/top-5-llm-prompts-for-re-writing-your-technical-documentation
29. Design Smarter Prompts and Boost Your LLM Output: Real Tricks from an AI Engineer's Toolbox | Towards Data Science, https://towardsdatascience.com/boost-your-llm-outputdesign-smarter-prompts-real-tricks-from-an-ai-engineers-toolbox/
30. What Is AI Prompt Security? - Palo Alto Networks, https://www.paloaltonetworks.com/cyberpedia/what-is-ai-prompt-security-
31. What is Iterative Prompting? A quick guide for Researchers using Generative AI - Indeemo, https://indeemo.com/blog/iterative-prompting-generative-ai
32. Prompt Versioning: Best Practices - Ghost, https://latitude-blog.ghost.io/blog/prompt-versioning-best-practices/
33. The Intersection of GDPR and AI and 6 Compliance Best Practices | Exabeam, https://www.exabeam.com/explainers/gdpr-compliance/the-intersection-of-gdpr-and-ai-and-6-compliance-best-practices/
34. AI Model Passport: Data and System Traceability Framework for Transparent AI in Health, https://arxiv.org/html/2506.22358v1
35. Enhancing LLM Generation with Knowledge Hypergraph for Evidence-Based Medicine, https://arxiv.org/html/2503.16530v1
36. Evaluating evidence-based health information from generative AI using a cross-sectional study with laypeople seeking screening information, https://pmc.ncbi.nlm.nih.gov/articles/PMC12149300/
37. The GRADE approach, Part 1: how to assess the certainty of the evidence - Medwave, https://www.medwave.cl/revisiones/metodinvestreport/8109.html?lang=en
38. -- GRADE - Systematic Reviews - CHSL LibGuides at Cleveland Health Sciences Library, https://chs.libguides.com/c.php?g=1149014&p=9263649

39. Overview of the GRADE approach, https://book.gradepro.org/guideline/overview-of-the-grade-approach
40. Chapter 7: GRADE Criteria Determining Certainty of Evidence - CDC, https://www.cdc.gov/acip-grade-handbook/hcp/chapter-7-grade-criteria-determining-certainty-of-evidence/index.html
41. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or - The BMJ, https://www.bmj.com/content/bmj/358/bmj.j4008.full.pdf
42. Assessing the Methodological Quality of Systematic Reviews - AMSTAR, https://amstar.ca/Amstar-2.php
43. LibGuides: Creating a PRISMA flow diagram: PRISMA 2020 - Subject Research Guides, https://guides.lib.unc.edu/prisma
44. What Is a PRISMA Flow Diagram - DistillerSR, https://www.distillersr.com/resources/systematic-literature-reviews/what-is-a-prisma-flow-diagram
45. The Role of External Validation in Artificial Intelligence Deployment: Strategies for Enhancing Performance in Real-World Clinical Settings | Simbo AI, https://www.simbo.ai/blog/the-role-of-external-validation-in-artificial-intelligence-deployment-strategies-for-enhancing-performance-in-real-world-clinical-settings-2073834/
46. Enhancing Clinical Data Validation with AI: A Comprehensive Approach to Accuracy and Efficiency, https://www.appliedclinicaltrialsonline.com/view/enhancing-clinical-data-validation-with-ai-a-comprehensive-approach-to-accuracy-and-efficiency
47. How do grade students with 100 % AI generated essays : r/Professors - Reddit, https://www.reddit.com/r/Professors/comments/1iknpm3/how_do_grade_students_with_100_ai_generated_essays/
48. Advice for students regarding Turnitin and AI writing detection - Academic integrity, https://academicintegrity.unimelb.edu.au/plagiarism-and-collusion/artificial-intelligence-tools-and-technologies/advice-for-students-regarding-turnitin-and-ai-writing-detection
49. Europe Smart Insulin Pens Market to Reach USD 223 Million by 2030 - MarkNtel, https://www.marknteladvisors.com/research-library/europe-smart-insulin-pens-market.html
50. Facilitators and Barriers to Smart Insulin Pen Use: A Mixed-Method Study of Multidisciplinary Stakeholders From Diabetes Teams in the United States | Request PDF - ResearchGate, https://www.researchgate.net/publication/363606261_Facilitators_and_Barriers_to_Smart_Insulin_Pen_Use_A_Mixed-Method_Study_of_Multidisciplinary_Stakeholders_From_Diabetes_Teams_in_the_United_States?_tp=eyJjb250ZXh0Ijp7InBhZ2UiOiJzY2llbnRpZmljQ29udHJpYnV0aW9ucyIsInByZXZpb3VzUGFnZSI6bnVsbH19
51. Advantages and disadvantages of connected insulin pens in diabetes management in, https://ec.bioscientifica.com/view/journals/ec/12/11/EC-23-0108.xml
52. Full article: The potential role of biosimilars in healthcare sustainability in Latin America, https://www.tandfonline.com/doi/full/10.1080/14712598.2025.2507173?src=
53. Biosimilar Drugs and the Hospital Formulary: A Canadian Experience - PMC, https://pmc.ncbi.nlm.nih.gov/articles/PMC6476576/
54. Full article: How do hospital characteristics and ties relate to the uptake of second-generation biosimilars? A longitudinal analysis of Portuguese NHS hospitals, 2015–2021 - Taylor & Francis Online, https://www.tandfonline.com/doi/full/10.1080/14737167.2023.2146579
55. A Look at the History of Biosimilar Adoption: Characteristics of Early and Late Adopters of Infliximab and Etanercept Biosimilars in Subregions of England, Scotland and Wales - A Mixed Methods Study - PubMed Central, https://pmc.ncbi.nlm.nih.gov/articles/PMC7803694/
56. The Impact of GDPR on AI Systems: Ensuring Ethical and Secure Data Processing in Healthcare | Simbo AI - Blogs, https://www.simbo.ai/blog/the-impact-of-gdpr-on-ai-systems-ensuring-ethical-and-secure-data-processing-in-healthcare-4013515/
57. AI Bias Mitigation: Detecting and Reducing Bias in AI Models - Sapien, https://www.sapien.io/blog/bias-in-ai-models-and-generative-systems

58. EU AI Act: first regulation on artificial intelligence | Topics - European Parliament, https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence
59. The EU AI Act's impact on medical devices - PharmaLex, https://www.pharmalex.com/thought-leadership/blogs/what-global-ai-regulations-mean-for-medical-device-manufacturers/
60. What Is a Prompt Injection Attack? - IBM, https://www.ibm.com/think/topics/prompt-injection
61. What Is a Prompt Injection Attack? [Examples & Prevention] - Palo Alto Networks, https://www.paloaltonetworks.com/cyberpedia/what-is-a-prompt-injection-attack
62. How Prompt Injection Works & 8 Ways to Prevent Attacks - Sprocket Security, https://www.sprocketsecurity.com/blog/prompt-injection
63. A Developer's Guide to Preventing Prompt Injection - Helicone, https://www.helicone.ai/blog/preventing-prompt-injection