Using Clinical Guidelines, domain ontology, and LLMs for Personalized Leukemia Treatment Recommendations

Xingru Xu[1], Michel Dumontier[1,2] and Chang Sun[1,2,*]

[1]Department of Advanced Computing Sciences, Faculty of Science and Engineering, Maastricht University, The Netherlands
[2]Institute of Data Science, Faculty of Science and Engineering, Maastricht University, The Netherlands

Abstract

Large Language Models (LLMs) offer new opportunities for clinical decision support, but face challenges in reliability, precise recommendations for individual patients, and adherence to medical guidelines. Challenges such as insufficient domain knowledge, generic outputs, and hallucination are risks to their clinical adoption. This paper proposes an approach that integrates LLMs with Clinical Practice Guidelines (CPGs) and medical ontologies to enhance personalized treatment recommendations. We compared four strategies to generate treatment recommendations with and without integrating clinical guidelines: (1) LLMs without any guideline input, (2) providing the full guideline document as textual input to LLMs with retrieval-augmented generation (RAG) technique; (3) converting guideline documents from PDF to markdown files capturing the structure of tables, diagrams, and references and using Chain-of-Thoughts to reason each decision steps; (4) structuring guidelines as graphs and linking medical concepts to ontologies as input to LLMs. We experimented on GPT-3.5 Turbo, GPT-4, and Llama 2. The evaluations assessed guideline adherence, treatment completeness, path alignment, and answer relevancy with Acute Lymphoblastic Leukemia as the primary use case. Additionally, we developed a user interface for health professionals to input patient descriptions and obtain treatment recommendations and explanations. Preliminary results demonstrate the feasibility of the graph-based approach in decision path tracing, graph-augmented reasoning, and natural language explanations to enhance transparency for clinician validation.

Keywords

Large Language Models, Clinical Practice Guidelines, Ontologies, Knowledge Graphs, Treatment Recommendation

1. Introduction

Large Language Models (LLMs) have been proposed as promising tools for enhancing clinical decision support systems with their ability to process vast amounts of medical literature, guidelines, and patient records to assist clinicians in diagnosis, treatment planning, and patient management. However, their adoption in clinical settings is hindered by unreliable outputs, lack of explainability, and risks of hallucination — where models may generate incorrect or misleading recommendations [1, 2]. To enhance the reliability of LLMs for clinical decision support, it is crucial to integrate authoritative, structured medical knowledge directly into their reasoning process. One approach is to provide LLMs with access to Clinical Practice Guidelines (CPGs). CPGs offer standardized, evidence-based treatment protocols that clinicians use to guide medical decisions [3].

However, current LLMs face challenges to accurately interpret and reliably apply CPGs. Unlike other medical knowledge sources, CPGs are often lengthy, complex, semi-structured documents that combine narrative text, cross-references, logic diagrams, workflow visualizations, tables, and conditional recommendations [4]. Simply converting CPGs to plain text and feeding it to an LLM can result in lost information, leading the model to miss key steps or generate wrong recommendations. Many recommendations in CPGs depend on "if-then" conditions, requiring logical reasoning to determine which pathway is relevant for a specific patient case. When CPGs are presented as raw text, LLMs struggle to navigate these dependencies and generalize without properly following the decision flow. Then, the majority of recommendations in CPGs are not stated in one place but instead refer to other

SeWebMeDa-2025: 8th International Workshop on Semantic Web solutions for large-scale biomedical data analytics, June 1, 2025, Portorož, Slovenia
*Corresponding author.
chang.sun@maastrichtuniversity.nl (C. Sun)

sections or external guidelines. LLMs often fail to resolve these references correctly. Furthermore, unlike other medical knowledge sources, CPGs require precise interpretation of dosages, contraindications, risk stratifications, and exception cases. LLMs, which rely on probabilistic text generation, may oversimplify nuanced recommendations or fail to capture clinically significant details. Finally, the lack of decision traceability in LLM-generated outputs poses a significant barrier to clinical adoption, as clinicians need transparent reasoning to validate the generated recommendations.

To address these challenges, we propose a framework that converts CPGs to graph structure and links them with relevant entities from medical ontologies. The graphs are fed into LLMs to improve the accuracy of treatment recommendations. Our approach centers on three innovations: (1) transforming decision diagrams in CPGs to graphs navigable by LLMs, formalizing clinical workflows for risk stratification, treatment staging, and decision logic; (2) integrating text, tables, and references into LLMs to ground recommendations; and (3) deploying domain knowledge from medical ontology to augment reasoning and minimize hallucinations. We designed four generation strategies: baseline model with structured prompting, RAG model, chain-of-thought model, and graph-based model. Three LLM models are experimented including GPT-3, GPT-4o, and Llama-2-70B. Our study focuses on Acute Lymphoblastic Leukemia (ALL), utilizing clinical guidelines from authoritative sources such as the National Comprehensive Cancer Network (NCCN). We evaluate the framework using synthetically generated patient datasets that mirror clinical scenarios, assessing performance through accuracy, guideline adherence, pathway alignment, and treatment completeness. We include explainability mechanisms, such as decision path tracing and rule-based reasoning, to enable clinicians to audit model outputs against CPGs. We demonstrate how structured guidelines and medical ontology integration enhance LLMs' capacity to deliver transparent and reliable treatment recommendations.

This paper is organized as follows: Section 2 reviews related work in medical LLM and guideline integration. Section 3 describes the proposed methodology and experimental architecture covering information retrieval and clinical guidelines integration. Section 4 details the experiment setting and evaluation methods. Section 5 discusses the results and implications. Finally, Section 6 summarizes the work and proposes future work.

2. Related Work

Translating CPGs into machine-actionable formats with integration with LLMs remains a significant challenge. Recent research has explored various approaches to address this issue, ranging from unstructured text-based methods to structured representations that enhance model reasoning. A straightforward approach is to incorporate guidelines as unstructured text, either through fine-tuning or real-time retrieval [5]. Fine-tuning can improve performance on relevant queries, but the model's knowledge is static, and it may hallucinate or misapply guidance. Retrieval-Augmented Generation (RAG) enables LLMs to fetch relevant guideline content dynamically from an external source, making it more adaptable to new and updated guidelines [6, 7]. However, it does not inherently enforce structured decision-making during clinical reasoning.

Beyond using raw text, researchers have explored structured representations of CPGs, such as decision trees, to enforce systematic adherence to medical guidelines[8, 9]. In this approach, LLMs function as reasoning engines, traversing a structured decision tree to make stepwise, logic-driven clinical decisions. Each node in the tree corresponds to a clinical decision point, guiding the model through a series of intermediate steps until it arrives at the recommended treatment. This method has been tested with multiple LLMs, including GPT-4, GPT-3.5, and PaLM-2, demonstrating improved alignment with correct treatment recommendations compared to zero-shot prompting. This highlights that hard-coding the guideline's decision logic can lead to more reliable outputs.

Another structured approach involves graph-based representations of CPGs, which capture complex relationships between medical concepts and decision pathways [9]. Guidelines are encoded as a graph, and the LLM selects a path through the graph that matches the patient's conditions. The LLM builds a reasoning path from patient data to a guideline-prescribed action, treating the guideline like a map of

connected decisions. The study also found that preserving the structure of guidelines (tables, flowcharts, condition-action pairs) is crucial. In summary, structured integration approaches treat CPGs not just as reference text but as algorithms or knowledge graphs that guide the model's reasoning.

3. Methodology

3.1. Guideline Extraction

We selected the National Comprehensive Cancer Network (NCCN) guidelines for Acute Lymphoblastic Leukemia (Version 1.2024) [10] as an external knowledge source for LLMs to provide treatment recommendations for different patients. The NCCN guidelines cover the entire patient management process from diagnosis to treatment to follow-up monitoring. The guidelines provide detailed diagnostic criteria, stratified risk assessment protocols, and treatment pathways for multiple patient subgroups.

The guideline contains various information - text, tables, references, and decision flowchart - accompanied by detailed supporting documentation. We first parse the guideline PDF files using Google Gemini 2.0 Flash Thinking, a large multi-modal language model with visual understanding capabilities and an extensive context window. The model was prompted to extract all textual explanations, tables, footnotes, references, and particularly the visual elements (e.g., decision flowcharts) to a semi-structured textual description in a markdown file.

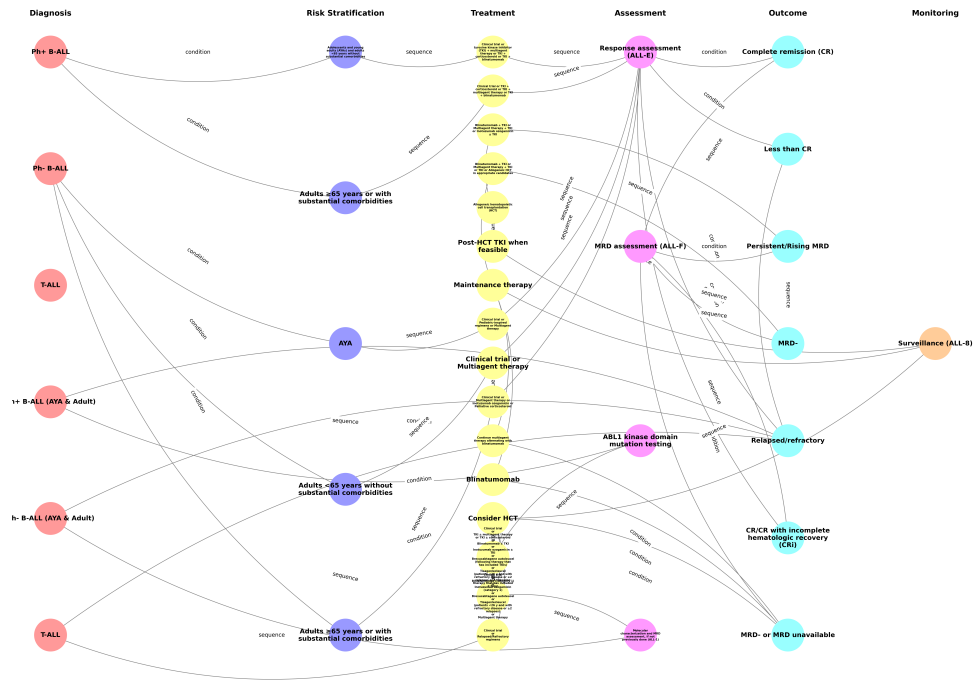

Figure 1: An example of a part of the graph structured from ALL Clinical Guideline

Furthermore, we utilized Anthropic's Sonnet-3.7-thinking model to transform the extracted components from the markdown file to a Directed Graph (DG) representation. An example of a part of the graph shown in Figure 1. In the graph, each node represents a decision point or treatment option in the clinical pathway, while each edge represents transitions between nodes with conditional logic. Properties include the patient characteristics and clinical variables that may influence decisions and references that are traceable to the source guidelines or articles. For the ALL case, we created three graphs for different ALL subtypes (Ph+ B-ALL, Ph- B-ALL, and T-ALL) to accurately reflect the different treatment pathways for each subtype. Finally, the graph structures were manually validated by the authors to ensure the accuracy of the extracted knowledge from the original guidelines, including

verifying the decision logic, checking the completeness of the treatment pathways and inclusion of the references and footnotes, and correcting the misinterpretation of the visual elements.

3.2. LLMs and Generation Strategies

We design and compare different strategies to retrieve from the CPGs and generate the recommendations, including a baseline LLM, RAG LLM model, RAG LLM with chain-of-thought reasoning, and graph-based RAG combined with LLM logical reasoning. We apply GPT-3.5 Turbo, GPT-4, and Llama-2-70B in the experiments.

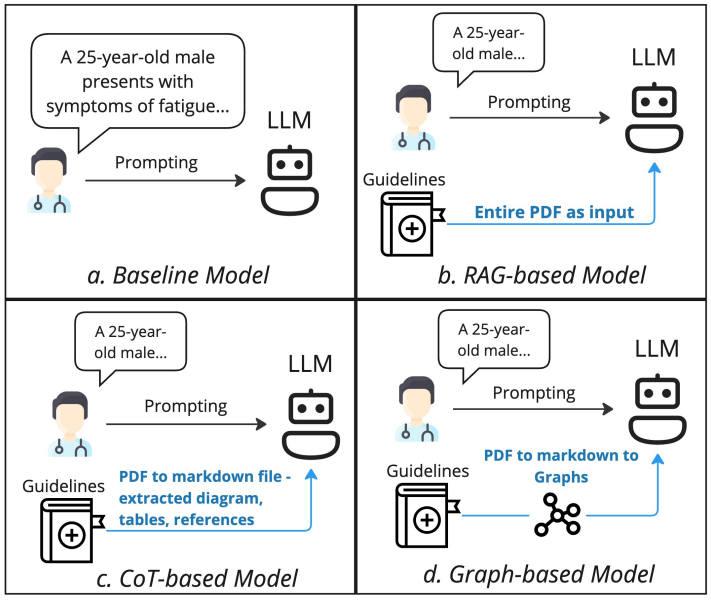

Figure 2: Four different strategies to integrate guidelines to LLM as input: a) baseline model without inputting any guideline; b) RAG-based model taking entire guideline in a PDF format as input; c) Chain-of-thought model extracting text, table, diagrams from the guideline file to a markdown file as input; d) Graph-based model converting the guideline to a graph and connecting medical terms with other ontology.

Baseline LLM Model The baseline model consists of an LLM without CPGs augmentation or retrieval mechanisms. The baseline model generates responses solely based on patient descriptions and its inherent pre-trained knowledge, showing its capabilities of providing accurate recommendations in the absence of external CPGs. We constructed a comprehensive prompting template for the model to extract key characteristics (e.g., age, ALL subtype, treatment history) from the patient description and shape the responses to follow the treatment pathway and consider the risk factors indicated in the guideline. The baseline model serves as a comparative standard to assess the impact of integrating guidelines in the following generation strategies. The following models use the same prompting template to prevent the potential influence of different promptings.

RAG-Based LLM Model The second model employs Retrieval-Augmented Generation (RAG) to retrieve relevant sections from CPGs during the response generation process. In this model, CPGs are injected directly as textual input, meaning that tables, references, decision pathways, and flowcharts are converted into plain text, resulting in the loss of decision logic and hierarchical structure. Compared to the baseline model, the RAG-based model can retrieve specific information from the guideline content based on patient description, which can help reduce hallucinations and improve the relevance and accuracy of the generated responses.

Chain-of-Thought (CoT) Enhanced LLM The CoT-enhanced LLM is based on the RAG-based model and converts flowcharts from CPGs into a graph representation and integrates stepwise reasoning. In this model, the LLM is provided with textual input from plain text, which was converted from a PDF version of the guideline to a markdown file. The tables, diagrams, and references are presented as text with their structures preserved. We provide a structured reasoning framework, including a multi-step decision process: 1) patient symptom analysis, 2) risk stratification, 3) treatment stage identification, and 4) personalized treatment recommendations. With the structured reasoning steps, CoT-enhanced LLM is enforced to follow the decision-making process defined in the CPG to reduce the possibility of irrational inferences or logical errors.

Graph-Guided LLM The fourth approach is Graph-Guided LLM, in which the guideline is transformed into a graph data structure, and the relevant medical concepts are linked with medical ontologies such as Human Phenotype Ontology [11] and Orphanet Rare Disease ontology [12]. The LLM extracts the defined key features and values from the patient description, such as age, comorbidity, and response to therapy, and constructs queries to retrieve relevant information from graphs. The decision path is guided based on the patient's conditions and retrieved outcome from the graph. Finally, the treatment recommendation is generated, including the extracted features from the patient, decision pathway selection, and how the path was navigated in each step.

3.3. Architectural Implementation

We implement the system with a user interface using Streamlit (shown in Figure 3). The interaction flow of the system was illustrated in Figure 4. The process begins when a user inputs a patient description in a chatbox. Then, the LLM will process and analyze the question and identify the patient characteristics that are required in the guideline. Then, the model will retrieve and query from the additional input, which is either the textual content or/and the graph constructed from CPGs, to find relevant treatment paths based on patient characteristics. For the graph-based models, the reasoning is generated for each decision step, including clinical rationale, supporting evidence from the graph, treatment decisions, and references to guidelines. On the UI, the user can choose the LLMs and generation strategies. In the generated responses, the extracted patient characteristics, selected pathway, pathway navigation, treatment recommendation, and the explanations of each step are displayed as results 3.

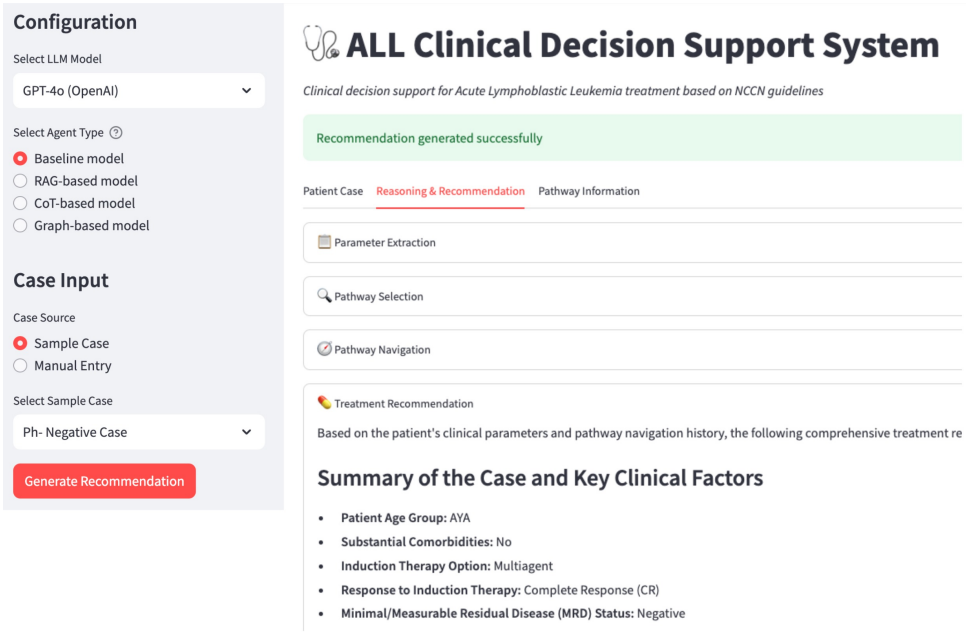

Figure 3: A screenshot of the user interface for the prototype of the ALL treatment recommendation tool.

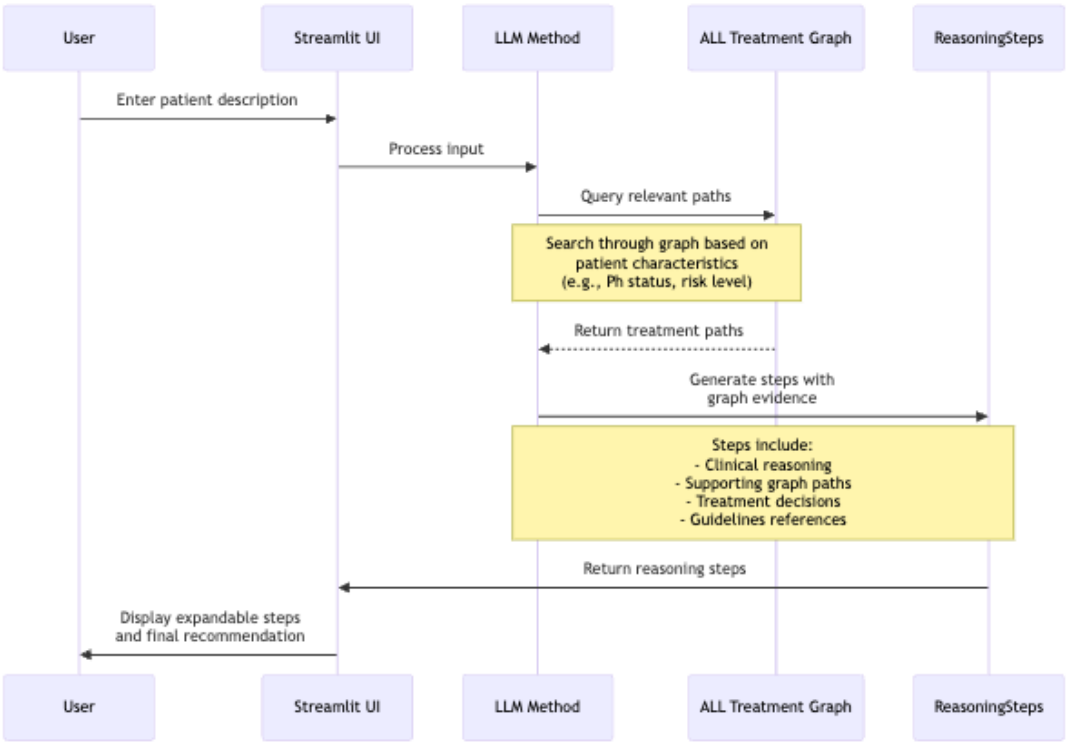

Figure 4: Sequence diagram illustrating the interaction flow between user input, LLM reasoning, and the ALL treatment graph in generating treatment recommendations using graph-based method. The system combines patient information with graph-based evidence to produce structured reasoning steps.

In practical clinical use, clinicians often consider multiple factors simultaneously, request additional information, or revise earlier decisions based on new data. We implement the system to these requirements by treating each clinical interaction as a series of events that can be processed, tracked, and audited independently. The system captures every decision and modification in the event log, storing a history of the decision-making process. The system can reconstruct the state of any clinical case so that the clinicians can examine how recommendations would have differed under various scenarios.

4. Experiments and Results

To evaluate different models, we generated synthetic data for 20 patients, including patient and condition descriptions, treatment pathways, and recommendations based on the guideline. We first extracted valid treatment paths from the guideline's graph data and then generated corresponding patient descriptions based on these paths using GPT-4o. The patient descriptions include patients' demographic information, medical history, clinical symptoms, and test results. Then, we generated treatment recommendations based on the diagnostic and treatment processes of each pathway as a reference for model output. The following shows an example of a patient description. The whole dataset is accessible at: https://github.com/MaastrichtU-IDS/guideline_graph_chatbot

"34-year-old male with newly diagnosed acute lymphoblastic leukemia. Flow cytometry shows B-lineage ALL (CD19+, CD20+, CD10+, CD34+, TdT+). FISH analysis confirms BCR::ABL1 fusion (p190 variant). No CNS involvement detected.Medical history notable for well-controlled hypertension on lisinopril 10mg daily. Performance status ECOG 1. Laboratory studies show normal liver and kidney function."

4.1. Evaluation Metrics

In our experiments, we used four evaluation metrics to measure the model's performance regarding guideline adherence, treatment completeness, path alignment, and answer relevance. All metrics are scaled to [0, 1], with higher values indicating better performance

Guideline Adherence (GA) This metric measures whether the generated responses from the model follow the specific treatment phases (induction, consolidation, maintenance, and surveillance) and are adherent to the terminology indicated in the CPGs. Let $G = \{g_1, g_2, ..., g_n\}$ represent the defined key terms mandated by the CPGs (such as TKI, tyrosine kinase inhibitor, transplantation), E represents expected responses (as true answers), and R denotes terms in the generated responses.

$$\text{GA} = \frac{1}{n}\sum_{i=1}^{n}\begin{cases}1 & \text{if } (g_i \in R \cap E) \vee (t \notin R \cup E)\\ 0.5 & \text{if } g_i \in R \setminus E\\ 0 & \text{if } g_i \in E \setminus R\end{cases} \tag{1}$$

Treatment Completeness (TC) This metric assesses whether the generated recommendations cover the complete treatment steps specified in the CPG, including initial treatment, follow-up monitoring, and potential follow-up treatment. This assessment is important to prevent the model from missing important treatment steps that could affect the quality of patient care.

$$\text{TC} = \frac{|R_{\text{step}} \cap G_{\text{step}}|}{|G_{\text{step}}|} \tag{2}$$

where R_{step} are generated treatment steps and G_{step} are required steps in the guideline.

Path Alignment (PA) This metric measures how closely the LLM's reasoning paths are consistent with the expected paths from the graph data structure from CPGs. We examine if the LLM agent follows the correct decision points (nodes in graphs) by calculating the longest common subsequence (LCS) of nodes between the LLM agent's path and the expected path as the measure of alignment. PA=1 indicates the LLM's reasoning path perfectly aligns with the expected guideline path, while PA=0 indicates no overlap between the two paths.

$$PA = \frac{LCS(R_{\text{path}}, E_{\text{path}})}{|E_{\text{path}}|}$$

Where $R_{\text{path}} = [r_1, r_2, \dots, r_m]$ denotes LLM's reasoning path, $E_{\text{path}} = [e_1, e_2, \dots, e_n]$ represents the expected guideline path. $LCS(R_{\text{path}}, E_{\text{path}})$ is the length of the longest common subsequence between R_{path} and E_{path}.

Answer Relevancy (AR) This metric measures how relevant the generated recommendations are to the given patient's case. A high score indicates that the generated recommendations can identify and address the specific characteristics of the patient, such as the ALL subtype, age, and risk factors.

$$\text{AR} = \frac{\sum_{i=1}^{n} \alpha \cdot \mathbb{I}(c_i \in R) + \lambda \cdot \mathbb{I}(v_i \in R)}{|C|} \tag{3}$$

where $C = \{c_1, c_2, ..., c_n\}$ denotes the key patient characteristics from the description, $V = \{v_1, v_2, ..., v_n\}$ represents the values of these characteristics. $\mathbb{I}(\cdot)$ is an indicator function (1 if true, 0 otherwise), and α and λ weight the characteristics terms and their values, respectively (default $\alpha = 1, \lambda = 1$).

4.2. Preliminary results

The performance results of using GPT-3.5, GPT-4, and Llama-2-70B in four different generation strategies are presented in the following tables. Among these, GPT-4 outperforms the other two models across all evaluation metrics. GPT-4 achieved baseline scores ranging from 0.462 to 0.513 across different metrics, compared to GPT-3.5 Turbo's range of 0.268 to 0.325 and Llama-2's range of 0.448 to 0.505. These results indicate that larger and more advanced language models may contain more medical knowledge and have better inherent capabilities in understanding and generating clinically relevant recommendations.

Table 1
Evaluation Results with GPT-3.5 Turbo

	Answer relevancy	Guideline Adherence	Treatment Completeness	Path Alignment
Baseline	0.325	0.284	0.268	N/A
RAG-based	0.412	0.388	0.365	N/A
CoT-based	0.567	0.539	0.553	0.574
Graph-based	0.629	0.612	0.621	0.642

Table 2
Evaluation Results with Llama-2

	Answer relevancy	Guideline Adherence	Treatment Completeness	Path Alignment
Baseline	0.505	0.464	0.448	N/A
RAG-based	0.562	0.538	0.525	N/A
CoT-based	0.627	0.609	0.613	0.624
Graph-based	0.669	0.652	0.661	0.672

Table 3
Evaluation Results with GPT-4

	Answer relevancy	Guideline Adherence	Treatment Completeness	Path Alignment
Baseline	0.513	0.475	0.462	N/A
RAG-based	0.580	0.549	0.541	N/A
CoT-based	0.642	0.631	0.628	0.639
Graph-based	0.690	0.662	0.685	0.698

All three LLMs demonstrated a consistent pattern of improvement when progressively advanced generation strategies were applied. The RAG-based approach outperformed the baseline model by showing the contribution of external guideline documents. Specifically, RAG-based generation improved GPT-4' Guideline Adherence from 0.475 to 0.549 and Treatment Completeness from 0.462 to 0.541. Similar trends were observed for Llama-2 and GPT-3.5 Turbo.

The CoT-based model resulted in further improvements by giving better reasoning and sequential decision-making, leading to increased Treatment Completeness and Path Alignment scores. For example, CoT-based generation yielded Path Alignment scores of 0.574 for GPT-3.5 Turbo, 0.624 for Llama-2, and 0.639 for GPT-4.

Among all generation strategies, the Graph-based approach consistently achieved the highest evaluation scores across all LLM models. GPT-4's Graph-based generation attained an Answer Relevancy of 0.690, Guideline Adherence of 0.662, Treatment Completeness of 0.685, and a Path Alignment score of 0.698. Similarly, Llama-2 achieved 0.669, 0.652, 0.661, and 0.672 on the same metrics, respectively. These results prove the efficacy of explicitly incorporating structured knowledge representations during the generation process.

5. Discussion

As observed from the experiments, the baseline model's performance showed limited capability in generating recommendations that adhered strictly to clinical guidelines. The incorporation of RAG improved performance by providing the models with access to relevant and unstructured guideline content. However, its effect was comparatively modest, likely due to the limitations of LLMs in processing lengthy and complex unstructured documents with heterogeneous input (text, table, figure, diagram, and references).

By giving structured input from guidelines, CoT model can produce consistent and adherent recommendations. By enforcing step-wise reasoning, CoT model resulted in significant gains in metrics such as Treatment Completeness and Path Alignment. These improvements demonstrate that reasoning-based prompting may lead to a better generation of clinical decision-making processes.

The Graph-based generation strategy outperformed all other methods, showing the advantages of representing clinical guidelines in structured graphs. Graph-based methods encourage models to generate not only complete and relevant but also strictly aligned with established treatment pathways defined in the guidelines. The improvements observed in Path Alignment scores, particularly in GPT-4, highlight its ability to maintain consistency and logical sequencing in complex treatment recommendations.

We observe a consistent trend across three models for generation strategies: Baseline < RAG < CoT < Graph-based. This trend indicates that the knowledge structuring strategies for LLMs to retrieve and reason are model-agnostic and can be generalized across LLM architectures.

6. Limitations and Future Work

This work can be improved in several aspects. For the test dataset in this study, we generated synthetic patient descriptions using GPT-4, which may not fully capture real-world patient scenarios. In future work, we plan to incorporate real patient cases from the MIMIC Clinical Database. The MIMIC database has clinical notes containing comprehensive patient information, including the brief hospital course, discharge summaries, prescriptions, patient illness histories, and treatment recommendations documented by clinicians. Integrating real patient descriptions and their actual clinical outcomes will enhance the reliability and diversity of the evaluation data and provide a more accurate assessment of model performance. For further validation, it will be valuable to ask clinicians to evaluate the generated recommendation and collect feedback from them.

Moreover, to generalize the proposed methods and enable their application to a broader range of diseases, a more flexible graph construction approach is required. Although the current method employs Gemini and other LLMs to automate the extraction process from guideline documents to graph structures, it still requires an amount of manual work. Specifically, manual corrections are often needed to adjust the extracted diagrams and validate their structures to ensure they accurately represent the decision logic of the original guidelines. Future work will focus on improving the methods to extract content and construct graphs in a more effective and accurate way. In addition, the current graph construction process is not fully integrated with disease and phenotype ontologies. At present, only medical entities are linked, and their definitions are added to the graph. Future efforts could aim to strengthen these connections to enhance richer semantic representation and interoperability.

Acknowledgement

This work has been supported by ICAI lab GENIUS (Generative Enhanced Next-Generation Intelligent Understanding Systems), a part of the NWO Long-Term Programme ROBUST initiated by the Innovation Centre for Artificial Intelligence (ICAI) and by REALM (Real-world data-enabled assessment for health regulatory decision-making), a project funded by Horizon Europe with grant number 101095435.

Declaration on Generative AI

The authors used GPT-4 for grammar and spelling checking. The author(s) reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] V. Liévin, C. E. Hother, A. G. Motzfeldt, O. Winther, Can large language models reason about medical questions?, Patterns 5 (2024).

[2] J. Ferdush, M. Begum, S. T. Hossain, Chatgpt and clinical decision support: scope, application, and limitations, Annals of Biomedical Engineering 52 (2024) 1119–1124.

[3] A. Qaseem, F. Forland, F. Macbeth, G. Ollenschläger, S. Phillips, P. van der Wees, B. of Trustees of the Guidelines International Network*, Guidelines international network: toward international standards for clinical practice guidelines, Annals of internal medicine 156 (2012) 525–531.

[4] D. Fast, L. C. Adams, F. Busch, C. Fallon, M. Huppertz, R. Siepmann, P. Prucker, N. Bayerl, D. Truhn, M. Makowski, et al., Autonomous medical evaluation for guideline adherence of large language models, NPJ Digital Medicine 7 (2024) 1–14.

[5] P. Lee, S. Bubeck, J. Petro, Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine, New England Journal of Medicine 388 (2023) 1233–1239.

[6] C. Zakka, R. Shad, A. Chaurasia, A. R. Dalal, J. L. Kim, M. Moor, R. Fong, C. Phillips, K. Alexander, E. Ashley, et al., Almanac—retrieval-augmented language models for clinical medicine, Nejm ai 1 (2024) AIoa2300068.

[7] C. Wang, J. Ong, C. Wang, H. Ong, R. Cheng, D. Ong, Potential for gpt technology to optimize future clinical decision-making using retrieval-augmented generation, Annals of biomedical engineering 52 (2024) 1115–1118.

[8] B. Li, T. Meng, X. Shi, J. Zhai, T. Ruan, Meddm: Llm-executable clinical guidance tree for clinical decision-making, arXiv preprint arXiv:2312.02441 (2023).

[9] D. Oniani, X. Wu, S. Visweswaran, S. Kapoor, S. Kooragayalu, K. Polanska, Y. Wang, Enhancing large language models for clinical decision support by incorporating clinical practice guidelines, in: 2024 IEEE 12th International Conference on Healthcare Informatics (ICHI), IEEE, 2024, pp. 694–702.

[10] National Comprehensive Cancer Network (NCCN), NCCN Clinical Practice Guidelines For Acute Lymphoblastic Leukemia, 2024. URL: https://www.nccn.org/guidelines/guidelines-detail?category=1&id=1410, accessed: 2024-03-08.

[11] F. Castellanos, J. Caufield, L. Chan, C. Chute, J. Cruz-Rojo, N. Dahan-Oliel, J. Davids, M. de Dieuleveult, V. de Souza, B. de Vries, et al., The human phenotype ontology in 2024: phenotypes around the world., Nucleic Acids Research 52 (2024).

[12] A. Rath, Annie Olry, Boulares Ouchenne, Caterina Lucano, David Lagorce, Marc Hanauer, Valérie Lanneau, Orphanet Rare Disease Ontology, 2023. URL: https://www.orphadata.com/data/ontologies/ordo/last_version/ORDO_en_4.4.owl.