

Pilot Assessment of Transparency of LLM-based Systems to Support Emergency Rooms

Michał Chojnicki^{1,*}, Katarzyna Kaczmarek-Majer^{2,3,*}, Paweł Burchardt^{4,5}, Yanwu Ren⁶ and Marek Z. Reformat^{6,7}

¹The Department of Medical Biology, Poznan University of Medical Sciences, Poland

²Systems Research Institute Polish Academy of Sciences, Newelska 6, 01-147 Warsaw, Poland

³University of Ostrava, Institute for Research and Applications of Fuzzy Modeling, 70103 Ostrava, Czech Republic

⁴Department of Cardiology, J. Struś Hospital, 61-285 Poznań, Poland

⁵Department of Hypertension, Angiology and Internal Medicine, Poznań University of Medical Sciences, 61-848 Poznań, Poland

⁶Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada

⁷University of Social Sciences, 90-113 Łódź, Poland

Abstract

One of the main challenges when developing medical decision support systems for the emergency room is adequately filtering the most relevant information. High workload, stress, and the necessity for urgent decisions require precise answers to the questions posed. Although LLM-based systems can provide abundant information, physicians need concise and relevant data in this particular clinical setting. In this study, we perform a pilot assessment of the transparency of selected LLM-based systems. The comparative analysis includes ChatGPT o1 model, which was asked to produce responses with varying temperatures and a pilot graph-based RAG specializing in cardiovascular diseases. A survey was conducted among 33 clinicians regarding the amount of information contained in the provided prompts. Physicians favored the most readable, specific, and helpful answers in emergency department conditions. Reliable medical data and the form in which answers are delivered are crucial for physicians working in the emergency room. We conclude that physicians have preferences for LLM responses at a specific temperature. Further research should be expanded to enable tailoring responses not only to the clinical situation but also to the experience of the asking physician.

Keywords

Trustworthiness assessment, Medical decision support systems, Large Language Models, Retrieval-Augmented Generation

1. Introduction

Attempts to implement Medical Decision Support Systems (MDSS) in emergency rooms based on text processing by Large Language Models (LLMs) have significantly intensified in recent years [1], [2], [3]. The work of a physician in emergency medicine requires that decisions be made particularly rapidly. Therefore, the decision support provided by LLMs must be substantive and precise. On the other hand, overly limited information delivery may necessitate additional queries directed to the LLM by medical personnel to obtain more comprehensive data. The motivation behind this study is to assess whether **variation in the temperature-related parameter of a large language model or a selection of model itself significantly affects the understanding of answers.**

Although various LLMs are able to deliver answers to medical questions including the control of their temperature, there is a need to perform advanced validation with the dedicated end users of such systems and to ensure that the deployed LLM-based systems are trustworthy. Trustworthiness requires that they will be technically accurate and follow appropriate ethical and legal principles. Following *Ethics Guidelines for Trustworthy AI* [4] the *EU Artificial Intelligence Act* [5], seven key requirements AI systems should meet to be deemed trustworthy: (i) human agency and oversight; (ii) technical robustness

EXPLIMED 2025 - Second Workshop on Explainable Artificial Intelligence for the medical domain - 25-30 October 2025, Bologna, Italy

*Corresponding author.

✉ michalc@ump.edu.pl (M. Chojnicki); k.kaczmarek@ibspan.waw.pl (K. Kaczmarek-Majer)

🆔 0000-0003-0422-9366 (K. Kaczmarek-Majer); 0000-0003-4783-0717 (M. Z. Reformat)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

and safety; (iii) privacy and data governance; (iv) transparency; (v) diversity, non-discrimination, and fairness; (vi) societal and environmental wellbeing; (vii) accountability. In this work, we focus on the transparency-related aspects of the AI-based systems and perform a pilot assessment across domain experts.

In this work, we propose a pilot version of the *MedicalGraphRAG* specialising in cardiovascular diseases, aiming to support emergency room clinicians (denoted as D in the remainder of this work), extending [6]. As baseline models, ChatGPT o1 with varying temperature [7] was also examined:

A. ChatGPT o1, temperature set to 0.2 in prompt;

B. ChatGPT o1, temperature set to 0.5 in prompt;

C. ChatGPT o1, temperature set to 0.8 in prompt.

The structure of the paper is as follows. In Section 2, we present the proposed pilot *MedicalGraphRAG* system to support emergency rooms. Section 3 describes the survey and its results. In Section 4, the main conclusions and future work are outlined.

2. The pilot MedicalGraph RAG in Cardiovascular Diseases

Our system transforms the medical literature into structured knowledge graphs to enable effective question answering with verifiable citations, i.e., references to the sections of the source information/documents. The process consists of three main phases: document processing and segmentation, knowledge graph construction, and the question-answering process supporting tracking of the information source. The system is designed to handle medical literature’s complexity and domain-specific nature while maintaining high accuracy and providing transparent source attribution.

The development of the *MedicalGraphRAG* system starts with the split of documents into fragments, the extraction of information, and the construction of knowledge graphs, as shown by the flow-chart in Figure 1.

2.1. Document Processing and Segmentation

Medical literature presents unique challenges for automated processing due to complex formatting, specialized terminology, and varying document structures. We address these challenges through a multi-stage processing pipeline that preserves semantic coherence while enabling efficient downstream processing.

2.1.1. Content Extraction

We extract content from PDF medical literature using MinerU [8, 9], an open-source tool specifically designed for precise document content extraction. Unlike traditional PDF parsers that often corrupt formatting and lose structural information, MinerU preserves document formatting while converting to Markdown, maintaining critical elements such as tables, figures, and hierarchical section structures that are essential for medical document comprehension.

2.1.2. Intelligent Text Segmentation

Given that medical documents typically exceed the token limits of current large language models (LLMs) and that LLM comprehension quality degrades significantly with excessively long contexts [10], intelligent text segmentation becomes crucial. We implement a relevance-based sliding window approach that maintains semantic coherence across segment boundaries.

Our segmentation algorithm operates as follows: starting with an empty window W , each new paragraph p_i is evaluated by an LLM using the prompt template:

"Given the current context [CONTEXT W], determine if the following paragraph is semantically related and should be included in the same section: [PARAGRAPH p_i]. Respond with 'INCLUDE' or 'NEW_SECTION'."

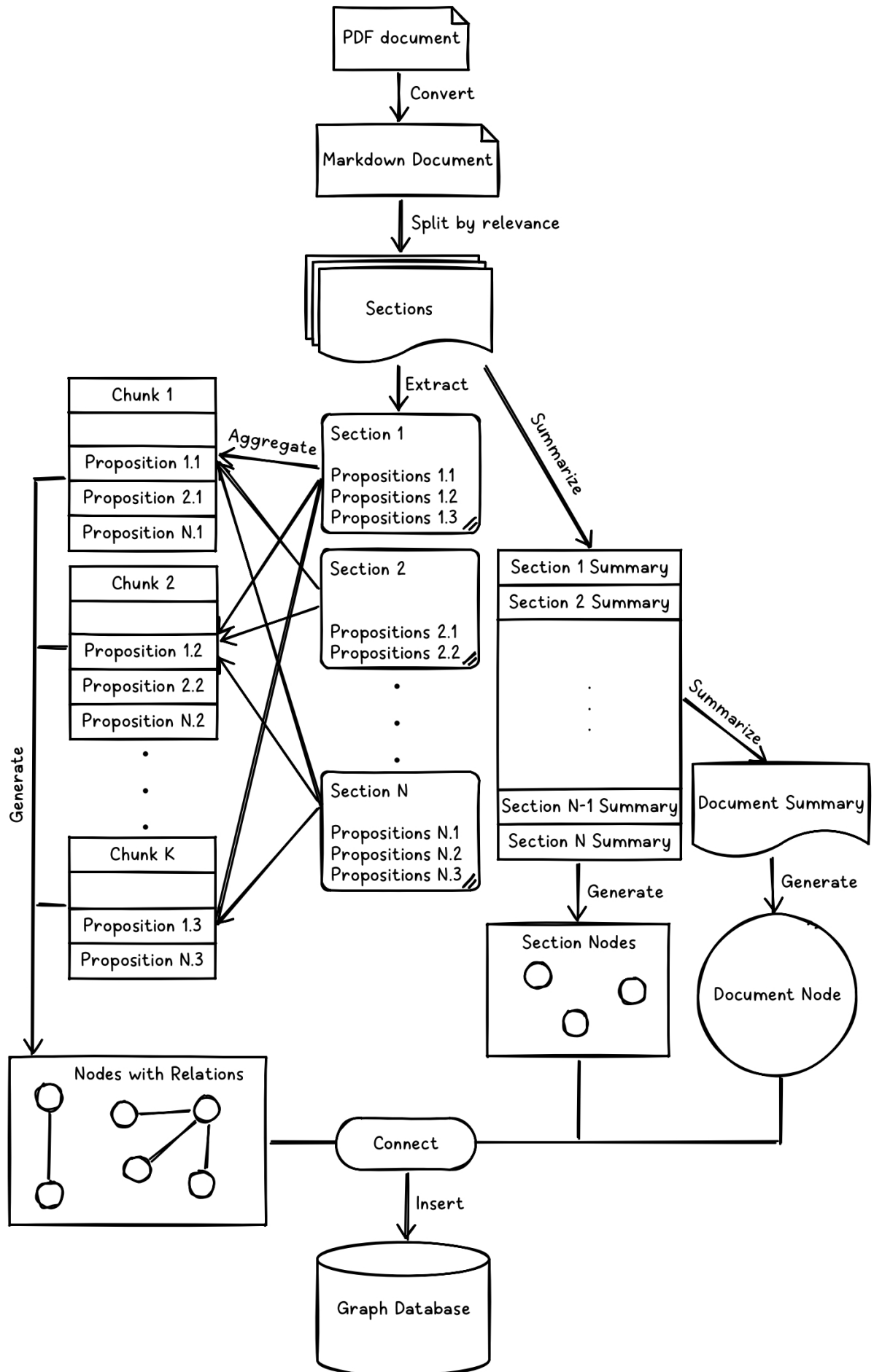


Figure 1: Flowchart for constructing the *MedicalGraphRAG*.

When the LLM determines that paragraph p_i is contextually relevant to the current window (section) content W , we update it $W \leftarrow W \cup \{p_i\}$. Otherwise, we finalize the current section $S_j = W$, reset the window (created a new section) $W = \{p_i\}$, and continue processing. This approach ensures that each segment maintains topical coherence while respecting token limitations.

2.2. Knowledge Graph Construction

Our knowledge graph construction operates at multiple semantic granularities to capture both fine-grained factual information and high-level conceptual relationships. The process creates a hierarchical knowledge representation that supports both detailed fact retrieval and contextual understanding.

2.2.1. Multi-Level Information Extraction

For each section S_i , we perform parallel extraction of two types of complementary information: *atomic propositions* and *summary of section*.

Atomic Proposition Extraction: Following the methodology of Chen et al. [11], we extract atomic propositions that represent the smallest meaningful factual units. Each proposition $prop_j$ is a simplified, self-contained statement that captures a single fact or relationship. For example, the complex sentence "This type of myocardial injury is characterized by myocyte necrosis and elevated troponin levels due to mechanisms other than myocardial ischemia and can be acute or chronic." is decomposed into propositions such as:

- $prop_1$: "This type of myocardial injury is characterized by myocyte necrosis."
- $prop_2$: "This type of myocardial injury is characterized by elevated troponin levels."
- $prop_3$: "Myocardial injury may be acute."
- $prop_4$: "Myocardial injury may be chronic."

Summarization: Simultaneously, each section undergoes structured summarization to extract high-level information relevant to potential medical queries. Our summarization prompt specifically targets key medical concepts including:

- Overall description
- Key medical methods
- Indications
- Mechanisms of action
- Efficacy and safety
- Advantages and limitations
- Clinical applications
- Patient outcomes
- Future directions

As a result, a summary of each section is created that contains information related to the abovementioned concepts.

2.2.2. Hierarchical Knowledge Organization

We create a three-level hierarchical structure to support different query types and retrieval granularities:

- **Document Level:** All section summaries for document D_k are aggregated using an LLM to create a comprehensive document-level summary node $DocNode_k$. This aggregation process identifies common themes, resolves potential contradictions, and creates a coherent overview of the entire document's content.
- **Section Level:** Each section S_i generates a corresponding section summary node $SecNode_i$ that captures the essential information within that specific section while maintaining links to the parent document node.
- **Proposition Level:** Atomic propositions extracted from each section are semantically grouped using our LLM-based clustering approach. Similar to the initial segmentation method, propositions are grouped based on semantic relevance, forming coherent chunks $C_j = \{prop_{j1}, prop_{j2}, \dots, prop_{jn}\}$.

2.2.3. Node and Relationship Generation

Within each proposition chunk C_j , we apply structured information extraction methods derived from the Camel framework ¹ to generate knowledge graph nodes and relationships. The process involves:

- Entity Recognition and Typing;
- Relationship Extraction;
- Node Deduplication.

2.2.4. Traceability and Provenance

A critical aspect of our system is maintaining bidirectional traceability between knowledge graph elements and source documents. Each generated node n_i stores:

- Source document identifier doc_id ;
- Section identifier sec_id ;
- Original text span.

This provenance information enables automatic citation generation and allows users to verify the reliability of extracted knowledge.

2.3. Question Answering with Citation Support

Our question-answering pipeline combines semantic retrieval with source attribution to provide accurate and verifiable answers to medical queries.

2.3.1. Query Processing and Restructuring

Medical queries often require domain-specific interpretation and may need restructuring to align with the knowledge graph's structure. We employ an LLM to transform user queries into a standardized medical inquiry format

¹<https://www.camel-ai.org>

2.3.2. Two-Stage Retrieval Process

To efficiently navigate the large-scale knowledge graph, we implement a coarse-to-fine retrieval strategy:

Document-Level Retrieval: The structured query is first matched against document-level summary nodes using semantic similarity. We identify the most relevant documents based on similarity score distribution.

Section-Level Refinement: Within the selected documents, we perform fine-grained matching against section-level nodes to identify the most relevant subsections.

This two-stage approach significantly reduces computational overhead while maintaining high recall.

2.3.3. Context Assembly and Answer Generation

Retrieved content is assembled into a structured context that includes relevant knowledge graph triplets. The assembled context is provided to a medical-domain fine-tuned LLM along with the original query to generate comprehensive answers.

2.3.4. Automatic Citation Generation

A key innovation of our system is the automatic generation of verifiable citations. Using the stored provenance information, we trace each piece of information used in answer generation back to its original source location.

3. Results

The purpose of the survey is to measure users' satisfaction with the answers provided by the selected LLM-based systems and check whether introducing explanations will add value to the system. Clinicians were asked to evaluate answers generated from ChatGPT o1 temperature set to 0.2 (model A), temperature set to 0.5 (model B), temperature set to 0.8 (model C) and answers by the pilot MedigalGraphRAQ (model D).

We asked each model (A-D) to answer the following question:

The patient has ST segment elevation at point J. When to diagnose acute coronary syndrome?

The respondents evaluated the four answers with the same criteria. Physicians evaluated the helpfulness, coherence, and presence of redundant information in the model responses.

3.1. Illustrative Example: End-to-end Answer Generation Process

The illustration of the answer generation process done by the *MedicalGraphRAG* is presented below.

First, the retriever compared all imported materials and returned the most relevant document, that is [12]. A second LLM pass ranked intra-document sections by semantic similarity to the query, after which the retriever collected the subject–predicate–object triples of the top passages.

(Acute coronary syndromes (ACS))–[PERTAINS_TO]–(ESC Guidelines 2023)
(Acute coronary syndromes (ACS))–[ASSOCIATED_WITH]–(ST-segment elevation)
(Acute coronary syndrome)–[SUGGESTS_DIAGNOSIS_OF]–(Acute chest discomfort)
(Acute coronary syndrome)–[PERFORM]–(12-lead ECG)
(Acute coronary syndrome)–[RECOMMENDED]–(Immediate invasive strategy)
(ST-segment elevation myocardial infarction (STEMI))–[RECOMMENDED]–(PPCI)
(Invasive strategy)–[RECOMMENDS]–(Acute coronary syndrome (ACS))
(ST-segment elevation myocardial infarction (STEMI))–[REQUIRES]–(Fibrinolysis)
(Acute coronary syndrome)–[REQUIRES]–(Electrocardiogram (ECG))
(Cardiac troponin)–[CHARACTERIZED_BY]–(Myocardial injury)

These triples were, in turn, supplied – together with the supporting passages – to LLM, prompting it to draft a clinical answer and put the document as references. The final generated answer is as follows:

Answer

ST-segment elevation at the J point in a patient is one of the key diagnostic criteria for the recognition of acute coronary syndrome, in particular ST-segment elevation myocardial infarction (STEMI). According to the European Society of Cardiology (ESC) guidelines, acute coronary syndrome can be diagnosed on the basis of electrocardiogram (ECG) changes, such as ST-segment elevation, together with elevated cardiac troponin levels indicating myocardial injury.

When ST-segment elevation is identified, an immediate invasive strategy is recommended, comprising primary percutaneous coronary intervention (PPCI) or reperfusion therapy, including fibrinolysis if PPCI is not available in a timely manner. The ESC guidelines emphasize the importance of rapid action to minimize myocardial damage and improve patient prognosis.

References: 2023 ESC Guidelines for the management of acute coronary syndromes

Translation based on: 2023 ESC Guidelines for the management of acute coronary syndromes (European Heart Journal, 2023 — doi.org/10.1093/eurheartj/ehad191).

Patients with acute coronary syndrome (ACS) may initially present with a wide variety of clinical symptoms, and it is important that both the public and healthcare professionals are aware of this..... [The remainder of the references are omitted].

3.2. Survey Results

We received responses from 33 physicians who completed their medical studies between 1979 and 2024. The study invitation explained that AI-based system would serve as a tool to support clinical decision-making. The purpose of the survey was to determine which responses, in their opinion, represented the best answers to the given question. The individual questionnaire included two questions:

1. *If such an answer was given in everyday work, how helpful is it?*
2. *Does it contain too much unnecessary information?*

Then, they were asked to answer and justify the following summary question:

- *Which of the above versions of the answer: A, B, C, or D - would be the best when working in the Hospital Emergency Department?*

Figures 2-3 present the main results. With the first survey question (*SQ1*), we aimed to assess whether the AI-generated answers would be helpful in everyday practice. To evaluate the comprehensiveness and usefulness of these answers in clinical decision-making, we employed a 10-point scale ranging from ‘not at all’ to ‘contains everything I need’. The vast majority of respondents selected the model with an intermediate temperature – 0.5, whose answers contained specific information regarding the diagnosis of an emergency condition. For the second question (*SQ2*), we asked respondents to evaluate the presence of unnecessary information. Model A had the lowest temperature setting and was rated as the most concise. Nevertheless, in each model, a significant percentage of physicians (ranging from 18% to 63%) indicated that the information provided was appropriate.

Figure 4 presents results for the third – summarizing – survey question (*SQ3*) about selecting the best answer, while Table 1 summarizes the reasons for such selection. Interestingly, most respondents selected model B (temperature 0.5). The remaining models were rated similarly (temperature 0.2, temperature 0.8, and *MedicalGraphRAG* system). Some individuals did not see a clear favourite (2 individuals). One person concluded that none of the answers met their expectations. Yet, when asked

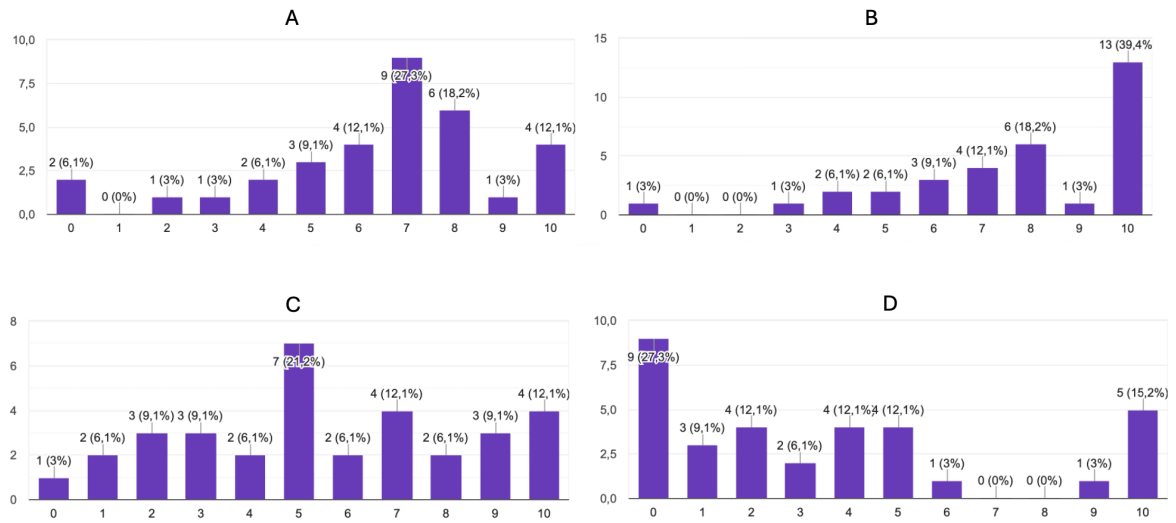


Figure 2: *SQ1*: "If such an answer was given in everyday work, how helpful is it?"

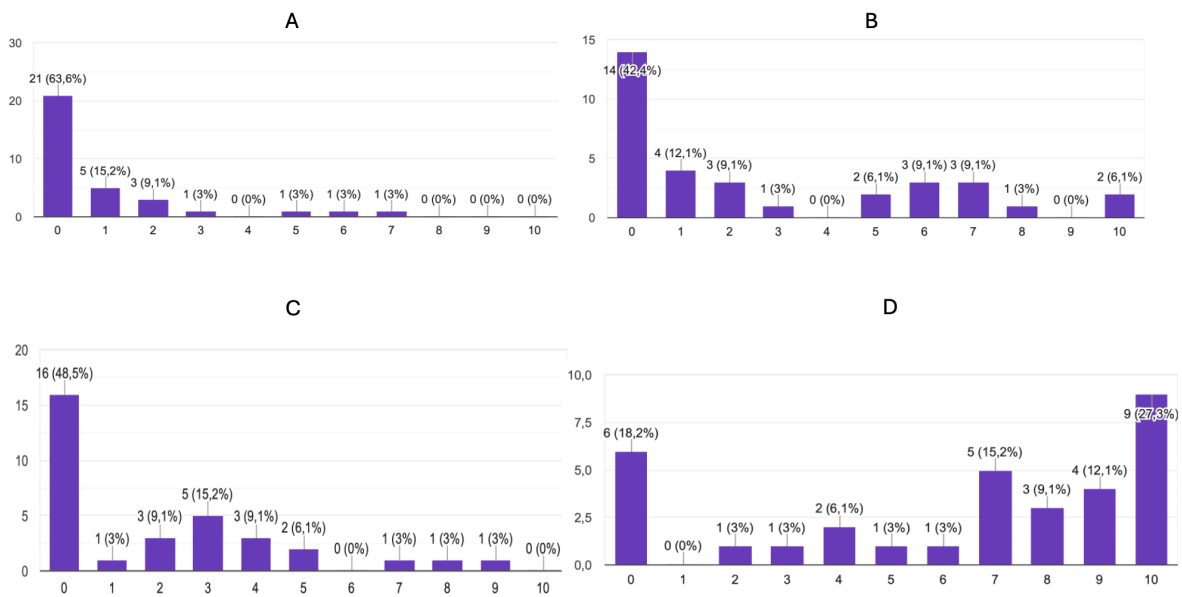


Figure 3: *SQ2*: "Does it contain too much unnecessary information?"

about the thoroughness of the answer – the overwhelming majority of participants pointed out the system *MedicalGraphRAG*, Table 1.

The present study indicates that, besides the essential substantive support provided by medical decision support systems, it is very important how the responses are presented. A group of physicians demands specific, concise responses from LLMs, while another group expects comprehensive answers that include an extensive clinical context.

This pilot study does not analyze all factors influencing the selection of responses at the current stage; however, it demonstrates that various forms of presentation are valued. These differences may

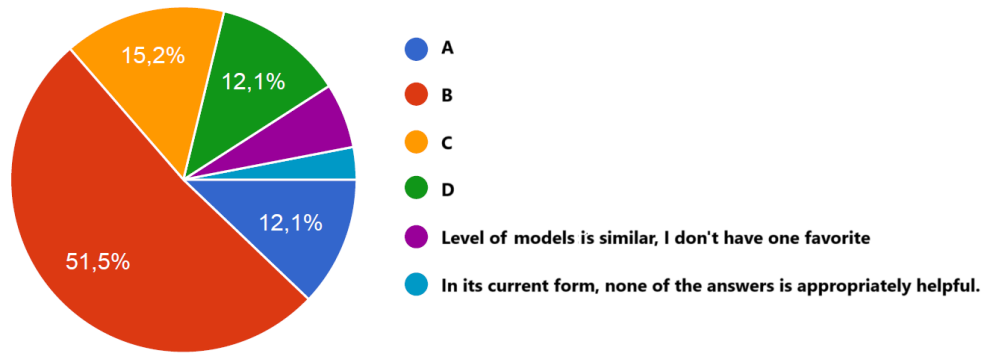


Figure 4: *SQ3*: "Which of the above versions of the answer: A, B, C, or D - would be the best when working in the Hospital Emergency Department?"

Table 1

Summary of why a particular answer for *SQ3* was selected

Reason	A	B	C	D
1. The most precise one, it contained everything I needed	33%	45%	33%	0%
2. The most comprehensive one, it explained all the aspects	17%	17%	17%	80%
3. Simple, clear message	50%	38%	50%	20%

stem from both professional experience and individual preferences.

Collaboration between physicians and decision-support systems, particularly regarding the level of detail in the provided content, should be tailored to the clinical context and individual users. One limitation of the present study is that it evaluates a single clinical scenario typical of an emergency department, assessed by physicians from various specialities. A cardiologist may consider one response valid if the scenario pertains to their field, whereas an otolaryngologist might prefer a different answer.

4. Conclusions and Future Work

Significant differences exist in the responses preferred by physicians from LLM systems depending on the model's temperature. Most physicians choose models whose responses are described as 'the most precise,' whereas a second group prefers comprehensive answers containing an extensive description of the suggested actions. Further research is needed to tailor responses to up-to-date medical knowledge in the clinical context and physicians' expectations. This study highlights the need for extensive validation of LLM-based systems, including their robustness, ethical aspects, transparency, and explicability.

Acknowledgments

This work is supported from the project „Research of Excellence on Digital Technologies and Wellbeing CZ.02.01.01/00/22_008/0004583” which is co-financed by the European Union. Katarzyna Kaczmarek-Majer is supported from the project "ExplainMe: Explainable Artificial Intelligence for Monitoring Acoustic Features extracted from Speech" (FENG.02.02-IP.05-0302/23) carried out within the First Team programme of the Foundation for Polish Science co-financed by the European Union under the European Funds for Smart Economy 2021-2027 (FENG). The authors thank all clinicians involved in the user study, helping with the evaluation of answers.

Declaration on Generative AI

During the preparation of this work, the author(s) used X-GPT-4 to generate samples for the survey. After using these tool(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

- [1] B. Glicksberg, P. Timsina, D. Patel, A. Sawant, A. Vaid, G. Raut, A. Charney, D. Apakama, B. Carr, R. Freeman, G. Nadkarni, E. Klang, Evaluating the accuracy of a state-of-the-art large language model for prediction of admissions from the emergency room, *J Am Med Inform Assoc* 1:31 (2024).
- [2] A. Shekhar, J. Kimbrell, A. Saharan, J. Stebel, E. Ashley, E. Abbott, Use of a large language model (llm) for ambulance dispatch and triage, *Am J Emerg Med* 89 (2025).
- [3] C. Y. K. Williams, T. Zack, B. Y. Miao, M. Sushil, M. Wang, A. E. Kornblith, A. J. Butte, Use of a large language model to assess clinical acuity of adults in the emergency department, *JAMA Network Open* 7 (2024) e248895–e248895.
- [4] European Commission and Directorate-General for Communications Networks, Content and Technology and Grupa ekspertów wysokiego szczebla ds. sztucznej inteligencji: Ethics guidelines for trustworthy AI, Publications Office, 2019.
- [5] European parliament and council: Regulation (eu) 2024/1689 laying down harmonised rules on artificial intelligence (artificial intelligence act), *Official Journal of the European Union*, 12 July 2024 (2024).
- [6] J. Wu, J. Zhu, Y. Qi, J. Chen, M. Xu, F. Menolascina, V. Grau, Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation, *arXiv preprint arXiv:2408.04187* (2024).
- [7] OpenAI, Introducing openai o1, Retrieved December 6, 2024 (2024).
- [8] B. Wang, C. Xu, X. Zhao, L. Ouyang, F. Wu, Z. Zhao, R. Xu, K. Liu, Y. Qu, F. Shang, B. Zhang, L. Wei, Z. Sui, W. Li, B. Shi, Y. Qiao, D. Lin, C. He, Mineru: An open-source solution for precise document content extraction, *arXiv preprint arXiv:2409.18839* (2024).
- [9] C. He, W. Li, Z. Jin, C. Xu, B. Wang, D. Lin, Opendatalab: Empowering general artificial intelligence with open datasets, *arXiv preprint arXiv:2407.13773* (2024).
- [10] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, P. Liang, Lost in the middle: How language models use long contexts, *Transactions of the Association for Computational Linguistics* 12 (2024) 157–173.
- [11] T. Chen, H. Wang, S. Chen, W. Yu, K. Ma, X. Zhao, H. Zhang, D. Yu, Dense x retrieval: What retrieval granularity should we use?, *arXiv preprint arXiv:2312.06648* (2023).
- [12] R. A. Byrne, X. Rossello, J. Coughlan, E. Barbato, C. Berry, A. Chieffo, M. J. Claeys, G.-A. Dan, M. R. Dweck, M. Galbraith, et al., Wytyczne esc 2023 dotyczące postępowania w ostrych zespołach wieńcowych, *Polish Heart Journal (Kardiologia Polska)* 81 (2023) 1–102.