

Modelling the Cost and Organisational Value of Language Model Systems in Primary Healthcare

Julia Trzaska

Wrocław University of Economics and Business

e-mail: julia.trzaska@ue.wroc.pl

ORCID: [0009-0003-2188-7384](https://orcid.org/0009-0003-2188-7384)

© 2026 Julia Trzaska

This work is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/4.0/>

Quote as: Trzaska, J. (2026). Modelling the Cost and Organisational Value of Language Model Systems in Primary Healthcare. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu*, 70(3), 16-27.

DOI: [10.15611/pn.2026.3.02](https://doi.org/10.15611/pn.2026.3.02)

JEL: I11, O33, D24

Abstract

Aim: This study aimed to examine the economic and organisational implications of a prototype system based on a large language model (LLM), designed to structure initial patient interview data in Polish primary healthcare.

Methodology: The exploratory study adopted a mixed-method approach, combining token consumption analysis across ten predefined patient scenarios with a semi-structured expert interview.

Results: The quantitative results demonstrated the stability and predictability of token usage, enabling preliminary modelling of computational costs under controlled conditions. The qualitative findings indicated potential organisational usefulness of the system, particularly in supporting information structuring prior to consultations.

Implications and recommendations: The findings suggest that language model-based solutions may support primary care processes in a cost-aware manner, provided that appropriate data protection measures and professional oversight are ensured. Further research should focus on real-world implementation, more diverse patient data, and extended system functionalities.

Originality/value: This study contributes to the literature by linking computational cost analysis with perceived organisational value in the context of primary healthcare, with a particular focus on the initial stage of the medical interview, which remains underexplored.

Keywords: large language models (LLM), token consumption, cost modelling, organisational decision support, primary health care

1. Introduction

The delegation of processes such as patient registration, appointment scheduling, and providing basic information has long been present in online environments as part of broader digitalisation trends across service sectors, including health care systems (Kellermann & Jones, 2013). In many service sectors the use of chatbots supporting users in obtaining information about pricing, locations or service rules has been steadily increasing (Adamopoulou & Moussiades, 2020). Solutions of this type are gradually becoming part of the organisational infrastructure, contributing to improved customer service and reducing the workload of administrative staff (Topol, 2019), however in primary healthcare the scope of such tools remains relatively limited. Although patients can often register for medical appointments electronically, the available solutions rarely extend beyond administrative functions. In practice, tools that support physicians in preparing for consultations are largely absent, particularly those aimed at the initial structuring of information provided by patients during the medical interview. At the same time, medical interviews in primary healthcare are conducted under significant time constraints, and the information communicated by patients is often unstructured and fragmented, which may delay the rapid identification of key symptoms and the efficient initiation of the diagnostic process. This raises the question of whether large language models could be used to preprocess patient-provided information and generate structured notes supporting the organisation of clinical information before the consultation takes place. From an organisational perspective the potential benefits of such solutions must be considered alongside their computational costs and the level of trust that can be placed in the generated information. This article contributes to the literature by examining the organisational usefulness and computational cost structure of language model-based systems designed to support the initial stage of the medical interview in Polish primary healthcare.

The aim of this study was to analyse the potential organisational usefulness and computational efficiency of a prototype system based on LLMs designed to support the initial stage of the medical interview in Polish primary healthcare. The author examined whether such systems can contribute to structuring patient-provided information before the consultation while maintaining predictable computational costs.

To address this objective, the study is guided by the following research questions:

RQ1: How stable and predictable is token consumption when generating structured medical notes using language models?

RQ2: How is the organisational usefulness of such notes perceived by a primary healthcare physician?

The remainder of the article is structured as follows. Section 2 reviews the literature on the use of artificial intelligence in healthcare organisations and the economic aspects of artificial intelligence (AI) implementation. Section 3 presents the research methodology. Section 4 reports the results of the computational efficiency analysis and expert evaluation. The final section discusses the findings and outlines implications for future research.

2. Literature Review

2.1. The Use of AI Tools in Improving Organizational Processes in Healthcare

The use of AI tools in healthcare organisations can take multiple forms and lead to diverse organisational implications. Such tools may contribute to the automation of selected processes (Jiang et al., 2017; Topol, 2019) as well as to decision support in handling complex and information-intensive tasks (Esteva et al., 2017; Kermayn et al., 2018). Beyond these functional aspects, the role of information itself becomes central, as AI-based solutions rely on the collection, processing, and structuring of data (Raghupathi & Raghupathi, 2014). From an organisational perspective, treating

information as a key resource is therefore an important factor in improving operations within health care settings. Its importance is further emphasised by the highly confidential nature of medical information, which requires that all solutions be adapted to the associated risks and responsibilities related to data collection and processing.

One of the most visible applications of AI in healthcare is the automation of selected processes, which provides a natural starting point for further analysis. Automation refers to the application of technologies aimed at improving control, predictability, and efficiency in complex organisational and economic systems. It supports organisations in maintaining competitiveness by enabling reliable, safe, and predictable operation of processes within optimal performance ranges (Mohapatra et al., 2023). Building on this concept, tools based on artificial intelligence, including solutions designed with language models, can be understood as an extension of automation technologies that support organisational activities through the processing and structuring of information. Language models enable the automation of text-based tasks, e.g. organising and summarising unstructured information, which are especially relevant in healthcare contexts. AI-based automation advances the knowledge obtained by computational analytics and enables machine-based decision-making (Sarker, 2022) and is increasingly applied to support decision-making and process optimisation across sectors (İşgüzar et al., 2024), including healthcare, yet its value scope extends beyond corporations, since AI has previously shown promise in the healthcare sector, where efficiency and effective information management are crucial. For instance, AI has proved to be valuable in supporting diagnostic processes, particularly in medical imaging (Poon et al., 2025).

In healthcare settings, this automation is closely linked with decision-making processes, as AI-based systems increasingly support professionals in tasks that require the analysis and interpretation of complex data. Rather than replacing human judgement, such systems are designed to assist decision-making by providing structured outputs and highlighting relevant information, thereby supporting more efficient and informed clinical and organisational decisions. In addition to supporting classification these systems aim to enhance transparency by indicating image regions that influence algorithmic outputs, demonstrating how AI tools can assist diagnostic workflows and clinical decision-making without replacing professional judgement (Kermany et al., 2018). Beyond imaging-based applications, research shows that AI methods based on natural language processing have been increasingly applied to the extraction of cancer-related concepts and symptoms from textual data. Rule-based approaches have been commonly used due to their effectiveness in accurately identifying and extracting relevant clinical information (Gholipour et al., 2023).

Recent developments in generative artificial intelligence have further expanded the application of language-based models beyond information extraction towards automated clinical documentation. Emerging solutions, often referred to as AI medical scribes or ambient clinical documentation systems, are designed to transcribe and summarise patient-physician interactions in order to reduce documentation burden and support clinical workflows, while emerging empirical evidence has begun to assess the quality and reliability of AI-generated clinical notes produced from such interactions (Sasseville et al., 2025). These developments demonstrate the technical likelihood of generative models in clinical documentation contexts. However, they primarily focus on post-consultation summarisation rather than on structuring patient-reported information prior to the clinical encounter. Moreover, the growing interest in AI-based solutions has also been reflected in recent Polish literature, which highlights the increasing adoption of artificial intelligence and other digital technologies in healthcare and family medicine, emphasising their potential to improve organisational efficiency while also recognising the ethical, legal, and implementation challenges associated with their responsible use (Możarowski et al., 2025; Suska, 2025).

Despite the growing body of research on AI applications in healthcare, several organisational and economic questions remain insufficiently explored. The existing literature predominantly focused on imaging-based diagnostics, assisted decision-making, and natural language processing aimed at information extraction. These approaches largely concentrate on classification, prediction, or

structured data extraction rather than on the generative structuring of narratives reported by patients. While recent studies have begun to examine AI-generated medical documentation, particularly in the context of ambient or AI-supported clinical note systems, these applications mainly address post-visit documentation. Considerably less attention has been devoted to the use of generative language models at the initial stage of the medical encounter, especially as a tool for structuring patient-reported information before the consultation takes place, while the organisational cost perspective of such solutions, particularly within the context of Polish primary healthcare, remains underexamined. This article addresses this research gap by analysing a prototype system designed to support the initial medical interview in Polish primary healthcare and by modelling its token consumption alongside its perceived organisational value through expert evaluation.

2.2. Computational Efficiency and Implementation Costs of AI-Based Solutions

The successful implementation of AI solutions in healthcare organisations requires a broader perspective than only assessing their technical capabilities, and is equally important to consider the costs of implementation and maintenance. Such systems require additional resources, including technical infrastructure, system integration, personnel training, testing, and continuous monitoring to ensure reliability and safety in clinical environments (Chui et al., 2023; Tabassi, 2023).

Large language models introduce a distinct cost structure compared to traditional business systems. Rather than relying on fixed licensing fees or one-time software purchases, their operational cost is typically based on token usage, including both input and output tokens processed by the model (OpenAI Developers, n.d.; Zhao et al., 2026). Tokens represent units of text, meaning that the length and structure of user input, as well as the generated output, directly influence the overall cost of using the system. As a result the cost of deploying language model-based solutions is not fixed but varies depending on usage patterns. In healthcare contexts this variability is particularly significant, as patient-provided information may differ substantially in length, complexity, and structure across consultation scenarios. Thus, total system cost may fluctuate depending on the number of users, interactions, and processed tokens, which directly affects the scalability of such solutions (Bommasani et al., 2022).

In healthcare settings this variability becomes particularly noticeable as it depends not only on the number of patients served, but also on differences in how patients communicate their symptoms. The length, structure, and level of detail in patient narratives may vary significantly, as different conditions often require different levels of explanation. Consequently, estimating average token usage becomes challenging even within a single primary care facility. This challenge may be further increased when voice-based input is introduced. While speech-to-text interaction can improve accessibility, particularly for older patients or those less familiar with digital interfaces, it may also increase the volume of generated text and, consequently, token consumption, leading to higher operational costs (OpenAI Developers, n.d.), directly affecting both the overall cost and the difficulty of predicting system usage.

In addition, the total number of patients assisted is closely related to the size and capacity of a given healthcare facility. A higher number of consultations leads to increased token usage and, therefore, higher operational costs. This issue is particularly relevant in healthcare systems with limited financial resources and budget constraints (OECD & European Observatory on Health Systems and Policies, 2024), where the lack of cost predictability may constitute a significant barrier to the adoption and scaling of AI-based solutions. The variability introduces challenges related to cost estimation and economic predictability. From an organisational perspective the ability to anticipate operational costs becomes an important factor in decision-making regarding the adoption and scaling of AI-based systems (Davenport & Ronanki, 2018), hence this study focused on computational efficiency by analysing token consumption patterns and providing preliminary cost projections based on controlled input scenarios.

3. Methodology

3.1. Research Design

The author employed a mixed-method design combining an exploratory and pilot case study approach, conducted within the context of Polish primary healthcare. The focus was on developing a prototype system based on a large language model, designed to generate structured summaries of patients' initial symptoms prior to a medical consultation. The intended function of the prototype system was to support physicians by providing a brief pre-consultation overview of the case, potentially reorganising the initial phase of the patient encounter. Rather than replacing the medical interview, the system was designed to structure the first layer of information provided by the patient. The prototype system was developed in Python using a Jupyter Notebook environment and integrated with the OpenAI GPT-4o-mini model via its Application Programming Interface (API). It was tested on ten hypothetical patient scenarios reflecting common primary care complaints, such as headaches or sore throat. The scenarios were discussed with an experienced primary healthcare physician to ensure their relevance to everyday clinical practice. The methodological framework consisted of two complementary components. First, a computational cost analysis was conducted to examine token distribution and cost predictability under controlled conditions. Second, a semi-structured expert interview was carried out to assess the perceived practical usefulness and organisational feasibility of the solution in the context of Polish primary healthcare. The combination of quantitative and qualitative methods was intentionally adopted to capture both the economic dimension of the solution and its perceived organisational usefulness. While token analysis provided measurable and comparable indicators, the expert interview allowed for contextual interpretation within real clinical practice. Such a mixed approach was considered necessary to avoid evaluating the prototype purely from a technical perspective.

3.2. Computational Efficiency Analysis

The computational efficiency analysis was conducted using ten predefined hypothetical patient scenarios reflecting common primary healthcare complaints. Each scenario included a short textual description of initial symptoms, duration of complaints, and, where applicable, information on chronic conditions or medication use. All scenarios were formulated in Polish, reflecting the linguistic context of Polish primary healthcare. The content of the scenarios was consulted with an experienced Polish primary healthcare physician to ensure their relevance to everyday clinical practice. The experimental procedure consisted of processing every scenario through the prototype system, developed in Python within a Jupyter Notebook environment and integrated with the OpenAI GPT-4o-mini model via API. For each case a standardised system prompt and identical structural framework were applied to ensure consistency across scenarios. To maintain analytical control and reproducibility, the model operated under deterministic settings, i.e. the temperature parameter was set to 0, which reduces randomness in generated responses. This configuration minimised variability in outputs and enabled controlled comparison of token consumption across the cases. All the responses were automatically recorded in a structured dataset, including total token count, input tokens, and output tokens for each API call. The collected data were exported to a spreadsheet format for further statistical analysis. The primary variable of interest was token consumption, treated as a proxy for computational cost. Descriptive statistics were calculated to assess variability, central tendency, and predictability of token usage through scenarios. Given the limited number of scenarios and controlled experimental conditions, the analysis should be interpreted as exploratory and pilot in nature. This approach allowed for modelling the cost distribution under defined input conditions and provided a basis for the economic interpretation presented in Section 4.1.

3.3. Expert Evaluation of Generated Notes

The qualitative component of the study involved an expert evaluation of the generated notes. The interview was conducted with a family medicine physician with over 20 years of professional experience in Polish primary healthcare, who participated anonymously. The interview was carried out in Polish and followed a semi-structured format, and was conducted by the author, audio-recorded, and accompanied by simultaneous notetaking. Next, a full transcription of the interview was prepared for analytical purposes. The collected material was analysed thematically, with particular attention given to recurring patterns and emerging themes related to the organisational usefulness of the prototype system. During the interview the expert was presented with the results in a structured format, including tables displaying the original patient scenarios and the corresponding notes generated by the prototype system, allowing to assess the outputs in direct relation to the provided inputs. The primary objective of the interview was to evaluate the perceived organisational usefulness of the system within Polish primary healthcare settings. The discussion addressed potential errors, limitations, practical applicability, and the conditions under which such a solution could be realistically implemented. The expert was selected based on accessibility and professional experience in primary care.

4. Results

This section presents the findings of both the quantitative and qualitative components of the study and addresses the research questions formulated in the introduction. The results are organised into two parts: the analysis of token consumption and the expert evaluation of the generated notes.

4.1. Token Consumption Analysis

This subsection presents the results of the experiment and analyses the performance of the diagnostic support prototype, addresses the computational efficiency of the model. All the data were derived from ten predefined patient scenarios. In addition to qualitative assessment, token usage was monitored to relate the generated content to computational costs.

The first step of the analysis involved examining computational efficiency measured by the number of tokens used during note generation. Token consumption constitutes a relevant indicator as commercial LLM services are priced per processed input and output token, where one token represents a unit of text processed by the model (OpenAI Developers, n.d.). This pricing structure directly links token usage with operational expenditure and scalability considerations.

As presented in Table 1, the average total token usage per case amounted to 190.6, including 139.9 input tokens and 50.7 output tokens. The observed standard deviation was relatively low, indicating stable and consistent system behaviour across scenarios. The minimum total token usage was 181 tokens, while the maximum reached 200 tokens, suggesting that even more extensive scenarios remained within a narrow range. The quartile distribution further confirmed this consistency. For the total token usage, the lower quartile equalled 187.25, the median 188.5, and the upper quartile 195.25. Similar patterns were observed for input and output tokens, with most of the cases clustering around comparable values, which indicates that the system generated responses of relatively similar length across different scenarios.

Based on the average number of input and output tokens and the pricing model of GPT-4o-mini (0.15 USD per 1 million input tokens and 0.60 USD per 1 million output tokens), the estimated cost of generating a single note amounted to approximately 0.000051 USD. For all ten analysed cases, the total computational cost was approximately 0.000514 USD. From an organisational perspective the results suggested a predictable cost structure associated with note generation. The relatively narrow

range of token usage enabled cost estimation in advance, which may support planning decisions in potential implementation contexts. These estimates represent only the operational API costs associated with token consumption and should not be interpreted as the total cost of implementing such a system in clinical practice, which would also include software development, system integration, infrastructure, maintenance, and personnel training.

Table 1. Descriptive statistics of token usage across scenarios

Statistic(s)	Total tokens	Input tokens	Output tokens
Mean	190.6	139.9	50.7
Standard deviation	5.76	2.73	3.53
Minimum	181	135	46
25th percentile	187.25	138.25	48.25
Median	188.5	140	50
75th percentile	195.25	142	53.5
Maximum	200	143	57

Source: author's own calculations.

To contextualise the token-based cost within Polish primary healthcare practice, workload scenarios were demonstrated, as Polish regulations do not define a maximum number of daily consultations (Narodowy Fundusz Zdrowia, 2013).

To illustrate the potential financial implications, three workload scenarios were modelled: 20, 30 and 40 consultations per working day for one physician. Under a workload of 20 consultations per day, the estimated daily cost would amount to approximately 0.00102 USD, corresponding to 0.02056 USD per month and 0.24672 USD per year. For 30 consultations per day, the daily cost would reach approximately 0.00154 USD, resulting in 0.03084 USD per month and 0.37008 USD annually. In the highest analysed scenario of 40 consultations per day, the estimated daily cost would amount to approximately 0.00206 USD, corresponding to 0.04112 USD per month and 0.49344 USD per year. Although the estimated costs appear marginal under controlled prototype conditions, real-world implementation would likely involve longer and less structured patient narratives, potentially increasing token consumption. The integration of additional components, such as speech-to-text processing, could further elevate computational expenditure.

Overall the results indicated that token consumption remained relatively stable across the analysed scenarios, suggesting a predictable computational cost under controlled conditions. This provided the basis for estimating operational expenses and assessing the scalability of the proposed solution.

4.2. Themes Emerging from the Expert Evaluation

This subsection presents the results of the qualitative analysis based on the expert evaluation of the generated notes. The aim of this part was to explore the perceived organisational usefulness of the proposed system and to identify the key themes emerging from the physician's perspective.

A recurring theme in the interview concerned the conditional usefulness of the proposed system depending on patient characteristics, particularly the age group of patients. The expert clearly distinguished between younger and older patients, emphasising that the structure and dynamics of the medical interview differ significantly between these groups. According to the physician, younger patients tend to communicate more efficiently and focus directly on the core problem, often presenting clear expectations regarding the purpose of the visit. As a result, consultations with younger individuals typically follow a more structured and goal-orientated trajectory. In contrast, consultations

with elderly patients were described as more narrative and expansive. The expert noted that older individuals frequently provided extensive contextual information, including peripheral details not always directly relevant to the diagnostic process. This broader storytelling approach may introduce a lack of structure and prolong the consultation, requiring additional effort from the physician to reorganise and prioritise the key clinical elements. In this context the potential time saved during the initial phase of the patient interview could differ between the two age groups and appeared to be more significant in consultations involving elderly patients, where scattered information tends to prolong the interaction.

The value of prior access to structured information before the patient enters the consultation room emerged as another important theme. According to the expert, even a brief pause to review the note would allow the physician to prepare more effectively for the visit. One implication concerned the possibility of acting more proactively in cases involving potentially contagious diseases. If the physician was aware in advance that a patient may present symptoms of an infectious condition, they could respond immediately, for instance, by providing a mask in the waiting area if the patient was not wearing one and adjusting the consultation process accordingly. A shorter and more focused interview in such cases may also reduce the time spent in close contact, thereby limiting the potential spread of infection within the appointment room.

Moreover, the physician indicated that prior access to information would increase professional comfort and provide greater peace of mind. Knowing what kind of case is approaching allows the doctor to focus on the most relevant aspects of the consultation from the very beginning, without dispersing attention across less essential topics.

Another important theme concerned the role of the system as support rather than a replacement for the medical interview. The physician emphasised that the tool could serve as a means of structuring the initial key information provided by the patient, thereby making the consultation process easier to follow, and at the same time being certain that further diagnostic questioning would still be necessary, as additional issues often emerge naturally during the visit, including during the physical examination phase. The physician stressed that the interview remains a crucial part of the appointment and that the doctor must continue to take an active role by asking follow-up questions as each case is different and there is no universal set of questions applicable to every patient. For this reason, the possible absence of certain details in the AI-generated notes would not be considered problematic according to the physician, who did not expect the system to capture everything, acknowledging that this was unrealistic. Instead, the note was perceived as a tool for streamlining repetitive or routine elements of the interview and providing a structured starting point for further assessment, clearly distinguished from the physician's diagnostic expertise. Ultimately, the expert expressed trust in the system, as it was seen not as altering professional judgement but as supplementing it and offering a useful baseline for the consultation.

An additional recurring topic concerned the time required for the physician to study the generated notes. When asked about the preferred moment for reviewing such information, the expert clearly stated that it could not become part of a general morning preparation routine covering all patients scheduled for the day. Instead, the non-negotiable moment would be immediately before each individual consultation, just prior to the patient entering the consultation room. The physician indicated that even one minute would be sufficient to review the note. However, this short but dedicated time slot was described as essential for ensuring that the system could function effectively within the existing workflow. Without such a structured moment of preparation the potential organisational value of the tool would be significantly reduced.

Another inflexible condition concerned the protection of sensitive medical information. The physician highlighted the necessity of providing a closed and private space for patient interaction with the system, as the information shared falls under medical confidentiality and is subject to personal data protection regulations.

The final theme extended beyond the physician's own workflow and concerned the patient's initial contact with the system. The expert emphasised the need for trained medical staff, potentially nurses working at the reception desk, to support patients during the first weeks or months of implementation, with such assistance being particularly important while patients become familiar with the tool. The process was perceived as likely to be more intuitive for younger individuals, whereas some elderly patients might require additional professional support, particularly in cases where understanding of digital tools is limited.

To summarise, the qualitative findings indicated that the generated notes may support the organisation of patient information and improve the efficiency of the initial stage of the consultation process. The system was perceived as a supportive tool that enhances, rather than replaces, the physician's role. In relation to the research questions, the quantitative results addressed RQ1 by demonstrating relatively stable and predictable token consumption under controlled conditions, while the qualitative findings addressed RQ2 by indicating a positive perception of the organisational usefulness of the generated notes in primary healthcare practice.

5. Discussion and Conclusions

The research topic remains highly relevant in an era where the term 'artificial intelligence' appears in everyday discussions about technology and society. Continued research and debate in the medical field are particularly important as healthcare systems rely heavily on information processing. On the one hand, many aspects of medical information follow recurring patterns, which makes them potentially suitable for automation with the use of large language models. On the other hand, medical data are highly sensitive and confidential, making privacy protection and careful system design a fundamental priority.

The quantitative results of this study indicated that token consumption in the tested prototype remained relatively stable and predictable, whilst the experiment was conducted using relatively short and controlled patient inputs, which resulted in marginal computational costs. In real-world scenarios, however, patient narratives are likely to be more diverse and less structured, which may influence token usage and overall system costs. This finding complements previous studies indicating that the operational costs of LLM-based systems are directly related to token consumption (Zhao et al., 2026). Unlike previous research discussing the computational efficiency of large language models more broadly, the presented study illustrates this relationship using controlled primary healthcare scenarios.

The qualitative component of the study, based on a semi-structured interview with a physician with twenty years of experience in Polish primary healthcare, suggested that such a solution could be beneficial in everyday clinical practice. According to the expert evaluation, structured notes generated prior to the consultation could help physicians familiarise themselves with the patient's situation before the visit and reduce informational chaos during the initial stage of the consultation. This observation is consistent with previous research suggesting that AI systems in healthcare should support, rather than replace, clinicians' professional judgement (Asan et al., 2020). Likewise, the expert participating in this study emphasised that the generated notes should serve as a structured starting point for the consultation and would still require verification and follow-up questions from the physician. The interview also highlighted potential organisational benefits, including improved management of visits involving patients with infectious diseases, where shorter and more focused interactions may help reduce exposure within the clinic.

This study had several limitations that should be acknowledged. First, the qualitative component was based on the perspective of a single expert, which limits the generalisability of the findings. In addition, the selection of the participant based on accessibility may introduce potential prejudice. Although the expert had extensive professional experience in Polish primary healthcare, the evaluation reflected an individual perspective and should therefore be interpreted as an initial exploratory assessment rather

than evidence applicable to the broader population of physicians. Second, the study was of an exploratory and pilot character, as the analysis was conducted on a limited number of predefined patient scenarios. The analysed cases were hypothetical and represented only a selected range of common health complaints rather than real clinical data. Consequently, the observed stability of token consumption should be interpreted with caution, as real-world patient narratives are likely to be more diverse in terms of language, structure, length, and completeness, which may influence computational costs and system performance.

Furthermore, the prototype did not include all elements of a complete medical interview, which indicates the need for further development and expansion of the system, particularly in terms of incorporating a broader range of clinically relevant questions.

Future implementations would likely involve higher computational costs, especially if additional components such as speech-to-text processing were introduced. Moreover, practical deployment would require appropriate technical infrastructure as well as trained personnel supporting patient interaction with the system. Future research should also validate the proposed approach using a larger and more diverse set of patient scenarios based on real clinical data. In addition, involving a broader group of primary healthcare physicians would enable the assessment of organisational usefulness from multiple professional perspectives and increase the strength of the qualitative findings.

From a theoretical perspective, this article contributes to the literature by linking the analysis of computational cost with the organisational value of AI-based solutions in primary healthcare. It also extends the existing knowledge by focusing on the initial stage of the medical encounter, which has received limited attention in prior studies, and directly addresses the research questions. The results indicate that, in relation to RQ1, token consumption can remain relatively stable and predictable under controlled conditions. In relation to RQ2, the organisational usefulness of the generated notes was perceived positively by the expert. From a practical perspective, the findings suggested that language model-based tools may support physicians in preparing for consultations and improve the organisation of patient information, although their implementation requires careful consideration of infrastructure, costs, and data privacy.

Thus, the article highlights the potential organisational value of language model-based tools in primary healthcare, whilst also indicating the need for further research and development before such solutions could be implemented in real clinical environments.

References

- Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, Technology, and Applications. *Machine Learning with Applications*, 2, 100006. <https://doi.org/10.1016/J.Mlwa.2020.100006>
- Asan, O., Bayrak, A., & Choudhury, A. (2020). Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *Journal of Medical Internet Research*, 22(6), e15154. <https://doi.org/10.2196/15154>
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S. von, Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2022). *On the Opportunities and Risks of Foundation Models*. arXiv:2108.07258. <https://doi.org/10.48550/arXiv.2108.07258>
- Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L., & Zimmel, R. (2023). *The Economic Potential of Generative AI. The Next Productivity Frontier*. McKinsey & Company. Retrieved April 3, 2026, from <https://www.mckinsey.com/~media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/the%20economic%20potential%20of%20generative%20ai%20the%20next%20productivity%20frontier/the-economic-potential-of-generative-ai-the-next-productivity-frontier.pdf>
- Davenport, T. H., & Ronanki, R. (2018). Artificial Intelligence for the Real World. *Harvard Business Review*.
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature*, 542(7639), 115-118. <https://doi.org/10.1038/nature21056>
- Gholipour, M., Khajouei, R., Amiri, P., Gohari, S. H., & Ahmadian, L. (2023). Extracting Cancer Concepts from Clinical Notes Using Natural Language Processing: A Systematic Review. *BMC Bioinformatics*, 24, 405. <https://doi.org/10.1186/s12859-023-05480-0>

- İşgüzar, S., Fendoğlu, E., & Şimşek, A. I. (2024). Innovative Applications in Businesses: An Evaluation on Generative Artificial Intelligence. *Amfiteatru Economic*, 26(66), 511-530. <https://doi.org/10.24818/EA/2024/66/511>
- Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial Intelligence in Healthcare: Past, Present and Future. *Stroke and Vascular Neurology*, 2(4). <https://doi.org/10.1136/svn-2017-000101>
- Kellermann, A. L., & Jones, S. S. (2013). What It Will Take to Achieve the As-Yet-Unfulfilled Promises of Health Information Technology. *Health Affairs*, 32(1). <https://doi.org/10.1377/hlthaff.2012.0693>
- Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M. K., Pei, J., Ting, M. Y. L., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., ... Zhang, K. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, 172(5), 1122-1131, e9. <https://doi.org/10.1016/j.cell.2018.02.010>
- Mohapatra, B., Mohapatra, S., & Mohapatra, S. (2023). *Process Automation Strategy in Services, Manufacturing and Construction* (chapter 2: *Business Process Automation*). Emerald Publishing Limited. <https://doi.org/10.1108/978-1-80455-143-120231004>
- Możarowski, W., Mateńko, J., Krupa, A., Kiełbratowska, J., Mozer, M., Potrykus, M., Pałuchowski, M., Pietruszka, W., Klasicki, P., & Krawczyk, A. (2025). Rola, cele i perspektywy rozwoju innowacji oraz nowych technologii cyfrowych w praktyce lekarza medycyny rodzinnej w Polsce. *Forum Medycyny Rodzinnej*, 19(3), 121-130. <https://doi.org/10.5603/fmr.104779>
- Narodowy Fundusz Zdrowia. (2013). *Zarządzenie Nr 69/2013/DSOZ Prezesa Narodowego Funduszu Zdrowia z dnia 27 grudnia 2013 r. w sprawie określenia warunków zawierania i realizacji umów o udzielanie świadczeń opieki zdrowotnej w rodzaju podstawowa opieka zdrowotna*. Retrieved April 3, 2026, from <https://www.nfz.gov.pl/zarzadzania-prezesa/zarzadzania-prezesa-nfz/zarzadzenie-nr-692013dsoz,5800.html>
- OECD & European Observatory on Health Systems and Policies. (2024). *Polska: Profil systemu ochrony zdrowia 2023*. OECD Publishing. <https://doi.org/10.1787/b12d3d03-pl>
- OpenAI Developers. (n.d.). *Pricing*. Retrieved April 3, 2026, from <https://developers.openai.com/api/docs/pricing/>
- Poon, E. G., Lemak, C. H., Rojas, J. C., Guptill, J., & Classen, D. (2025). Adoption of Artificial Intelligence in Healthcare: Survey of Health System Priorities, Successes, and Challenges. *Journal of the American Medical Informatics Association*, 32(7), 1093-1100. <https://doi.org/10.1093/jamia/ocaf065>
- Raghupathi, W., & Raghupathi, V. (2014). Big Data Analytics in Healthcare: Promise and Potential. *Health Information Science and Systems*, 2(1), 3. <https://doi.org/10.1186/2047-2501-2-3>
- Sarker, I. H. (2022). AI-Based Modelling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems. *SN Computer Science*, 3(2), 158. <https://doi.org/10.1007/s42979-022-01043-x>
- Sasseville, M., Yousefi, F., Ouellet, S., Naye, F., Stefan, T., Carnovale, V., Bergeron, F., Ling, L., Gheorghiu, B., Hagens, S., Gareau-Lajoie, S., & LeBlanc, A. (2025). The Impact of AI Scribes on Streamlining Clinical Documentation: A Systematic Review. *Healthcare*, 13(12), 1447. <https://doi.org/10.3390/healthcare13121447>
- Suska, E. (2025). Sztuczna inteligencja w medycynie. *Pedagogika Społeczna*, (3), 83-90. <https://doi.org/10.35464/1642-672X.PS.2025.3.08>
- Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. National Institute of Standards and Technology (U.S.). <https://doi.org/10.6028/NIST.AI.100-1>
- Topol, E. J. (2019). High-Performance Medicine: The Convergence of Human and Artificial Intelligence. *Nature Medicine*, 25(1), 44-56. <https://doi.org/10.1038/s41591-018-0300-7>
- Zhao, W., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., & Wen, J.-R. (2026). A Survey of Large Language Models. *Frontiers of Computer Science*, 20. <https://doi.org/10.1007/s11704-026-60308-3>

Modelowanie kosztów i wartości organizacyjnej systemów opartych na modelach językowych w podstawowej opiece zdrowotnej

Streszczenie

Cel: Celem artykułu jest analiza ekonomicznych i organizacyjnych implikacji prototypowego systemu opartego na dużym modelu językowym (LLM), zaprojektowanego do strukturyzowania danych z wywiadu wstępnego pacjenta w polskiej podstawowej opiece zdrowotnej.

Metodyka: Badanie ma charakter eksploracyjny i wykorzystuje podejście mieszane, łączące analizę zużycia tokenów w dziesięciu zdefiniowanych scenariuszach pacjentów z półustrukturyzowanym wywiadem eksperckim.

Wyniki: Wyniki ilościowe wskazują na stabilność i przewidywalność zużycia tokenów, co umożliwia wstępne modelowanie kosztów obliczeniowych w warunkach kontrolowanych. Wyniki jakościowe sugerują potencjalną użyteczność organizacyjną systemu, szczególnie w zakresie wspierania strukturyzacji informacji przed konsultacją lekarską.

Implikacje i rekomendacje: Uzyskane wyniki wskazują, że rozwiązania oparte na modelach językowych mogą wspierać procesy w podstawowej opiece zdrowotnej w sposób uwzględniający koszty, pod warunkiem zapewnienia odpowiedniej ochrony danych oraz nadzoru specjalistycznego. Dalsze badania powinny koncentrować się na wdrożeniach w warunkach rzeczywistych, wykorzystaniu bardziej zróżnicowanych danych pacjentów oraz rozszerzeniu funkcjonalności systemu.

Oryginalność/wartość: Artykuł wnosi wkład do literatury poprzez powiązanie analizy kosztów obliczeniowych z oceną użyteczności organizacyjnej rozwiązań opartych na sztucznej inteligencji w podstawowej opiece zdrowotnej, ze szczególnym uwzględnieniem wstępnego etapu wywiadu medycznego, który pozostaje niedostatecznie zbadany.

Słowa kluczowe: duże modele językowe, zużycie tokenów, modelowanie kosztów, wsparcie decyzji organizacyjnych, podstawowa opieka zdrowotna
