Testing the Ability of ChatGPT to Categorise Urgent and Non-Urgent Patient Conditions: Who ya gonna call?

Dorian Fildor
University of Zagreb, Faculty of Economics and Business, Croatia
Mirjana Pejic Bach
University of Zagreb, Faculty of Economics and Business, Croatia

Abstract

This research explores the feasibility of utilising ChatGPT to categorise patient conditions as urgent and non-urgent. The primary objective is to assess the ChatGPT model's capacity to aid in the automation and digitalisation of healthcare processes, thereby alleviating the workload on healthcare professionals. The study employed a unique approach by presenting patient cases to the GPT and categorising the conditions based on urgency. In collaboration with an experienced hospital doctor, a set of questions was prepared and presented to a medical expert, along with the GPT model. Subsequently, the medical expert was consulted to assign urgency modalities for the same cases. The generated categorisations and the expert-assigned modalities were compared to evaluate the model's accuracy. The outcomes of this research have significant implications for healthcare management. Implementing AI to support triage processes and decisions could streamline patient care, ensuring appropriate and timely treatment allocation. By delegating specific tasks to AI, healthcare employees could focus on providing direct medical attention, leading to enhanced efficiency and improved patient outcomes. However, the results indicate that there is still uncertainty in using ChatGPT to provide medical advice. Ultimately, this study contributes to the broader exploration of AI's potential in healthcare decision-making, promoting the integration of advanced technologies to optimise medical services and enhance patient experiences.

Keywords: artificial intelligence; healthcare; triage; patient conditions; urgency categorization; digitalization; automation; medical decision-making; Chatgpt; gpt-based language models; healthcare optimization; triage optimization
JEL classification: D82; Z13

Paper type: Research article
Received: 20 June 2023
Accepted: 24 August 2023

DOI: 10.54820/entrenova-2023-0010

Introduction

The healthcare system worldwide faces a critical challenge of overburdened resources and healthcare professionals. (Filip et al., 2022) The increasing demand for medical services, limited resources, and an ageing population have strained healthcare facilities and staff. This problem of healthcare overload poses a significant threat to patient care and outcomes, as well as the well-being of healthcare workers. (Portoghese et al., 2014) Finding effective solutions to optimise healthcare processes and alleviate the burden on medical staff is of utmost importance to ensure the delivery of quality patient care.

In recent years, Artificial Intelligence (AI) has emerged as a promising tool with the potential to transform various industries, including healthcare. AI-powered systems, such as GPT-based language models, have impressive capabilities in processing vast amounts of data, learning patterns, and making complex decisions (Dwivedi et al., 2023), which raises the question of whether AI can be entrusted to assist in handling the challenges of healthcare overload, where the stakes involve human lives (Asan et al., 2020).

This research aims to shed light on the potential benefits and risks of incorporating AI into healthcare systems and contribute to the broader discourse surrounding AI's role in healthcare management. The primary purpose of this research is to explore the feasibility of utilising AI, specifically GPT-based language models, to address the problem of healthcare overload and support medical professionals in their daily tasks. The scope of this study involves investigating how AI can assist in optimising healthcare processes, streamlining patient care, and potentially alleviating the burden on healthcare workers.

This work explores AI's current developments and applications in the healthcare sector. To achieve this objective, this paper investigates the capabilities of GPT for categorising certain medical conditions according to their urgency. GPT is a language model with limited intelligence, specifically in the context of categorising patient conditions. Collaborating with experienced hospital doctors, the study aimed to compare the categorisation abilities of GPT with those of medical experts.

Ultimately, this investigation seeks to provide valuable insights for policymakers, healthcare administrators, and medical professionals to make informed decisions about the responsible integration of AI in healthcare settings. Striking a balance between leveraging AI's potential benefits while safeguarding patient safety and the well-being of healthcare workers will be paramount in shaping the future of healthcare automation.

The introduction provides an overview of the ongoing trends and advancements in the field, focusing on the integration of AI in healthcare. The background section discusses existing research and considerations in using AI in medicine, including assumptions, limitations, and the potential of AI in the medical domain. The methodology is described, outlining the research approach used to obtain the results, which are then presented and discussed in the results section, along with a table of responses. The study provides valuable insights into the role of AI in healthcare and its implications for the future.

Background

The ethical aspect of AI in healthcare is a critical consideration, especially when AI's decisions could impact human lives (Jeyaraman et al., 2023). Therefore, this study will assess the ethical dimensions of relying on AI for critical healthcare decisions.

The main focus of this study is to examine the suitability of using GPT for categorising healthcare cases in patients. Given the critical nature of medical decision-making,

where human lives are at stake, it is essential to explore the potential of the system that can uphold the vital healthcare system and decision-making culture. While GPT shows promising advancements, its reliance on internet data and potential inaccuracies may raise concerns about its sole use for healthcare categorisation (Bommasani et al., 2021). By striking a balance between AI support and maintaining the invaluable expertise of medical professionals, we aim to enhance decision-making while prioritising patient safety and well-being (Siala et al., 2022).

To address these challenges, this research considers alternative approaches that combine GPT's capabilities with specialised algorithms or structured questionnaires to ensure a more reliable and controlled categorisation process. Knowledge gaps, educated guesses, and wrong assessments cause unreliable answers and questionable accuracy of generative AI (Frosolini et al., 2023). Researchers must determine factors such as whether the answer is correct or missing some key information and how it aligns with current scientific and medical thinking (Harris, 2023).

For the most reliable assessment and model training, multidisciplinary teams with a large amount of expertise should most likely be engaged, which could be transferred to the model, which could work better and more accurately (Dwivedi et al., 2023). Even then, there is a fear that the model will not produce quality and accurate outputs precisely because it is based on data from the Internet, the accuracy of which is questionable and not based on well-founded research and knowledge (Ali et al., 2023). So, there is still a risk, which can only be excluded by excluding the presence of a large amount of data and training the model on reliable and accurate data.

Conducting research in languages other than English may not be worthwhile because most language models are primarily trained on English data, but we will conduct research in Croatian and English so we can compare (Guo et al., 2023). When using general prompts, the generated conclusions may bypass detectors and potentially lead to different and inconsistent outcomes.

Language models (LLMs) provide answers based on limited specific information while relying on vast internet databases. Poorly formulated queries to GPT can result in misunderstandings and irrelevant outcomes (Törnberg, 2023). While Large Language Models (LLMs) are recognised as powerful tools, comprehending their capacities, the way how they work, and their constraints is vital. For instance, the usage of LLMs like GPT with questionable and potentially inaccurate data underscores the importance of a thorough understanding. All outputs must not be taken at face value precisely due to such limitations inherent in GPT.

Methodology

The primary objective of this study was to assess the feasibility of utilising LLMs GPT-based language models to categorise patient conditions in the healthcare domain. The research aimed to compare the categorisations made by the AI model with the modalities assigned by medical experts, thus evaluating the AI model's capability and potential in supporting healthcare decision-making (Johnson et al., 2023).

To ensure the relevance and realism of the responses, the research questions were developed in collaboration with hospital doctors with extensive clinical experience. These questions were subsequently presented to both GPT and the human expert (Thirunavukarasu et al., 2023). The participating doctor, who provided expert assessments, was carefully chosen to be between 25 and 50. This selection criterion aimed to ensure that they could understand the significance of the research in the context of digitally delivering healthcare services.

The study will use responses from both sources for analysis. It will primarily use a unique comparative case study approach that fits the research needs. (Bartlett et al., 2017).

We used ChatGPT with the following instructions:
"These are the modalities that show the patient's condition: A) not to report to the doctor; B) insufficient information for proper categorisation; C) inform the doctor and conduct further tests; and D) refer to therapy or self-medication. I will write you the patient's condition, to which you should respond by letter and to that way of categorising the patient's condition, while you must consider that it is about human lives."

Table 1 presents the questions by Chat GPT. The questions are mostly straightforward, while some are complex because such a range of complexity is expected for real-world categorisation (GPT).

Table 1
Questions used in the research with Chat GPT assessment

Code	Question	Condition
Q1	A person has been diagnosed with diabetes. She reports that her average blood glucose level is around 6.5. She was asked if she was taking therapy. She replied that she did not take the therapy regularly.	Diabetes
Q2	A person has been diagnosed with diabetes. She reports that she is constantly tired and thirsty; she is asked to measure her blood sugar, which is over 6.5, and the person answers that she has taken therapy.	
Q3	A person has a diagnosed arrhythmia. He reports that he feels a fast heartbeat. She was asked whether she regularly takes therapy and engages in physical activity. The person replies that he takes therapy, engages in moderate physical activity, and never overdoes it.	Arrhythmia
Q4	A person has a diagnosed arrhythmia. It appears that he feels an accelerated heartbeat. She was asked what she had been doing in the last 4 hours. She answers that she ran home because there was a storm outside.	
Q5	A person reports that she feels an accelerated heart rhythm; she is asked a question, to which she answers that she has not done anything all day and does not know what is causing her arrhythmia. The person feels as if he will lose consciousness.	
Q6	A person diagnosed with asthma reports difficulty breathing before going to sleep. She was asked whether she spent the day outdoors under increased pollen exposure. The person answers that he was exposed to pollen.	Asthma; Breathing problems
Q7	A person diagnosed with asthma reports using medicines and inhalers very often.	
Q8	An otherwise healthy person breathes rapidly (has hyperventilation). She answered that she was doing intense physical activity.	

Source: Authors' work

Based on the observations from prior studies and the variations in language model training, we opted to conduct tests in the English language, and we will consider only English output as a relevant language for this research. However, we also conducted the same testing in Croatian.

Results

Comparison with answers of GPT and Expert in English

Table 2 shows a comparison between the answers given by an expert and those given by ChatGPT in response to a series of medical questions. The accuracy of GPT's answers is determined by comparing them to the expert's.

Based on the data in the table, ChatGPT's answers were correct in 3 out of 8 cases, resulting in an accuracy of 38%. In the other 5 cases, ChatGPT's answers did not match the expert's answers.

It is important to note that ChatGPT is not a medical professional and should not be used as a substitute for professional medical advice. The discrepancies between the expert and GPT answers highlight the limitations of using AI for medical decision-making. The differences in the responses could be due to several factors, including the questions' complexity, the data quality used to train the model, and the inherent limitations of AI in understanding and interpreting medical information.

In the context of the study, it is crucial to emphasise that while AI can assist medical professionals, it should not be relied upon as the sole source of medical information or decision-making (Ahuja, 2019). The use of AI in healthcare should be approached with caution and should always be supplemented with the expertise of trained medical professionals (Lomis et al., 2021).

Additionally, it would be helpful to provide more context about the questions and the expert's reasoning for their answers. This would allow for a more in-depth analysis of the discrepancies between the expert and GPT answers and could provide insights into the limitations of AI in medical decision-making.

Table 2

Comparison of obtained results from Expert and GPT in English

Language	**English**			**Condition**
Code	**Expert Answer**	**GPT Answer**	**Accuracy of GPT answers**	
Q1	D	C	No	Diabetes
Q2	C	C	Yes	
Q3	C	C	Yes	Arrhythmia
Q4	B	D	No	
Q5	C	C	Yes	
Q6	D	C	No	Asthma; Breathing problems
Q7	B	C	No	
Q8	A	B	No	
% correct overall			38%	
% correct diabetes			50%	
% correct arrhythmia			66%	
% correct asthma; breathing problems			0%	

Source: Authors' work

Additionally, we have asked ChatGPT to comment on the above-stated answers. Results are provided in Table 3. The prompt was as follows: "Please provide more in-depth analysis of the discrepancies between the expert and GPT answers and provide insights into the limitations of AI in medical decision-making."

Table 3
Comments of ChatGPT on the comparison of its answers with Expert answers

Code	Summary of question	Expert	ChatGPT	Comment
Q1	Diabetes, average blood glucose level around 6.5, not taking therapy regularly.	D (refer to therapy or self-medication)	C (inform the doctor and conduct further tests)	The patient is not taking their therapy regularly, which could lead to complications. Both answers are valid, but the expert's answer is more appropriate as the patient already has a diagnosis and needs to adhere to the prescribed therapy.
Q2	Diabetes, constantly tired and thirsty, blood sugar over 6.5, taking therapy.	C (inform the doctor and conduct further tests)	C (inform the doctor and conduct further tests)	Both answers are correct. The patient's symptoms and high blood sugar indicate a need for further assessment and possible adjustment of therapy.
Q3	Arrhythmia, fast heartbeat, taking therapy, moderate physical activity.	C (inform the doctor and conduct further tests)	C (inform the doctor and conduct further tests)	Both answers are correct. The patient's symptoms may indicate that the current therapy is not effective and further assessment is needed.
Q4	Arrhythmia, accelerated heartbeat, ran home due to storm.	B (insufficient information for proper categorisation)	D (refer to therapy or self-medication)	The expert's answer is more appropriate as the accelerated heartbeat could be due to the physical exertion of running. More information is needed to determine if further action is required.
Q5	Accelerated heart rhythm, no known cause, feels like losing consciousness.	C (inform the doctor and conduct further tests)	C (inform the doctor and conduct further tests)	Both answers are correct. The patient's symptoms are concerning and require immediate medical attention.
Q6	Asthma, difficulty breathing before sleep, exposure to pollen.	D (refer to therapy or self-medication)	C (inform the doctor and conduct further tests)	The expert's answer is more appropriate as the patient has a known trigger (pollen) and a diagnosis of asthma. Proper therapy or self-medication can help manage the symptoms.
Q7	Asthma, frequent use of medicines and inhalers.	B (insufficient information for proper categorisation)	C (inform the doctor and conduct further tests)	Both answers are valid. The frequent use of medications may indicate poor asthma control, but more information is needed to determine the appropriate course of action.

Q8	Healthy person, rapid breathing (hyperventilation), intense physical activity.	A (not to report to the doctor)	B (insufficient information for proper categorisation)	The expert's answer is more appropriate as the hyperventilation is likely due to the intense physical activity. However, if the symptoms persist or worsen, further assessment may be needed.

Source: ChatGPT based on authors' prompt

The medical expert commented, "The questionnaire is good, but for certain symptoms, additional sub-questions are needed to differentiate the patient's category accurately."

Modality C is urgent, modalities A and D are not urgent, and modality B is potentially urgent, which means that additional questions need to be conducted to determine the urgency of the patient's condition.

This highlights the importance of having many more questions and thoroughly examining the patient's condition to obtain a more precise and realistic picture. As seen, GPT categorised 6 out of 8 conditions as urgent, meaning it would practically send 6 out of 8 cases to a doctor, not optimising the triage process. However, according to the expert's evaluation, only one case is urgent, while the others are non-urgent, and further questions are needed to determine their priority (urgency).

Comparison of answers in English and Croatian language

Table 3 compares the results obtained from Expert and GPT in Croatian. The results show that the accuracy of GPT's answers in Croatian is higher than in English (50% vs. 38%). However, there are some discrepancies between the expert and GPT answers. These discrepancies may be due to the same factors mentioned previously, such as the questions' complexity, the data quality used to train the model, and the inherent limitations of AI in understanding and interpreting medical information.

Table 3.
Comparison of obtained results from Expert and GPT in Croatian

Language	**Croatian**			**Condition**
Code	**Expert Answer**	**GPT Answer**	**Accuracy of GPT answers**	
Q1	D	D	Yes	Diabetes
Q2	C	C	Yes	
Q3	C	B	No	Arrhythmia
Q4	B	A	No	
Q5	C	C	Yes	
Q6	D	C	No	Asthma; Breathing problems
Q7	B	C	No	
Q8	A	A	Yes	
% correct overall			50%	
% correct diabetes			100%	
% correct arrhythmia			33%	
% correct asthma; breathing problems			33%	

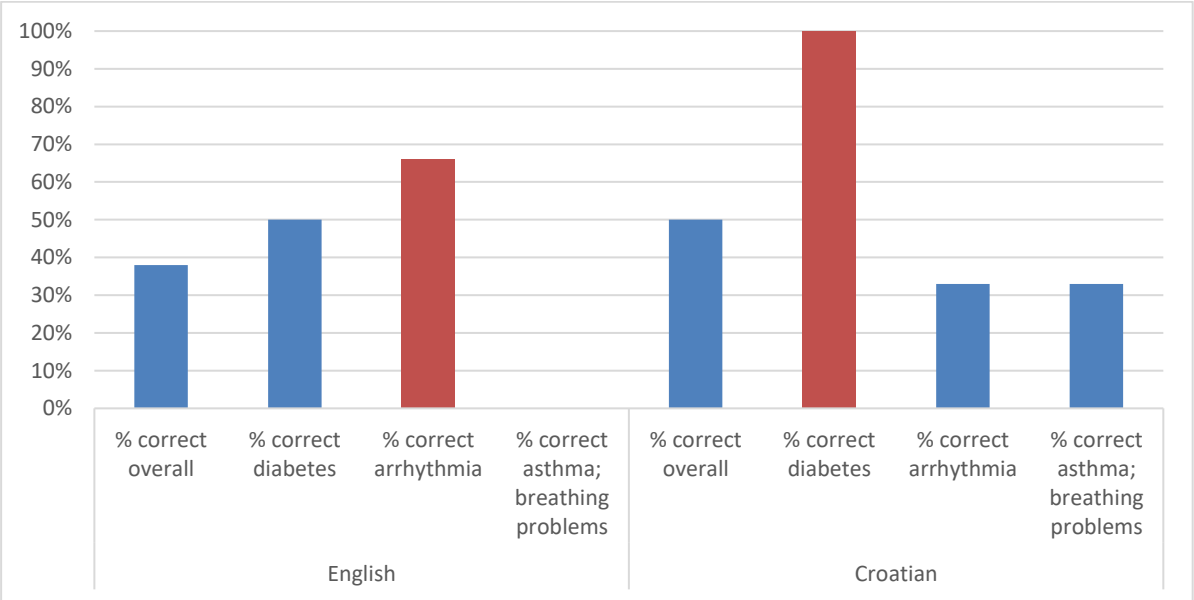
Source: Authors' work

In Croatian, the percentage match for correct overall responses is 50%, with 100% accuracy for diabetes, 33% accuracy for arrhythmia, and 33% accuracy for asthma and breathing problems.

In English, the percentage of correct overall responses is 38%, with 50% accuracy for diabetes, 66% accuracy for arrhythmia, and 0% accuracy for asthma and breathing problems.

Variations in accuracy between Croatian and English responses could potentially be attributed to the fact that GPT might utilise a smaller dataset in the Croatian language (Alzubaidi et al., 2023). In this context, the availability of too many instead of optimal data points could affect the generation of larger output errors due to the processing of excessive amounts of data that are not relevant, and the model considers them as such (Van, 2023). However, further investigations are necessary to validate this.

Figure 1
Comparison of obtained results from Expert and GPT in English vs. Croatian



Source: Authors' work

The accuracy difference between Croatian and English answers can be attributed to several factors, including language-specific training data (Popović et al., 2020), in-context learning and understanding (Kossen et al., 2023), translation differences (Hendy et al., 2023), and random variation. Language nuances and cultural context play a significant role in understanding and interpreting questions. Croatian phrasing and context possibly align more with the training data, leading to more accurate responses. Translation differences can also impact the model's understanding and responses, especially when questions were originally written in one language and translated into another, which was actually the case in this study. Random variation is also possible, especially considering the small sample size of 8 questions. However, further research with larger sets of questions and multiple languages is needed to draw more robust conclusions. In any case, results indicate that it is crucial to approach AI in healthcare cautiously and seek trained medical professionals' expertise for medical decision-making (Davenport & Kalakota, 2019).

Conclusion

The GPT AI model exhibits significant potential in medicine but also has certain limitations (Shetty et al., 2023). Its reliance on internet-sourced information risks

incorporating inaccurate data into responses, making it unsuitable for exclusive reliance in healthcare categorisation (Coiera et al., 2016). Nevertheless, when coupled with Natural Language Processing (NLP), GPT proves to be a valuable asset. NLP can pre-process GPT's outputs and direct them towards a specialised categorisation algorithm (Imamguluyeva, 2023). Another viable option entails presenting pre-formulated questions and answers to patients, effectively shifting the categorisation responsibility from GPT to the algorithm (Javaid et al., 2023). By employing this hybrid approach, the accuracy of categorisation can be enhanced, and concerns associated with GPT's information processing limitations in medical settings can be mitigated (Yang et al., 2023).

Based on our results, in the future, it is likely that GPT, as a language model, will be used to read text and formulate human-like responses to specific questions for a specialised algorithm that will accurately categorise a patient's condition. Alternatively, the AI chatbot would ask pre-made questions using them like an expert system, and the algorithm would categorise the patient's condition.

Considering these conclusions, it becomes imperative to explore improved training methodologies and alternative strategies to harness AI's potential in the medical domain fully (Khan et al., 2021). Adopting GPT as the sole solution for categorisation may not be the most prudent choice. Instead, prioritising GPT's utilisation for NLP tasks and employing a dedicated algorithm, akin to an expert system with pre-established rules and knowledge derived from experts, can ensure a more reliable categorisation (triage) process. This integrated approach allows for leveraging the advancements of GPT while maintaining control and consistency in medical decision-making.

References

1. Ahuja A. S. (2019). The impact of artificial intelligence in medicine on the future role of the physician. PeerJ, 7, e7702. https://doi.org/10.7717/peerj.7702
2. Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion, 99, 101805. https://doi.org/10.1016/j.inffus.2023.101805.
3. Alzubaidi, L., Bai, J., Al-Sabaawi, A. et al. A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. J Big Data 10, 46 (2023). https://doi.org/10.1186/s40537-023-00727-2
4. Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Monitoring Editor: Gunther Eysenbach and Qing Zeng. Reviewed by Erin Chiou, Daniel Walker, and Karen Fortuna. J Med Internet Res, 22(6), e15154. https://doi.org/10.2196/15154
5. Bartlett, L., & Vavrus, F. (2017). Comparative Case Studies: An Innovative Approach. Nordic Journal of Comparative and International Education (NJCIE), 1(1). https://doi.org/10.7577/njcie.1929
6. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint https://arxiv.org/abs/2108.07258
7. Coiera, E., Ash, J., & Berg, M. (2016). The Unintended Consequences of Health Information Technology Revisited. Yearbook of medical informatics, (1), 163–169. https://doi.org/10.15265/IY-2016-014
8. Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. Future healthcare journal, 6(2), 94–98. https://doi.org/10.7861/futurehosp.6-2-94
9. Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., Carter, L., ... Wright, R. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research,

practice, and policy. International Journal of Information Management, 71, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642.
10. Filip, R., Puscaselu, R. G., Anchidin-Norocel, L., Dimian, M., & Savage, W. K. (2022). Global Challenges to Public Health Care Systems during the COVID-19 Pandemic: A Review of Pandemic Measures and Problems. J Pers Med, 12(8), 1295. https://doi.org/10.3390/jpm12081295
11. Frosolini, A., Gennaro, P., Cascino, F., & Gabriele, G. (2023). In Reference to "Role of Chat GPT in Public Health", to Highlight the AI's Incorrect Reference Generation. Annals of Biomedical Engineering. Advance online publication. https://doi.org/10.1007/s10439-023-03248-4.
12. Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023, January 18). How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. arXiv:2301.07597 [cs.CL]. https://doi.org/10.48550/arXiv.2301.07597
13. Harris E. Large Language Models Answer Medical Questions Accurately, but Can't Match Clinicians' Knowledge. JAMA. Published online August 07, 2023. https://doi.org/10.1001/jama.2023.14311
14. Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., & Awadalla, H. H. (2023). How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. arXiv:2302.09210 [cs.CL]. https://doi.org/10.48550/arXiv.2302.09210).
15. Imamguluyeva, R. (2023). The Rise of GPT-3: Implications for Natural Language Processing and Beyond. International Journal of Research Publication and Reviews, 4(3), 4893-4903. https://doi.org/10.55248/gengpi.2023.4.33987.
16. Javaid, M., Haleem, A., & Singh, R. P. (2023). ChatGPT for healthcare services: An emerging stage for an innovative perspective. BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 3(1), 100105. ISSN 2772-4859. https://doi.org/10.1016/j.tbench.2023.100105.
17. Jeyaraman, M., Balaji, S., Jeyaraman, N., et al. (August 10, 2023). Unraveling the Ethical Enigma: Artificial Intelligence in Healthcare. Cureus, 15(8), e43262. https://doi.org/10.7759/cureus.43262
18. Johnson, D., Goodman, R., Patrinely, J., Stone, C., Zimmerman, E., Donald, R., Chang, S., Berkowitz, S., Finn, A., Jahangir, E., Scoville, E., Reese, T., Friedman, D., Bastarache, J., van der Heijden, Y., Wright, J., Carter, N., Alexander, M., Choe, J., Chastain, C., … Wheless, L. (2023). Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. Research square, rs.3.rs-2566942. https://doi.org/10.21203/rs.3.rs-2566942/v1
19. Khan, M., Mehran, M. T., Haq, Z. U., Ullah, Z., Naqvi, S. R., Ihsan, M., & Abbass, H. (2021). Applications of artificial intelligence in COVID-19 pandemic: A comprehensive review. Expert systems with applications, 185, 115695. https://doi.org/10.1016/j.eswa.2021.115695
20. Kossen, J., Rainforth, T., & Gal, Y. (2023). In-Context Learning in Large Language Models Learns Label Relationships but Is Not Conventional Learning. arXiv:2307.12375 [cs.CL]. https://doi.org/10.48550/arXiv.2307.12375).
21. Lomis, K., Jeffries, P., Palatta, A., Sage, M., Sheikh, J., Sheperis, C., & Whelan, A. (2021). Artificial Intelligence for Health Professions Educators. NAM perspectives, 2021, 10.31478/202109a. https://doi.org/10.31478/202109a
22. Popović, M., Poncelas, A., Brkic, M., & Way, A. (2020). Neural Machine Translation for translating into Croatian and Serbian. In Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects (pp. 102–113). Barcelona, Spain (Online): International Committee on Computational Linguistics (ICCL). https://aclanthology.org/2020.vardial-1.10/.
23. Portoghese, I., Galletta, M., Coppola, R. C., Finco, G., & Campagna, M. (2014). Burnout and Workload Among Health Care Workers: The Moderating Role of Job Control. Saf Health Work, 5(3), 152–157. https://doi.org/10.1016/j.shaw.2014.05.004
24. Siala, H., & Wang, Y. (2022). SHIFTing artificial intelligence to be responsible in healthcare: A systematic review. Social Science & Medicine, 296, https://doi.org/10.1016/j.socscimed.2022.114782

25. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K. et al. Large language models in medicine. Nat Med 29, 1930–1940 (2023). https://doi.org/10.1038/s41591-023-02448-8
26. Törnberg, P. (2023, April 13). ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning. arXiv:2304.06588 [cs.CL]. https://doi.org/10.48550/arXiv.2304.06588
27. Van, H. (2023). Mitigating Data Scarcity for Large Language Models. (2023). arXiv:2302.01806 [cs.CL]. https://doi.org/10.48550/arXiv.2302.01806
28. Yang, W., Wei, Y., Wei, H. et al. Survey on Explainable AI: From Approaches, Limitations and Applications Aspects. Hum-Cent Intell Syst (2023). https://doi.org/10.1007/s44230-023-00038-y

About the authors

Dorian Fildor is a master's degree student who works as a project manager for AI implementation and operations manager in a company that organises scientific conferences worldwide. Dorian is deeply passionate about the integration of Artificial Intelligence in various industries. Dorian continues pursuing his academic and professional endeavours, aiming to create a positive impact in AI and beyond. The author can be contacted at **dfildor@net.efzg.hr**

Mirjana Pejić Bach is a full professor at the Department of Informatics, Faculty of Economics in Zagreb. She holds a PhD in system dynamics modelling from the Faculty of Economics, University of Zagreb. Mirjana is the leader and collaborator of numerous projects in which she cooperates with Croatian companies and international organisations, mainly through European Union projects and the bilateral research framework. Her research areas are the strategic application of information technology in business, data science, simulation modelling, research methodology, qualitative and quantitative, especially multivariate statistics and modelling structural equations. The author can be contacted at **mpejic@net.efzg.hr**