

Zbornik 29. mednarodne multikonference
INFORMACIJSKA DRUŽBA – IS 2026
Zvezek K

Proceedings of the 29th International Multiconference
INFORMATION SOCIETY – IS 2026
Volume K

Uporaba umne zdravstvene
AI in Healthcare

Urednika / Editors

Matjaž Gams / Žiga Kolarič

<https://is.ijs.si>

October 2026 / 8 October 2026
Ljubljana, Slovenia

Urednika:

Matjaž Gams

Odsek za inteligentne sisteme, Institut »Jožef Stefan«, Ljubljana

Žiga Kolar

Odsek za inteligentne sisteme, Institut »Jožef Stefan«, Ljubljana

Založnik: Institut »Jožef Stefan«, Ljubljana

Priprava zbornika: Mitja Lasič, Vesna Lasič

Oblikovanje naslovnice: Vesna Lasič

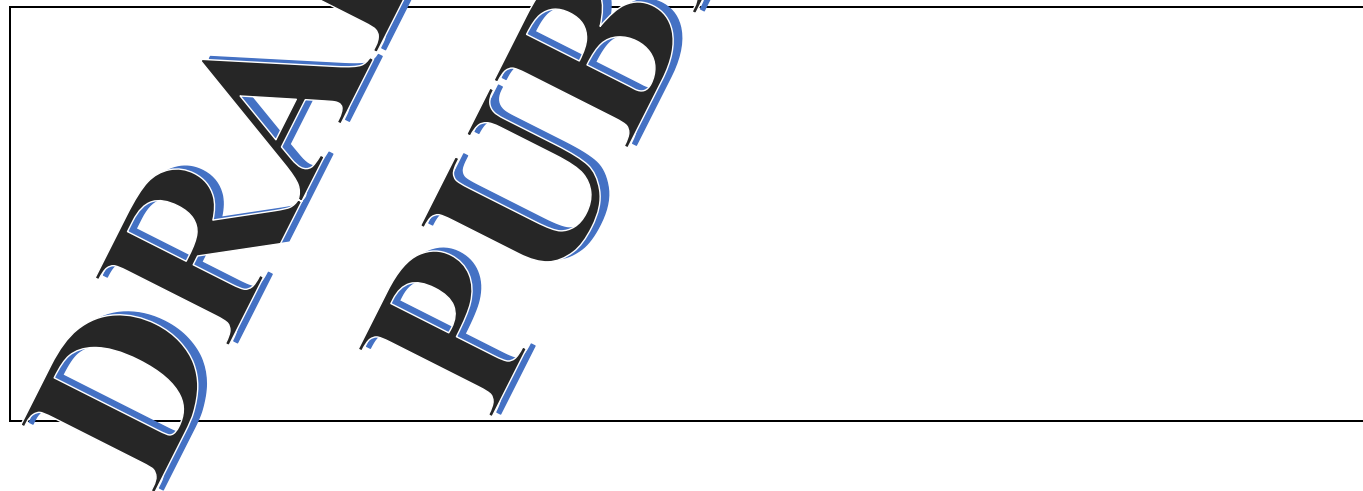
Dostop do e-publikacije:

<https://library.ijs.si/Stacks/Proceedings/InformationSociety>

Ljubljana, oktober 2026

Informacijska družba

ISSN 2630-371X



PREDGOVOR MULTIKONFERENCI INFORMACIJSKA DRUŽBA 2026

29. mednarodna multikonferenca *Informacijska družba* se odvija v času izjemne rasti umetne inteligence, njenih aplikacij in vplivov na človeštvo. Vsako leto bolj vstopamo v novo dobo, v kateri generativna umetna inteligenca ter drugi inovativni pristopi oblikujejo poti k superinteligenci in singularnosti, ki bosta krojili prihodnost človeške civilizacije. Naša konferenca je tako hkrati tradicionalna znanstvena in akademsko odprta, pa tudi inkubator novih, pogumnih idej in pogledov.

Letošnja konferenca poleg umetne inteligence vključuje tudi razprave o perečih temah današnjega časa: ohranjanje okolja, demografski izzivi, zdravstvo in preobrazba družbenih struktur. Razvoj UI ponuja rešitve za številne sodobne izzive, kar poudarja pomen sodelovanja med raziskovalci, strokovnjaki in odločevalci pri oblikovanju trajnostnih strategij. Zavedamo se, da živimo v obdobju velikih sprememb, ko je ključno, da z inovativnimi pristopi in poglobljenim znanjem ustvarimo informacijsko družbo, ki bo varna, vključujoča in trajnostna.

V okviru multikonference smo letos združili dvanajst vsebinsko raznolikih srečanj, ki odražajo širino in globino informacijskih ved: od umetne inteligence v zdravstvu, demografskih in družinskih analiz, digitalne preobrazbe zdravstvene nege ter digitalne vključenosti v informacijski družbi, do raziskav na področju kognitivne znanosti, zdrave dolgoživosti ter vzgoje in izobraževanja v informacijski družbi. Pridružujejo se konference o legendah računalništva in informatike, prenosu tehnologij, mitih in resnicah o varovanju okolja, odkrivanju znanja in podatkovnih skladiščih ter seveda Slovenska konferenca o umetni inteligenci.

Poleg referatov bodo okrogle mize in delavnice omogočile poglobljeno izmenjavo mnenj, ki bo pomembno prispevala k oblikovanju prihodnje informacijske družbe. »Legende računalništva in informatike« predstavljajo domači »Hall of Fame« za izjemne posameznike s tega področja. Še naprej bomo spodbujali raziskovanje in razvoj, odličnost in sodelovanje; razširjeni referati bodo objavljeni v reviji *Informatica*, s podporo dolgoletne tradicije in v sodelovanju z akademskimi institucijami ter strokovnimi združenji, kot so ACM Slovenija, SLAIS, Slovensko društvo Informatika in Inženirska akademija Slovenije.

Vsako leto izberemo najbolj izstopajoče dosežke. Letos je nagrado *Michie-Turing* za izjemen življenjski prispevek k razvoju in promociji informacijske družbe prejel **Denis Trček**, priznanje za raziskovalni dosežek leta pa **Jure Leskovec**. »Informacijsko limono« za najmanj primerno informacijsko storitev je prejel informacijski sistem za dolgotrajno oskrbo. »Informacijsko jagodo« za najboljši sistem ali storitev v letih 2025/2026 pa je prejela elektronska osebna izkaznica. Iskrene čestitke vsem nagrajencem!

Naša vizija ostaja jasna: prepoznati, izkoristiti in oblikovati priložnosti, ki jih prinaša digitalna preobrazba, ter ustvariti informacijsko družbo, ki koristi vsem njenim članom. Vsem sodelujočim se zahvaljujemo za njihov prispevek — veseli nas, da bomo skupaj oblikovali dosežke, ki bodo krojili usodo Slovenije in sveta.

Mojca Ciglarič, predsednica programskega odbora
Matjaž Gams, predsednik organizacijskega odbora

FOREWORD TO THE MULTICONFERENCE INFORMATION SOCIETY 2026

The 29th International Multiconference *Information Society* takes place at a time of remarkable growth in artificial intelligence, its applications, and its impact on humanity. Each year we move further into a new era in which generative artificial intelligence and other innovative approaches are opening paths toward superintelligence and the singularity — two developments that will shape the future of human civilization. Our conference is at once a traditional scientific meeting, an academically open forum, and an incubator of new and bold ideas.

In addition to artificial intelligence, this year's conference also addresses pressing issues of our time: environmental protection, demographic challenges, healthcare, and the transformation of social structures. The development of AI offers solutions to many contemporary challenges and underscores the need for collaboration among researchers, professionals, and policymakers in shaping sustainable strategies. We are aware that we live in a period of profound change, when innovative approaches and deep knowledge are essential if we are to build an information society that is safe, inclusive, and sustainable.

This year's multiconference brings together twelve thematically diverse meetings that reflect the breadth and depth of the information sciences. They range from artificial intelligence in healthcare, demographic and family studies, the digital transformation of nursing, and digital inclusion in the information society, to research in cognitive science, healthy longevity, and education in the information society. They are joined by conferences on the Legends of Computing and Informatics, technology transfer, myths and truths of environmental protection, knowledge discovery and data warehouses, and, of course, the Slovenian Conference on Artificial Intelligence.

Alongside the papers, round tables and workshops will enable an in-depth exchange of views that will contribute substantially to shaping the future information society. *Legends of Computing and Informatics* serves as a national Hall of Fame for outstanding individuals in the field. We will continue to foster research and development, excellence, and collaboration. Extended papers will be published in the journal *Informatica*, building on a long-standing tradition and on cooperation with academic institutions and professional associations, including ACM Slovenia, SLAIS, the Slovenian Society Informatika, and the Slovenian Academy of Engineering.

Each year we single out the most distinctive achievements. This year the *Michie–Turing Award* for an outstanding lifetime contribution to the development and promotion of the information society was presented to **Denis Trček**, and the Award for Research Achievement of the Year to **Jure Leskovec**. The “Information Lemon” for the least appropriate information service went to the long-term care information system. The “Information Strawberry” for the best system or service in 2025/2026 went to the electronic identity card. Our warmest congratulations to all the award winners.

Our vision remains clear: to recognize, use, and shape the opportunities of digital transformation, and to create an information society that benefits all of its members. We thank all participants for their contributions. We look forward to shaping, together, achievements that will determine the fate of Slovenia and of the world.

Mojca Ciglarič, Chair of the Program Committee
Matjaž Gams, Chair of the Organizing Committee

KONFERENČNI ODBORI

CONFERENCE COMMITTEES

International Programme Committee

Vladimir Bajic, South Africa
Heiner Benking, Germany
Se Woo Cheon, South Korea
Howie Firth, UK
Olga Fomichova, Russia
Vladimir Fomichov, Russia
Vesna Hljuz Dobric, Croatia
Alfred Inselberg, Israel
Jay Liebowitz, USA
Huan Liu, Singapore
Henz Martin, Germany
Marcin Paprzycki, USA
Claude Sammut, Australia
Jiri Wiedermann, Czech Republic
Xindong Wu, USA
Yiming Ye, USA
Ning Zhong, USA
Wray Buntine, Australia
Bezalel Gavish, USA
Gal A. Kaminka, Israel
Mike Bain, Australia
Michela Milano, Italy
Derong Liu, Chicago, USA
Toby Walsh, Australia
Sergio Campos-Cordobes, Spain
Shabnam Farahmand, Finland
Sergio Crovella, Italy

Organizing Committee

Matjaž Gams, chair
Mitja Luštrek
Lana Zemljak
Vesna Koricki
Mitja Lasič
Blaž Mahnič

Programme Committee

Mojca Ciglarich, chair
Bojan Orel,
Franc Solina,
Viljan Mahnič,
Cene Bavec,
Tomaž Kalin,
Jozsef Györkös,
Tadej Bajd
Jaroslav Berce
Mojca Bernik
Marko Bohanec
Ivan Bratko
Andrej Brodnik
Dušan Caf
Saša Divjak
Tomaž Erjavec
Bogdan Filipič
Andrej Gams
Matjaž Gams
Mitja Luštrek
Marko Grobelnik
Nikola Guid

Marjan Heričko
Borka Jerman Blažič Džonova
Gorazd Kandus
Urban Kordeš
Marjan Krisper
Andrej Kuščer
Jadran Lenarčič
Borut Likar
Janez Malačič
Olga Markič
Dunja Mladenich
Franc Novak
Vladislav Rajkovič
Grega Repovš
Ivan Rozman
Niko Schlamberger
Gašper Slapničar
Stanko Strmčnik
Jurij Šilc
Jurij Tasič
Denis Trček
Andrej Ule

Boštjan Vilfan
Baldomir Zajc
Blaž Zupan
Boris Žemva
Leon Žlajpah
Niko Zimic
Rok Piltaver
Toma Strle
Tine Kolenik
Franci Pivec
Uroš Rajkovič
Borut Batagelj
Tomaž Ogrin
Bojan Blažica
Matjaž Kljun
Robert Blatnik
Erik Dovgan
Špela Stres
Anton Gradišek
Borut Trpin
Anka Slana Ozimič

KAZALO / TABLE OF CONTENTS

Uporaba UI v zdravstvu / AI in Healthcare	1
PREDGOVOR / FOREWORD	3
PROGRAMSKI ODBORI / PROGRAMME COMMITTEES	4
HomeDOctor 2.0: Explainable Stateful Medical Intake within the European ChatMED Ecosystem / Bureau Mathieu, Gaudin Jonathan, Kocuvan Primož, Horvat Tadej, Gams Matjaž	5
Estimating the GDP-Relevant Effects of AI in Slovenian Healthcare through Reduced Sick Leave / Kocuvan Primož, Gams Matjaž	9
Overcoming Bias and Building Trust: Design of a Prospective Evaluation of Explainable AI in Melanoma Detection Across Diverse Skin Phototypes / Herke Anna-Katharina	17
Detecting Spin in Clinical Trials with Large Language Models / Ajdovec Tjaš, Robnik-Šikonja Marko, Šuster Simon	21
Evaluating Automatic Speech Recognition Performance in Individuals with Parkinson's Disease and Healthy Controls / Andonov Stefan, Damjanović Radomir, Trojachanec Dineva Katarina, Tojčanska Biljana, Ribarski Pance, Loshkovska Suzana, Simjanoska Misheva Monika, Mishev Kostadin, Stemenović Jelena, Lukić Stevo	26
Digital Health Across EU Borders: Data Protection and Artificial Intelligence in MyHealth@EU / Kocuvan Primož, Gams Matjaž	32
Physician-Rated Accuracy, Safety and Communication Quality of Large Language Model Responses to Epilepsy Questions in an Online Patient Forum: Two Parallel Evaluations / Lazić Lazar, Svetozarević Mihailo, Stojković Mina, Janković Milan, Svetozarević Isidora, Janković Sonja, Lukić Stevo	37
Benchmarking ChatGPT-5 Against ILAE Standards for Epilepsy-Associated Depression / Vučković Snežana, Janković Milan, Biševac Boban, Lukić Stevo	43
On Robustness of LLMs in Healthcare: A Combinatorial Testing Approach to Prompt Perturbations / Mujić Emir, Nica Iulia, Wotawa Franz	52
Indeks avtorjev / Author index	59

Zbornik 29. mednarodne multikonference
INFORMACIJSKA DRUŽBA – IS 2026
Zvezek K

Proceedings of the 29th International Multiconference
INFORMATION SOCIETY – IS 2026
Volume K

Uporaba umne zdravstvene
AI in Healthcare

Urednika / Editors

Matjaž Gams Žiga Kolarič

<https://is.ijs.si>

October 2026 / 8 October 2026
Ljubljana, Slovenia

FOREWORD

We are entering a period in which artificial intelligence (AI) is no longer a disruptive promise in medicine but a working instrument. In our view, AI has reached the capability of artificial general intelligence (AGI) and is moving toward superintelligent (SI or SAI) systems. For healthcare that means a change of era, not another incremental tool.

The most recent results make the scale of that change visible. An AI-designed drug candidate, rentosertib, developed for a rare lung disease, has in 2026 shown early clinical evidence that markers of biological age can fall on several independent aging clocks. Google DeepMind's AlphaGenome Atlas now maps the predicted molecular effect of billions of possible single-letter changes in the human genome. Rare-disease variants that used to take years of specialist reading can be ranked in software, and the same maps can be searched for drug targets. Medicine is acquiring a searchable layer over biology.

These are early findings, not finished therapies. They are nevertheless sufficient to say that extraordinary years lie ahead for patients: faster discovery, earlier diagnosis of rare conditions, and treatments designed with models rather than only with trial and error.

Responsibility remains. Privacy, bias, and clinical accountability do not disappear because the models are stronger. AI should support physicians in high-stakes decisions. What is no longer tenable is to treat proven medical AI as an experiment to be postponed.

A practical demonstration is already running. The LLM-based personal medical platform HomeDOCTOR, developed in the ChatMED line of work, has been in regular use in Slovenia since 2025, without operational failure of the kind often predicted for such systems. Prototypes have been tested in North Macedonia and Serbia. Continuous everyday use is evidence that large language models, properly supervised, can be reliable and of sufficient quality for real patients. National-scale adoption of systems of this class has been estimated to yield benefits on the order of €100 million.

This conference is convened in that spirit: to present current work, to keep claims tied to evidence, and to prepare healthcare for the extraordinary period that AI is opening.

Franz Wotawa

Monika Smiljanovska

Stevo Lukić

Matjaž Gams

PROGRAMSKI ODBOR / PROGRAMME COMMITTEE

Matjaž Gams

Monika Simjanoska Misheva

Stevo Lukić

Franz Wotawa

HomeDOctor 2.0: Explainable Stateful Medical Intake within the European ChatMED Ecosystem

Mathieu Bureau	Jonathan Gaudin	Primož Kocuvan	Tadej Horvat	Matjaž Gams
IUT de Cachan, BUT GEII	IUT de Cachan, BUT GEII	Jožef Stefan Institute	Jožef Stefan Institute	Jožef Stefan Institute
Université Paris-Saclay	Université Paris-Saclay	Ljubljana, Slovenia	Ljubljana, Slovenia	Ljubljana, Slovenia
Cachan, France	Cachan, France	Primož.Kocuvan@ijs.si	tadej.horvat@ijs.si	matjaz.gams@ijs.si
mathieubureau94@gmail.com	Jonathan.gaudin.pro@gmail.com			

Abstract

HomeDOctor 2.0 is a stateful conversational medical-intake assistant developed at the Jožef Stefan Institute within the European ChatMED ecosystem. Building on the knowledge-augmented counselling approach of HomeDOctor 1.0, we separate patient dialogue from information extraction, patient state, and case generation. A conceptual knowledge graph connects symptoms and patient context to follow-up questions, urgency assessment, and care recommendations. The implementation uses separate language-model calls and fixed prompts for structured extraction and response generation. Patient facts accumulate across turns, while diagnostic suggestions remain separate from the recorded history. We describe the state-update rules, prompting strategy, red-flag handling, and session privacy policy, and illustrate their relationship through a two-turn intake example. The contribution is a transparent account of the architecture and its implementation, providing a basis for technical inspection and further evaluation. Diagnostic accuracy and clinical safety have not yet been established through systematic clinical validation.

Keywords

HomeDOctor, conversational AI, medical intake, knowledge graph, patient state, structured extraction, clinical decision support, large language models, healthcare recommendation, red-flag detection, ChatMED.

1 Introduction

Patients usually describe symptoms in everyday language and may omit their duration, severity, associated symptoms, or relevant medical history. A conversational intake assistant must collect these details over several turns and retain them when asking follow-up questions. The resulting information should also be available in a form that a clinician can inspect without reconstructing the entire conversation.

HomeDOctor 1.0 integrated medical knowledge into GPT-4o for personal health counselling [1]. In HomeDOctor 2.0, we extend this approach with an explicit patient state and a reusable patient case. The work forms part of the European ChatMED ecosystem, which supports research collaboration and generative-AI adoption in healthcare, with the Jožef Stefan Institute as a partner [2].

Our objective is to explain how conversational input becomes structured information and how that information supports the next question or care recommendation. We present the architecture, describe the underlying transformations, and connect them through a worked example. Explainability in this paper refers to the visibility of patient facts, intermediate representations, and data flow. It does not imply that the internal reasoning of the language model is accessible or clinically verified.

2 Background and related work

Healthcare chatbots provide conversational access to information and services, with transparency and human-AI interaction as important design concerns [3]. Medical question-answering research has demonstrated the capabilities of large language models (LLMs), while also showing the need to evaluate factuality, potential harm, and clinical appropriateness [4]. Stress tests of GPT-3-based healthcare agents identified erroneous information and unsafe recommendations [5]. Omiye et al. review further limitations affecting the use of LLMs in medicine [6]. These findings motivate explicit handling of patient information and evaluation of individual system components.

Knowledge graphs make entities and their relationships explicit [7]. We use this form to explain the medical and conversational entities in HomeDOctor 2.0. Retrieval-augmented generation supplies external knowledge to a model [8], whereas session state preserves the patient-specific information collected during an interaction. Our contribution combines these established ideas in a medical-intake workflow. It concerns the organisation of the application and its interfaces, without introducing a new retrieval algorithm or a validated medical ontology.

WHO guidance on large multimodal models highlights unreliable outputs, automation bias, and privacy risks [9]. We therefore position HomeDOctor as support for information collection before consultation. Its recommendations require appropriate clinical evaluation before they can be treated as reliable medical decisions.

3 Architecture and knowledge representation

We developed the representation from the HomeDOctor 1.0 design [1], our intake implementation, and the web application workflow. We first identified medical entities, including symptoms, medication, conditions, and red flags. We then connected them to messages, conversation history, patient state, and generated outputs. Each connection represents a data transformation or a dependency in the intake process.

Figure 1 presents the resulting conceptual knowledge graph. Patient context and symptom collection feed follow-up

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, 5–9 October 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
<http://doi.org/10.70314/is.2026.aihc.3>

questions, information extraction, and urgency assessment. The recommendation layer includes possible diagnoses, care-path guidance, and patient reports. These outputs are generated from the available information and remain separate from patient-reported facts. The graph describes the architecture and its intended scope, rather than a clinical inference engine.

Figure 2 shows the web workflow. A FastAPI endpoint receives a message and session identifier, extraction updates the

patient state in Redis, and the state is formatted as a patient case. The case then supports the next response and optional diagnostic processing. Conversation history preserves the dialogue, while patient state provides a concise representation of the clinical information already collected. In the figure, persistence refers to continuity across turns within a session.

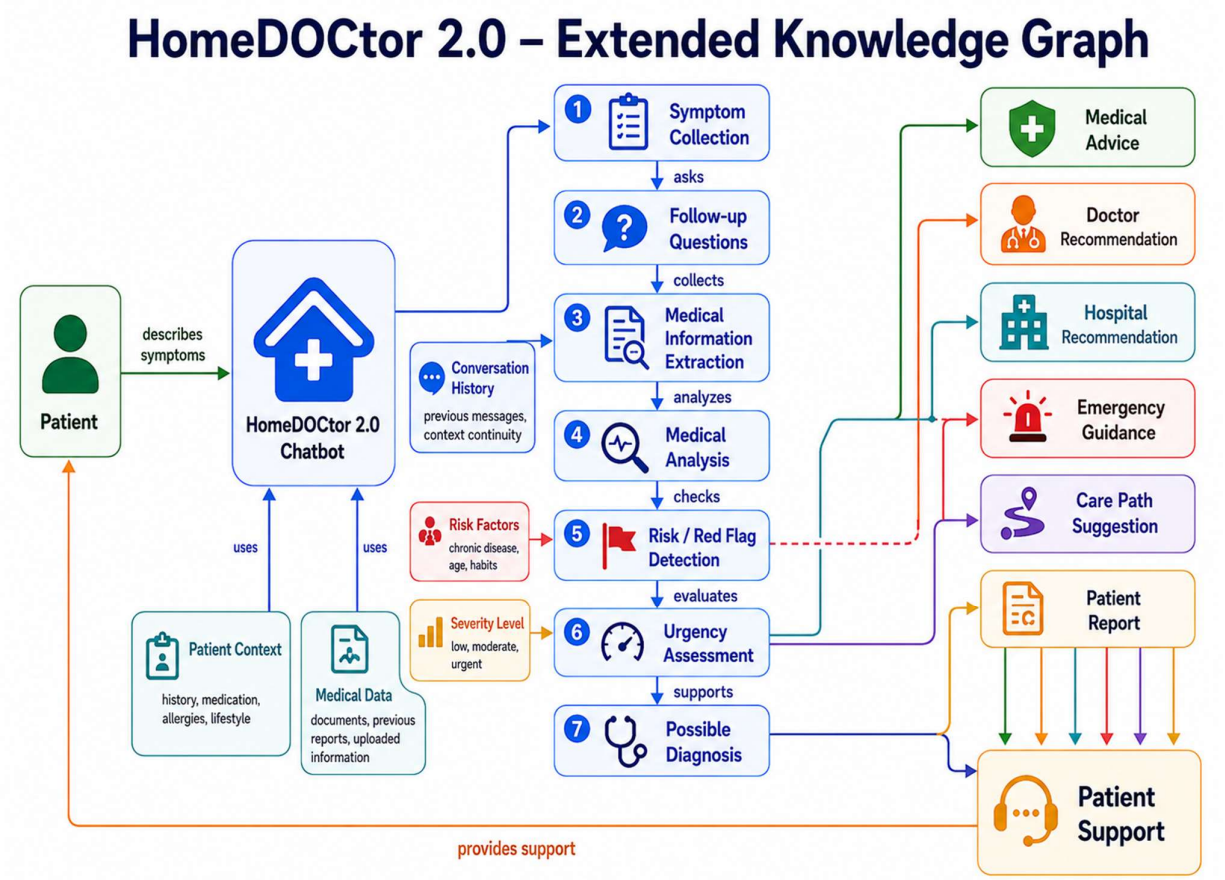
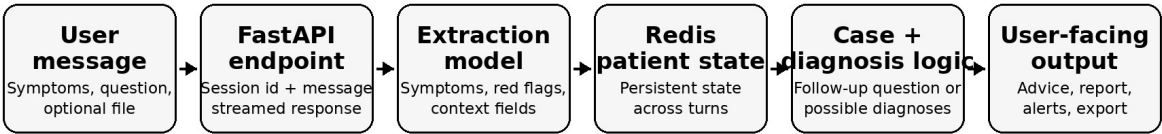


Figure 1: HomeDOctor 2.0 conceptual knowledge graph connecting patient context, symptom collection, urgency assessment, and recommendation outputs.

HomeDOctor 2.0 conversational workflow



The model does not replace medical diagnosis: red flags and uncertainty should direct the user toward professional care.

Figure 2: HomeDOctor 2.0 web workflow. Patient state is retained across turns for the duration of the conversation.

4 Implementation

4.1 Models and prompting

We currently use GPT-5.6 Luna for both information extraction and response generation. The two roles use separate calls and fixed instructions. Preparation for local inference with a Qwen or Gemma family model is ongoing, and we do not report local-model results here.

The extraction call combines `EXTRACTOR_PROMPT` with the preceding assistant message and the latest patient reply. This context helps interpret short answers to follow-up questions. The model returns structured JSON parsed into the `IntakeExtraction` schema. Response generation receives `HOME_DOC_PROMPT`, a language instruction, the conversation history, and a temporary prompt constructed from the current patient case. This prompt directs the model to consider missing information, respond with brief empathy, and ask one useful follow-up question. After diagnostic processing, it also requests a care recommendation. Temporary prompts are rebuilt for each call rather than accumulated in the conversation history.

We supplement model prompts with retrieval from medical sources and a database of disease–symptom associations expressed as percentages. The associations provide supporting medical context. We have not calibrated them as probabilities of an individual patient’s diagnosis. Exact replication also depends on the full prompt texts, schema definitions, retrieval corpus and settings, database version, and diagnostic-module configuration, which are not included in this paper.

4.2 Patient state and recommendation generation

The `update_state` function matches symptoms by name. For an existing symptom, newly extracted attributes replace stored values when they are neither null nor empty. Otherwise, the previous values remain available. New symptoms are appended. Medication, past-condition, addiction, allergy, and negative-finding lists are merged by set union. Age and gender are retained when absent from the new extraction, while `red_flag` and `medical_relevance` take the latest extracted values.

The `generate_patient_case` function formats the stored information into a structured text summary. It includes demographics, medication, history, allergies, negative findings, and each symptom’s location, onset, duration, severity, and type. Unreported values appear as “unknown”. Case generation is deterministic formatting of patient state, so it does not introduce a separate diagnostic interpretation. The conversational model receives the updated case on each turn.

The `Konzilij` module accepts the patient case and returns a diagnostic summary and recommendation. The wrapper includes up to three candidate diagnoses with likelihood scores, self-help information, and contextual text. These outputs become additional context for response generation without overwriting `PatientState`. Our command-line development version uses a fixed first-turn trigger for this module. A criterion based on the completeness of the case remains to be defined for systematic evaluation.

4.3 Red flags and urgent care

Red-flag detection uses the LLM with support from the disease–symptom database. The extractor records a Boolean `red_flag` in the patient state. Our intended emergency path gives a positive flag priority over routine follow-up and optional diagnostic processing, directing the patient to urgent professional care. The example in Section 5 illustrates this requirement.

In our command-line development version, the explicit `red_flag_protocol` call is disabled and the formatted case omits the flag. The demonstration handler displays an urgency message but cannot call emergency services. The intended response therefore requires an enabled routing path that passes the flag directly to the patient-facing warning and is tested end to end.

4.4 Privacy and session management

Our user manual discourages personal identifying details and prohibits them in some contexts. This is an input policy, without automatic anonymization. The web application holds encrypted session state in Redis memory during the conversation, writes no conversation data to disk, and deletes the state when the conversation ends. This limits retained patient data but prevents retrospective analysis of conversation records. Application-side deletion does not determine an external model provider’s retention policy.

5 Illustrative intake interaction

This constructed example connects patient input, state updates, and the intended care recommendation. It is not a recorded patient conversation or a measured system result.

Turn 1, patient: “I have had a headache since this morning, about 4 out of 10.”

Extracted state: symptom = headache, duration = since this morning, severity = 4/10. Age, medication, onset pattern, and associated symptoms remain unknown. The red-flag assessment is incomplete.

Follow-up: “Do you also have fever, a stiff neck, or discomfort in bright light?”

Turn 2, patient: “Yes, my temperature is 39°C, my neck is stiff and light hurts my eyes. I am 52 and take amlodipine.”

Updated state: the headache, duration, and severity are retained. Fever, neck stiffness, light sensitivity, age = 52, and medication = amlodipine are added. The expected red-flag assessment is positive.

Patient case, abridged: age: 52. Medication: amlodipine. Symptoms: headache since this morning, severity 4/10, fever of 39°C, neck stiffness, and light sensitivity. Allergies and past conditions: unknown.

Expected recommendation: “This combination of symptoms needs emergency medical assessment. Call 112 now.” The warning signs are consistent with emergency guidance for headache [10], and 112 is the emergency number in Slovenia [11]. Further intake and diagnostic deliberation should not delay this advice. The patient must make the call.

6 Discussion and limitations

The worked example shows why patient state is useful: the second reply adds clinically relevant information without requiring the patient to repeat the original complaint. The case retains both sets of facts, making the input to recommendation generation available for inspection. Keeping candidate diagnoses in a separate output distinguishes generated hypotheses from patient-reported history. These are architectural properties, and their effect on clinical outcomes remains to be evaluated.

The patient case provides an interface for clinical review: a clinician can inspect the available information and identify details requiring clarification. A consistent case format alone cannot establish fidelity, because a summary can omit nuance or preserve an extraction error.

Our current merge rules have limitations. Matching by symptom name depends on consistent terminology, and set union does not resolve corrections or withdrawal of previous statements. A newly extracted false red flag can replace an earlier true value. These cases require explicit handling of contradictions and safety-relevant information across turns. The emergency route described in Section 4.3 also requires end-to-end verification.

We operated HomeDOCTOR 1.0 online for approximately eighteen months. We did not retain conversation records or record conversation counts under our privacy policy, so we cannot report the number of users or encounters. Users could submit feedback voluntarily. This was not a systematic physician or user survey, and we did not establish a predefined definition or monitoring protocol for major medical incidents. We therefore draw no conclusions about clinical safety or overall satisfaction from the period of operation or the feedback channel.

We used internal benchmarks to examine conversation quality, but this paper does not present their composition, sample size, scoring protocol, or numerical results. Our evidence is consequently limited to the implementation description and illustrative case. A systematic evaluation should examine information extraction, state updates, case fidelity, and recommendation appropriateness separately, because a fluent final response can conceal an error introduced earlier in the workflow.

To isolate the contribution of patient state, we propose a paired comparison using the same scripted cases, model, medical sources, and dialogue instructions. One configuration would receive conversation history alone, while the other would also receive the state and case. Clinician annotations at each turn would distinguish reported facts, denied findings, and unknown information. Measures should cover retained facts, unsupported additions, omissions, and handling of corrections. Repeated runs would reveal variation in model outputs, with separate results for complete and incomplete histories.

Emergency tests should examine the complete route from extraction through state updates to the visible recommendation. A positive flag is insufficient if it is subsequently lost or does not change the advice. A fixed set of urgent and clinically similar non-urgent cases, assessed against predefined clinician criteria, would support sensitivity and false-alarm estimates. Language variation and symptom corrections should test whether urgency is preserved across turns.

Constructed, clinician-annotated cases would allow repeatable testing without retaining ordinary user histories. This approach is compatible with our session privacy policy, while prospective clinical validation would still be needed.

7 Conclusion

We presented HomeDOCTOR 2.0 as a stateful medical-intake architecture within the ChatMED ecosystem. Its main contribution is the explicit connection between conversational input, structured patient state, case generation, and recommendation support. The knowledge graph and worked example make these relationships understandable, while the implementation description identifies the prompts and update rules that govern them. Further work will address state corrections, emergency routing, reproducible configuration, and clinician-led evaluation before drawing conclusions about clinical reliability.

Acknowledgments

We thank the Jožef Stefan Institute, its Department of Intelligent Systems, the HomeDOCTOR contributors, and the ChatMED partners. We also thank medical collaborators connected with the Ljubljana Health Center, the Medical Chamber of Slovenia, and the Faculty of Medicine, University of Niš. The internship was supported by IUT de Cachan / Université Paris-Saclay and the supervision team at the Jožef Stefan Institute.

This research was supported by the European Union through the Horizon Europe programme, under the ChatMED project (Grant Agreement ID: 101159214).

References

- [1] M. Zadobovšek, P. Kocuvan, and M. Gams. 2024. HomeDOCTOR App: Integrating Medical Knowledge into GPT for Personal Health Counseling. Information Society 2024, Ljubljana, Slovenia. <https://doi.org/10.70314/is.2024.chtm.8>.
- [2] European Commission. ChatMED: Bridging Research Institutions to Catalyze Generative AI Adoption by the Health Sector in the Widening Countries. CORDIS, Grant 101159214. <https://cordis.europa.eu/project/id/101159214> (accessed 8 September 2026).
- [3] M. Jovanović, M. Baez, and F. Casati. 2021. Chatbots as conversational healthcare services. IEEE Internet Computing 25, 3, 44–51. <https://doi.org/10.1109/MIC.2020.3037151>.
- [4] K. Singhal et al. 2023. Large language models encode clinical knowledge. Nature 620, 172–180. <https://doi.org/10.1038/s41586-023-06291-2>.
- [5] F. Lechner, A. Lahnala, C. Welch, and L. Flek. 2023. Challenges of GPT-3-Based Conversational Agents for Healthcare. Proceedings of RANLP 2023, 619–630. <https://aclanthology.org/2023.ranlp-1.67/>.
- [6] J. A. Omiye, H. Gui, S. J. Rezaei, J. Zou, and R. Daneshjou. 2024. Large Language Models in Medicine: The Potentials and Pitfalls: A Narrative Review. Annals of Internal Medicine 177, 2, 210–220. <https://doi.org/10.7326/M23-2772>.
- [7] A. Hogan et al. 2021. Knowledge Graphs. ACM Computing Surveys 54, 4, Article 71. <https://doi.org/10.1145/3447772>.
- [8] P. Lewis et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems 33, 9459–9474. <https://arxiv.org/abs/2005.11401>.
- [9] World Health Organization. 2024. Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models. WHO, Geneva. ISBN 978-92-4-008475-9.
- [10] NHS. 2024. Headaches. <https://www.nhs.uk/symptoms/headaches/> (accessed 8 September 2026).
- [11] Government of the Republic of Slovenia. Klici v sili [Emergency calls]. <https://www.gov.si teme/klic-v-sili/> (accessed 8 September 2026).

Estimating the GDP-Relevant Effects of AI in Slovenian Healthcare through Reduced Sick Leave

Primož Kocuvan
primoz.kocuvan@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Matjaž Gams
matjaz.gams@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

This paper presents a scenario analysis of one potential macroeconomic channel through which AI adoption in Slovenian healthcare may affect GDP: reductions in sickness-related work absence. It explores how AI-enabled improvements in diagnostic accuracy, workflow efficiency, and patient management can reduce inefficiencies in healthcare delivery, particularly by decreasing absenteeism and associated productivity losses. We do not separately quantify administrative gains or resource-allocation effects; these remain outside the present accounting. A case calculation based on current Slovenian economic and healthcare data reports several distinct, non-additive outputs: a wage-based valuation of restored working time, a separate household-consumption channel based only on the unreplaced share of restored earnings, an alternative national-accounts sensitivity valuation based on GDP per employed person-day, and external pessimistic and optimistic productivity benchmarks adopted from prior work. The analysis therefore distinguishes rather than combines the channels and valuation perspectives through which AI-enabled reductions in sickness absence may have macroeconomic significance. The scenario results suggest that AI-enabled healthcare improvements could contribute to healthcare transformation and broader economic performance, conditional on the assumed reduction in sickness absence.

Keywords

GDP, economy, gross benefit, AI, healthcare

1 Introduction

Healthcare systems worldwide face persistent pressures arising from rising costs, workforce shortages, demographic aging, and the growing prevalence of chronic diseases, all of which strain the sustainability and performance of health services. These pressures also influence broader economic outcomes, given the close interdependence between population health, labour productivity, and national economic performance. Aging populations and chronic disease burdens increase demand for complex care, elevate healthcare expenditures, and reduce labour force participation through illness-related work absence, thereby affecting gross domestic product (GDP) and public finances. In response to these multifaceted challenges, technological innovation, particularly artificial intelligence (AI), has gained prominence as a potential lever for improving efficiency, quality, and resilience in healthcare delivery.

In Slovenia, developments over the past decade and a half illustrate that increasing financial and human resources alone has not translated into proportional improvements in healthcare performance. Between 2010 and 2024, the number of healthcare employees grew substantially, with the number of physicians increasing markedly and the overall healthcare workforce expanding accordingly, while current health expenditure rose to approximately EUR 6.26 billion, representing 9.3% of GDP. Public spending on healthcare has thus become one of the largest and fastest-growing components of government expenditure. Despite these sustained increases in funding and staffing, key performance indicators remain problematic: waiting times for specialist visits and elective procedures remain long, and the number of patients waiting beyond acceptable thresholds continues to pose a major accessibility challenge. These trends suggest that traditional policy responses based on expanding budgets and personnel within existing organizational structures have not been sufficient to improve service accessibility or efficiency and, in some respects, have coincided with a worsening user experience.

In the context of Slovenia, this divergence between rising inputs and declining outcomes raises an important economic and policy question. If AI can reduce sickness absence, decrease administrative and clinical inefficiencies, and improve decision-making in healthcare delivery, the resulting gains may materialize not only within the health system but also at the macroeconomic level. Reduced lost workdays increase effective labour supply, while only the unreplaced share of restored earnings can support additional household consumption; potential fiscal-pressure effects remain outside the present accounting. Against this background, the central question addressed in this paper is the following: what is the potential macroeconomic benefit of introducing AI into Slovenian healthcare, and to what extent could efficiency gains translate into measurable improvements in national economic output?

The paper contributes by providing a macroeconomic translation of AI-enabled reductions in sickness absence into separate, non-additive scenario outputs benchmarked against GDP. It distinguishes a wage-based valuation of restored working time, a separate induced household-consumption channel, a GDP-per-employed-person-day sensitivity valuation, and external productivity benchmarks adopted from prior Slovenian research. This separation makes explicit which quantities are calculated in the present study, which are inherited from prior work, and which should not be combined because they represent alternative valuations of the same underlying recovery in working time.

Our system HomeDOctor [5], can be perceived as a practical attempt in this direction.

2 Related work

Healthcare systems globally are confronting escalating costs, workforce shortages, and increasing demand for high-quality

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Information Society 2026, Ljubljana, Slovenia

© 2026 Copyright held by the owner/author(s).

<https://doi.org/10.70314/is.2026.aihc.4>

services, placing pressure on both the sustainability of health systems and overall economic performance. These systemic challenges—exacerbated by aging populations and rising chronic disease prevalence—have motivated interest in technological innovations that can increase efficiency and reduce costs. In this context, AI has emerged as a key innovation with potential to transform clinical workflows, enhance diagnostic accuracy, automate administrative tasks, and improve operational performance across healthcare settings [26], [7], [16]. Contemporary analyses highlight AI's capacity not only to assist clinical decision-making but also to streamline processes that traditionally consume significant time and resources, potentially improving care delivery while lowering overall system expenditure. For example, predictive analytics and automated documentation tools can free clinicians from routine tasks and reduce administrative burden, while AI-supported triage and imaging interpretation can accelerate diagnosis and limit unnecessary procedures [26], [7].

AI technologies from machine learning diagnostic models to high-dimensional predictive systems are already being explored for deployment in hospitals and primary care environments because they can enhance decision quality and optimize workflow efficiency [18]. Evidence from systematic reviews and emerging cost-effectiveness studies suggests that AI interventions have the potential to reduce unnecessary costs and improve health outcomes by expediting processes and reducing inefficiencies, although methodological challenges in economic evaluations remain significant [25]. Specifically, there is a dearth of standardized economic evaluation frameworks for AI technologies, and many existing studies suffer from heterogeneous outcomes and limited long-term cost data [25]. Yet, despite these limitations, preliminary findings indicate that AI can achieve cost savings by decreasing diagnostic errors, optimizing resource allocation, and improving care coordination.

Beyond the health services domain, broader economic analyses indicate that AI adoption can contribute positively to productivity growth at national and sectoral levels. Emerging models of the AI economy suggest that increased automation and data-driven decision support can enhance labour productivity, with potential correlations between AI intensity and macroeconomic indicators such as GDP per capita [8]. Similarly, research in economic efficiency has identified AI as a driver of innovation and productivity growth in multiple sectors, emphasizing its role in reducing production inefficiencies and enabling new value creation pathways [3]. These findings align with macroeconomic projections that underscore structural productivity gains as critical drivers of long-term economic growth, including in high-income economies where labour force growth is slowing.

The potential for AI to contribute to national economic output is further supported by empirical studies of healthcare management systems, which highlight associations between efficient health service delivery and broader socio-economic development [11]. However, measuring AI's full economic impact remains complex, as some AI-generated gains—particularly those related to quality improvements and intangible productivity enhancements—may not be fully captured within traditional GDP accounting frameworks or national statistics [4]. This underscores the difficulty of quantifying the aggregate economic value of AI innovations without comprehensive and dynamic modeling approaches.

In the context of Slovenia—an economically developed and digitally engaged EU member state—understanding the linkages

between AI innovation in healthcare and wider economic performance is especially relevant. If AI can reduce the burden of chronic illness, decrease lost workdays, and improve system efficiency, then the resulting productivity gains may propagate through household consumption, labour supply, and other channels that contribute to GDP growth. Moreover, theoretical studies on the economic implications of biased or inequitable AI systems highlight that suboptimal implementations can impose economic burdens, particularly during health emergencies, reinforcing the importance of robust governance and ethical design in pursuing AI-driven efficiency gains [13]. The present study therefore develops a conceptual and quantitative framework to estimate the pathways through which AI in healthcare could influence Slovenia's GDP, offering insights for policymakers and health system stakeholders seeking evidence-based strategies for integrating AI into health and economic planning.

3 Methodology

3.1 Data and valuation framework

We anchor the financial analysis in the calendar year 2024, as this is the most recent year for which absenteeism, earnings, household-sector, and GDP variables are available in consistent official form from NIJZ, ZZZS, and SURS [15, 10, 21, 23, 22]. The analytical framework builds on the empirical estimates reported by Smodiš et al. on the economic effects of reduced sickness absence [20]. Their analysis reports potential annual economic savings from reduced sick leave ranging from approximately EUR 17 million in a pessimistic scenario to EUR 322 million in an optimistic scenario.

The present analysis extends this absenteeism-based framework by translating reduced sickness absence into separate macroeconomic scenario valuations. First, we calculate a conservative net wage-based valuation of restored working time. This valuation uses average net earnings and the assumed reduction in working sick-leave days under AI-supported healthcare improvements. It should be interpreted as a conservative wage-based representation of recovered working time, not as a full gross value-added estimate. Second, we estimate a separate household-consumption channel. Because workers on sick leave continue to receive partial income replacement, only the unreplaced share of restored earnings is treated as additional disposable household income. This amount may support additional consumption, but it is not added to the direct productivity valuations in order to avoid double counting.

The analysis distinguishes quantities calculated in the present study from benchmark valuations adopted from prior work. The present-study calculations comprise: (i) the wage-based valuation of restored working time, denoted by W_s ; (ii) the induced household-consumption channel, denoted by C_{ind} ; and (iii) an alternative national-accounts sensitivity valuation based on GDP per employed person-day, denoted by $V_{GDP/day}(r)$ and introduced in the robustness analysis. These quantities are alternative economic representations of recovered working days and are not additive.

For comparison with prior Slovenian estimates, we denote the pessimistic and optimistic productivity benchmarks reported by Smodiš et al. [20] as $V_{LB} = 17,692,124$ EUR and $V_{UB} = 321,996,664$ EUR, respectively. Throughout this paper, the symbols V_{LB} and V_{UB} refer exclusively to these external benchmark valuations. They are not functions of the reduction parameter r used in the present study and should not be interpreted as the

Table 1: Baseline variables and benchmark valuations used in the analysis (Slovenia, 2024)

Label	Value	Comment
D_{cal}	19,477,932	Calendar sick-leave days
N_e	1,104,000	Total employment, national accounts
$\bar{w}_{net}$	EUR 1,526.02	Average monthly net earnings
$\bar{w}_{gross}$	EUR 2,394.92	Average monthly gross earnings
$\bar{d}_{cal}$	20.51 days	Average calendar sick-leave days per employee
p_{rep}	0.80	Assumed average replacement rate
c_h	0.867	Household consumption share of disposable income
Y	EUR 67,418 bn	Gross domestic product
r	0.05	Assumed reduction in working sick-leave days
D_w	16,844,940	Lost working days
V_{LB}	EUR 17,692,124	Pessimistic productivity valuation reported by Smodiš et al. [20] for their pessimistic scenario; used as an external benchmark and not derived from the present study's r scenarios.
V_{UB}	EUR 321,996,664	Optimistic productivity valuation reported by Smodiš et al. [20] for their optimistic uptake scenario; used as an external benchmark and not derived from the present study's r scenarios.

Note: D_{cal} and $\bar{d}_{cal}$ are NIJZ calendar-day indicators. D_w is the separate ZZZS working-day measure used directly in the wage-based and household-consumption calculations. N_e is the SURS annual average total employment used only in the GDP/day sensitivity valuation; it is not the denominator behind $\bar{d}_{cal}$. V_{LB} and V_{UB} are external pessimistic and optimistic productivity benchmarks adopted from Smodiš et al. [20]. They correspond to the pessimistic and optimistic scenarios defined in that study and are not calculated from the 2%, 5%, or 10% reduction scenarios used in the present sensitivity analysis. All alternative valuations are reported separately and are not additive.

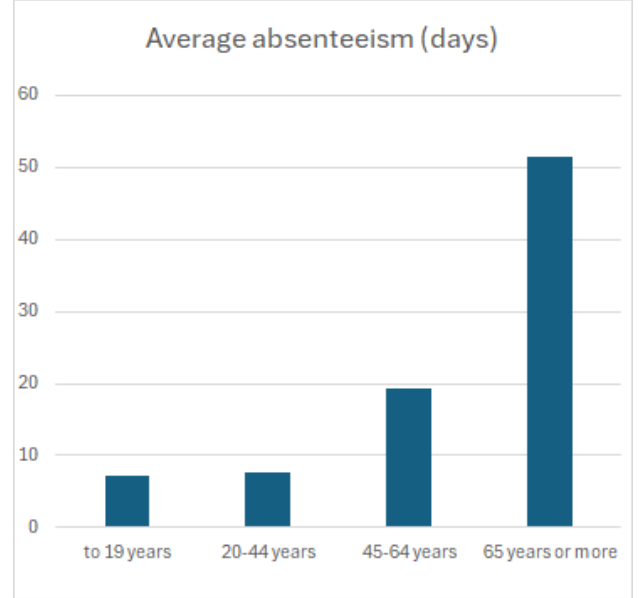
lower and upper endpoints of the present study's 2%, 5%, and 10% sensitivity scenarios.

The purpose of this framework is therefore not to construct a single aggregate benefit estimate, but to report distinct and interpretable valuations of the economic implications of reduced sickness absence while avoiding double counting.

Reduced sickness absence can affect national economic performance through both supply-side and demand-side mechanisms. On the supply side, fewer lost working days increase effective labour input. On the demand side, the unreplaced share of restored earnings can increase disposable household resources and thereby support household consumption. Potential fiscal effects, healthcare-budget effects, and general equilibrium adjustments remain outside the present accounting. The estimates should therefore be read as partial-equilibrium scenario calculations rather than as full economy-wide impact estimates.

Table 1 summarizes the key statistical inputs for Slovenia in 2024 using data from the National Institute of Public Health (NIJZ), the Health Insurance Institute of Slovenia (ZZZS), and the Statistical Office of the Republic of Slovenia (SURS) [15, 10, 21, 23, 22]. Several variables require clarification because official sources use different absenteeism measures. The variable $\bar{d}_{cal}$ denotes the NIJZ calendar-day indicator of sickness absence per employee and equals 20.51 in 2024 [15]. By contrast, D_w denotes the ZZZS working-day series, which records 16,844,940 lost working days in 2024 [10]. Because calendar-day and working-day absenteeism are based on different measurement conventions, they are not treated as interchangeable. The working-day measure D_w is used directly in the wage-based and household-consumption calculations.

The replacement-rate parameter p_{rep} represents a central simplifying assumption for income replacement during sick leave. Under Slovenian rules, compensation depends on the reason and duration of absence, and ordinary illness and injury cases are generally covered by the employer for the first 30 working days before compulsory health insurance takes over [9]. The value

**Figure 1: Average absenteeism in days per age group in 2024 [14]**

$p_{rep} = 0.80$ is therefore not interpreted as a universal legal constant, but as a central modelling assumption for the disposable-income channel. This distinction is important because a reduction in sick leave does not convert the full restored wage amount into new disposable household income. Only the unreplaced share, given by $(1 - p_{rep})$, can be treated as incremental household resources available for additional consumption.

The parameter c_h denotes the household consumption share of disposable income, derived from SURS household-sector data. It is used to approximate the part of incremental disposable income that is likely to translate into household consumption. Finally, r denotes the assumed proportional reduction in working sick-leave days. In the baseline calculation, we use $r = 0.05$, corresponding to a 5 percent reduction. This should be interpreted as a conditional scenario parameter for AI-supported healthcare improvements, such as better triage, diagnostic support, administrative automation, and improved patient routing, rather than as a measured causal effect.

The average net hourly wage is calculated as:

$$w_h = \frac{\bar{w}_{net}}{160 \text{ h}} \quad (1)$$

where 160 hours approximates a standard monthly working-time denominator. The induced household-consumption channel is then estimated as:

$$C_{ind} = D_w \times 8 \times w_h \times r \times (1 - p_{rep}) \times c_h \quad (2)$$

The wage-based valuation of restored working time is:

$$W_s = D_w \times 8 \times w_h \times r \quad (3)$$

Equation 2 estimates the separate household-consumption channel arising from the unreplaced share of recovered earnings. Equation 3 gives the wage-based valuation of restored working time. Neither quantity is added to the external productivity benchmarks V_{LB} and V_{UB} reported by Smodiš et al. [20].

3.2 Scope and interpretation of the AI-enabled mechanism

This study does not evaluate a specific AI model or clinical AI system. Instead, it provides a scenario-based health-economic assessment of one plausible AI-enabled pathway in healthcare: the reduction of sickness-related work absence through improved triage, diagnostic support, administrative automation, patient routing, and continuity of care. The assumed reduction in working sick-leave days should therefore be interpreted as a conditional modelling parameter, not as a measured causal effect of AI adoption.

The estimates reported in this paper quantify the economic order of magnitude that could arise if AI-supported healthcare improvements translated into measurable reductions in sickness absence. Empirical validation would require future pilot deployments or pre/post implementation studies linking concrete AI use cases to intermediate healthcare KPIs such as waiting times, diagnostic delay, triage throughput, administrative workload, episode duration, and return-to-work patterns.

4 BI Methodology, Evaluation Design and Robustness Checks

To strengthen the Business Intelligence (BI) contribution and improve replicability, we define the BI component as an integrated measurement and evaluation methodology rather than only as a KPI dashboard. The framework comprises four connected components: (i) a layered data architecture linking healthcare operations, AI-system utilisation, sickness-absence outcomes, and economic data; (ii) a data-linkage and KPI layer supporting repeated provider-level measurement; (iii) an evaluation and attribution design for distinguishing changes associated with AI deployment from underlying temporal and organisational variation; and (iv) scenario and sensitivity analyses translating attributable changes in sickness absence into alternative economic valuations.

4.1 BI data architecture and data linkage

The proposed Business Intelligence framework is designed as a layered architecture connecting healthcare operations with sickness-absence outcomes and macroeconomic valuation. The purpose of the architecture is to establish a traceable analytical chain from AI use in healthcare processes to observable operational outcomes, subsequent changes in sickness absence, and finally their economic valuation. The BI component should therefore not be interpreted merely as a reporting dashboard, but as a data-integration and evaluation framework supporting repeated measurement of AI-enabled healthcare effects.

The first layer consists of *operational healthcare and AI-use data*. At provider level, this layer may contain process indicators such as appointment and triage timestamps, waiting times, patient throughput, administrative workload, diagnostic or referral delays, and other routinely recorded healthcare-process variables. Where AI-supported systems are deployed, system logs should additionally record measures of AI exposure and utilisation, such as the number of AI-assisted triage interactions, documentation-support events, decision-support requests, or the proportion of relevant workflows in which the AI system was used. These variables provide an observable measure of intervention intensity rather than treating AI adoption as a simple binary condition.

The second layer contains *health and labour-outcome data*, with particular emphasis on sickness absence. Relevant variables

include the number of sickness-absence episodes, working sick-leave days, episode duration, return-to-work timing, and, where available, the distribution of short- and long-duration absence. For the national scenario calculations presented in this paper, aggregate working sick-leave data are obtained from ZZZS and NIJZ. For evaluation of an actual AI deployment, the same conceptual framework should be implemented at a finer temporal and organisational level, such as provider-month or provider-quarter observations. Where legally and technically feasible, pseudonymised episode-level linkage could provide a more detailed assessment while avoiding the use of directly identifiable personal information.

The third layer is the *economic valuation layer*. It combines sickness-absence outcomes with labour-market and national-accounts indicators obtained from SURS, including employment, wages, GDP, and household-sector variables. Changes in recovered working days can then be translated into the alternative economic valuations defined in this study.

In practical BI implementation, the data sources would be harmonised through common dimensions such as time period, healthcare provider, organisational unit, and intervention status. A provider-by-period analytical table would allow healthcare-process indicators, AI-use intensity, and sickness-absence outcomes to be compared consistently over time. National wage, employment, consumption, and GDP variables would subsequently be joined at the corresponding temporal level for economic valuation. This design separates operational measurement from macroeconomic valuation while maintaining a reproducible data lineage from source data to the final indicators.

4.2 Evaluation and attribution design

The BI architecture is intended not only to describe outcomes after AI implementation but also to support empirical evaluation of whether observed changes are associated with AI exposure. For an actual deployment, the preferred analytical grain is a provider-by-period panel, for example provider-month or provider-quarter observations. Each observation would combine healthcare-process indicators, AI-use intensity, sickness-absence outcomes, and relevant organisational covariates for the same provider and time period. Where legally and technically feasible, pseudonymised episode-level linkage could additionally be used to analyse episode duration and return-to-work outcomes while avoiding the use of directly identifiable information.

AI exposure should be measured from system logs rather than represented only by a binary implementation indicator. Depending on the application, exposure measures may include the share of eligible encounters supported by AI, the number of AI-assisted triage interactions, documentation-support utilisation, decision-support requests, or another workflow-specific utilisation measure. These variables provide the intervention measure that links the operational data layer to the evaluation model.

For deployments in which AI utilisation varies substantially after implementation, the binary treatment interaction can be supplemented or replaced by a continuous AI-exposure measure. This allows the BI framework to test whether greater utilisation is associated with larger changes in process or sickness-absence outcomes. Pre-intervention trends should be inspected to assess the plausibility of the parallel-trends assumption, and robustness checks should include alternative baseline periods, alternative outcome definitions, and, where sample size permits, event-time estimates around implementation.

If no suitable comparison provider or unit is available, an interrupted time-series design can be used as a secondary evaluation strategy by comparing post-implementation changes in outcome level and trend with the established pre-implementation trajectory. Because such a design provides weaker control for simultaneous external changes, results from this specification should be interpreted more cautiously than estimates obtained from a credible comparison-group design.

The attribution chain is therefore specified explicitly as *AI exposure* → *healthcare-process change* → *sickness-absence outcome* → *economic valuation*. Process indicators such as waiting time, triage throughput, diagnostic delay, and administrative workload represent intermediate outcomes. Working sick-leave days, episode duration, and return-to-work timing represent downstream labour outcomes. Only the estimated change in sickness absence associated with the evaluated intervention is subsequently passed to the economic valuation layer. The valuation equations therefore translate an estimated change in recovered working days into economic terms; they do not themselves establish that AI caused the underlying reduction.

This separation between measurement, attribution, and valuation provides a reproducible data lineage. Operational and AI-system logs establish intervention exposure, linked healthcare and absence data provide evaluation outcomes, the attribution model estimates the intervention-associated change, and SURS labour-market and national-accounts variables are subsequently applied to value that change economically.

4.3 BI measurement design: data sources and KPI layer

Our estimate links healthcare-side efficiency gains to separate macroeconomic valuations benchmarked against aggregate national indicators. To make this link auditable and decision-relevant, we propose a BI measurement layer that connects (a) operational healthcare process data, (b) absence and labour-market indicators, and (c) national accounts data.

Data sources - the core data sources are: (1) national absenteeism statistics (NIJZ/ZZZS), (2) labour-market and wage aggregates (SURS), and (3) healthcare process indicators where available (e.g., administrative workload proxies, waiting-time indicators, and throughput measures at provider level). For AI deployment monitoring, system logs can provide usage intensity measures (number of triage interactions, documentation automation rate, clinical decision-support utilisation), which provide the AI-exposure variables used in the attribution design described above.

KPI set - we define a minimal KPI set for evaluating AI adoption in healthcare from a BI perspective: (a) total sick-leave days and working sick-leave days (level and trend), (b) average duration per episode and distributional tail risk (long absences), (c) cost proxies (employer burden vs social insurance burden), (d) process KPIs (administrative time share, triage throughput, appointment lead times), (e) macro link KPIs (GDP per employed person-day; wage aggregates; consumption share). This KPI layer allows quarterly monitoring and supports pre/post comparisons against a defined baseline window.

4.4 Scenario and sensitivity analysis (robustness)

Because the result depends on behavioural and organisational adoption, we report a sensitivity range rather than a single point

estimate. Let D_w be the estimated annual number of working sick-leave days in Slovenia and r the proportional reduction achievable under AI-supported interventions. The recovered working days are $r \cdot D_w$.

A key modelling choice is the economic value assigned to each recovered working day. To show the sensitivity of the results to this choice, we report two alternative valuation approaches rather than interpreting them as components of a single total:

(i) Wage-based valuation: the value of a recovered working day is approximated using average net earnings.

(ii) GDP-per-employed-person-day sensitivity valuation: the value of a recovered working day is approximated by GDP per employed person-day, calculated as $Y / (N_e \times 220)$, where Y denotes annual GDP, N_e denotes annual average total employment, and 220 is the assumed number of working days per employed person per year.

The corresponding scenario valuations are:

$$V_{\text{wage}}(r) = D_w r \left(\frac{\bar{w}_{\text{net}}}{160} \times 8 \right) \quad (4)$$

$$V_{\text{GDP/day}}(r) = D_w r \left(\frac{Y}{N_e \times 220} \right). \quad (5)$$

Equations 4 and 5 apply the same assumed reduction in working sick-leave days but assign different economic values to each recovered day. They should therefore be interpreted as alternative valuations of the same underlying quantity of recovered working time, not as additive effects.

Table 2 reports the implied valuation range under three illustrative reductions in working sick-leave days (2%, 5%, 10%). These estimates quantify the sensitivity of the macroeconomic valuation to the assumed intervention strength and valuation method.

Table 2 also reconciles the principal monetary estimates reported in the paper. The EUR 64.3 million and EUR 233.8 million estimates both correspond to the present study's central 5% reduction scenario, but they use different valuation methods: average net earnings and GDP per employed person-day, respectively. The EUR 467.6 million estimate applies the GDP-per-employed-person-day method to the larger 10% reduction scenario. By contrast, the EUR 322.0 million estimate is V_{UB} , the external optimistic productivity benchmark reported by Smodiš et al. [20]; it is based on the scenario assumptions of that study and is not generated by the 5% or 10% reduction scenarios used here. Consequently, these values represent alternative scenario valuations or external benchmarks and should neither be summed nor interpreted as successive components of a single benefit estimate.

Interpretation - the table provides a robustness range for the direct valuation component. Additional healthcare-system savings should be reported separately and carefully checked for overlap to avoid double counting. A full economy-wide analysis (e.g., CGE) can further refine these ranges by modelling leakages, imports, substitution effects, and general equilibrium price adjustments.

5 Results

Applying Equation 3 to the ZZZS working-day baseline D_w under the central assumption of a 5% reduction in working sick-leave days yields a wage-based valuation of restored working time of $W_s = 64.26$ million EUR per year. This estimate values recovered working time using average net earnings and should not be interpreted as a complete measure of gross value added.

Table 2: Comparison of present-study valuations and external productivity benchmarks for reduced working sick-leave days. Estimates are alternative, scenario-specific, and non-additive.

Source	Scenario	r	Estimate (EUR m.)	Valuation method	Interpretation
This study	Conservative	2%	25.7	Wage-based	Lower-bound valuation of restored working time using average net earnings.
This study	Central	5%	64.3	Wage-based	Central wage-based valuation of restored working time.
This study	Optimistic	10%	128.5	Wage-based	Wage-based valuation under the largest assumed reduction in working sick-leave days.
This study	Conservative	2%	93.5	GDP per employed person-day	National-accounts sensitivity valuation of recovered working days.
This study	Central	5%	233.8	GDP per employed person-day	Central national-accounts sensitivity valuation of recovered working days.
This study	Optimistic	10%	467.6	GDP per employed person-day	Upper end of the GDP/day sensitivity range under a 10% reduction in working sick-leave days.
Smodiš et al. [20]	Pessimistic	–	17.7	Prior-study productivity valuation (V_{LB})	Pessimistic benchmark reported by Smodiš et al.; not derived from the present study's r scenarios.
Smodiš et al. [20]	Optimistic uptake	–	322.0	Prior-study productivity valuation (V_{UB})	Optimistic benchmark reported by Smodiš et al.; shown for comparison and not added to the valuations above.

Note: The estimates are alternative, non-additive valuations and should not be summed. The wage-based and GDP-per-employed-person-day estimates are calculated in the present study using the common 2024 baseline and illustrative reductions in working sick-leave days of 2%, 5%, and 10%. The V_{LB} and V_{UB} estimates are taken from Smodiš et al. [20] and follow the scenarios defined in that study; consequently, they are not directly equivalent to the r scenarios used in the present sensitivity analysis.

The same central scenario produces an induced household-consumption estimate of $C_{ind} = 11.14$ million EUR per year using Equation 2. This amount represents only the consumption potentially supported by the unreplaced share of restored earnings under the assumed income-replacement and household-consumption parameters. It is therefore reported as a separate demand-side transmission channel and is not added to the wage-based valuation.

The robustness analysis provides an alternative national-accounts valuation of the same recovered working days. As shown in Table 2, the 5% central scenario produces a GDP-per-employed-person-day sensitivity valuation of 233.8 million EUR, compared with the wage-based valuation of 64.3 million EUR. Under the 10% reduction scenario, the corresponding GDP-per-employed-person-day valuation rises to 467.6 million EUR. These differences arise from the valuation method and assumed reduction in working sick-leave days; they do not represent separate benefits that can be accumulated.

For external comparison, Smodiš et al. [20] report a pessimistic productivity benchmark of $V_{LB} = 17.69$ million EUR and an optimistic productivity benchmark of $V_{UB} = 321.997$ million EUR. These benchmarks follow the scenarios and valuation assumptions of that study and are not functions of the 2%, 5%, or 10% reduction parameter r used in the present analysis. In particular, V_{UB} should be interpreted as an external optimistic benchmark rather than as the overall upper endpoint of the present study's sensitivity range.

Relative to Slovenia's 2024 GDP of 67.418 billion EUR, the external Smodiš optimistic benchmark V_{UB} corresponds to approximately 0.478% of GDP. This ratio is reported only to express the magnitude of that benchmark relative to national output; it does not imply a mechanically equivalent increase in measured GDP.

Taken together, the results should therefore be interpreted as conditional and non-additive scenario outputs. The wage-based valuation, the GDP-per-employed-person-day sensitivity valuation, the external Smodiš benchmarks, and the household-consumption channel provide different perspectives on the economic significance of reduced sickness absence. They should not be summed into a single estimate without a broader economy-wide model capable of explicitly representing interactions and potential double counting.

6 Future work

What we have calculated in this study is a preliminary, static quantified assessment, based on available data and simplified assumptions about labour productivity and consumption effects. Such calculations, while informative, are necessarily sensitive to underlying economic and clinical conditions specific to a country and may not capture the full general equilibrium effects of systemic changes such as the introduction of AI in healthcare.

A more comprehensive and robust analysis would be to employ an economic simulation framework capable of capturing the interdependencies between sectors, agents, and factor markets. One of the most widely used methodologies for such purposes is the Computable General Equilibrium (CGE) model, which combines economic theory with detailed data to estimate how an entire economy responds to exogenous shocks or policy changes [2, 1]. CGE models are grounded in the principles of general equilibrium theory and consist of a system of equations that describe the behaviour of economic agents—such as households, firms, government, and the rest of the world—under the assumption of optimizing behaviour and market clearing. They allow the simulation of how changes in one part of the economy propagate through prices, production, consumption, trade, and factor allocation to yield a new equilibrium after a shock is introduced [2, 1].

In the context of our research, a CGE model would enable an integrated assessment of the economy-wide effects of introducing AI in healthcare, including both direct and indirect channels. For example, beyond the direct productivity gains from reduced sick leave, such a model could trace how increased labour supply, changes in household consumption, shifts in healthcare demand and supply, and resulting price adjustments affect other sectors, government revenues, household welfare, and macro variables such as total output and employment. CGE models have been successfully applied to diverse economic questions—including tax policy analysis in Slovenia and Slovakia, and investment impact studies—demonstrating their versatility in capturing complex economy-wide interactions [12, 19].

To operationalize a CGE approach, modern computational tools such as GAMS (General Algebraic Modeling System) can be coupled with Python via APIs to facilitate both model specification and simulation execution [6, 24]. GAMS provides a high-level

modeling environment tailored for solving large systems of non-linear equations and is widely used for formulating and solving CGE models, often with extensions such as the MPSGE library to more directly represent general equilibrium systems. Recent developments, such as GAMSPy, further enable integration with Python for data management, calibration, and post-processing [17], thereby enhancing reproducibility and accessibility for economic researchers.

Future work will involve constructing and calibrating a CGE model for Slovenia that includes health and labour market modules, permitting rigorous scenario analysis of AI adoption in healthcare. Such an extension would provide a more nuanced and policy-relevant simulation of how technological changes affect the broader economy, including sectoral reallocation, price adjustments, and welfare impacts, offering decision-makers evidence that goes beyond partial equilibrium or static calculations.

Before a full CGE extension, future work should empirically validate the healthcare-side mechanism by connecting specific AI interventions, such as AI-supported triage or documentation support, to observed changes in waiting times, diagnostic delay, sick-leave duration, and return-to-work outcomes.

7 Discussion and Conclusion

7.1 Discussion

The key contribution of this study lies not in identifying new healthcare cost savings, but in showing how AI-enabled improvements in healthcare performance can propagate through labour supply and consumption channels and be expressed as separate macroeconomic valuations. By extending absenteeism-based productivity analyses with a demand-side component and benchmarking the results against GDP, the paper provides a policy-relevant perspective that complements health-system-focused evaluations. From a macroeconomic standpoint, healthcare digitalization may therefore function as a productivity-enhancing investment rather than solely as a cost-containment measure.

Importantly, the estimated effect does not rely on radical system-wide transformation. Instead, it arises from relatively small but plausible reductions in lost workdays and administrative inefficiencies, consistent with incremental deployment of AI-based decision support tools. Such tools, typically LLM-based, can alleviate cognitive and administrative overload among medical staff, enabling clinicians to focus more effectively on core clinical tasks and patient interaction. From the patient perspective, AI-assisted systems can enhance accessibility, responsiveness, and continuity of care, particularly in routine and triage-related interactions. Partial adoption could therefore already generate non-negligible economic benefits while improving working conditions and patient experience, strengthening the case for gradual and targeted investment in AI-enabled healthcare solutions.

The analysis is subject to several limitations. First, it relies on a static, partial-equilibrium framework and does not capture general equilibrium effects, sectoral reallocation, or price adjustments in a fully dynamic economy. Second, the assumed reductions in sickness absence are based on prior empirical estimates and illustrative scenarios; actual outcomes would depend on clinical adoption, organizational change, and behavioral responses. Third, productivity and consumption effects are reported separately; combining them into a single total without a broader macroeconomic model could overstate impacts.

In addition, AI in healthcare raises important regulatory and ethical considerations. Within the European Union, deployment

is governed by a stringent regulatory environment combining the Artificial Intelligence Act (AI Act) and the General Data Protection Regulation (GDPR). High-risk AI systems are subject to conformity assessments, transparency requirements, and ongoing oversight, which may increase implementation costs and slow adoption. Data protection constraints, legal uncertainty regarding secondary use of health data, and governance requirements must therefore be considered alongside potential economic benefits.

To mitigate these risks, healthcare AI systems may be designed to minimize regulatory exposure by avoiding persistent storage of personal health data and relying on non-retentive interactions with external AI services. While such design choices may limit functionality, they illustrate how economic benefits from AI adoption could be pursued alongside regulatory compliance and risk reduction.

7.2 Conclusion

This paper estimates the potential macroeconomic significance of AI-enabled reductions in sickness absence in Slovenian healthcare by translating recovered working time into several explicitly separated economic valuations. Under the central 5% reduction scenario, the analysis yields a wage-based valuation of approximately 64.3 million EUR and a separate induced household-consumption channel of approximately 11.1 million EUR. Applying the alternative GDP-per-employed-person-day valuation to the same 5% scenario yields approximately 233.8 million EUR, while the corresponding 10% sensitivity scenario yields approximately 467.6 million EUR. For comparison, the pessimistic and optimistic productivity benchmarks reported by Smodiš et al. [20] are approximately 17.7 million EUR and 322.0 million EUR, respectively; these external benchmarks follow different scenario assumptions and are not derived from the present study's reduction parameter. All reported values are therefore conditional and non-additive. The analysis demonstrates the possible economic order of magnitude of reduced sickness absence while providing a BI framework through which future AI deployments could link observed system utilisation, healthcare-process changes, sickness-absence outcomes, and economic valuation.

Acknowledgements

This research was supported by the European Union through the Horizon Europe programme, under the ChatMED project (Grant Agreement ID: 101159214). Additional support was provided by the Slovenian Research Agency through research core funding (Grant No. P2-0209). We also thank Tadej Horvat, Filip Ivanišević, and Ivana Karasmanakis for helping and providing expertise in the technical and medical field.

References

- [1] Christoph Böhringer and Andreas Löschel. 2006. Computable general equilibrium models for sustainability impact assessment: status quo and prospects. (2006). doi:<https://doi.org/10.1016/j.ecolecon.2006.03.006>.
- [2] 2016. Computable general equilibrium modelling: introduction. Overview of CGE models and their use in policy analysis. (2016). <https://www.gov.scot/publications/cge-modelling-introduction/>.
- [3] Mirjana Dejanović. 2025. The role of artificial intelligence in enhancing economic efficiency and innovation. *Forum for Economic and Financial Studies*, 3, 1, 1541. doi:10.59400/feefs1541.
- [4] European Journal of Computer Science and Information Technology. 2025. Artificial intelligence in healthcare economics: transforming cost efficiency, member experience, and preventative care. *European Journal of Computer Science and Information Technology*. doi:<https://doi.org/10.37745/ejcsit.2013/vol13n9104114>.

- [5] Matjaž Gams, Tadej Horvat, Žiga Kolar, Primož Kocuvan, Kostadin Mishev, and Monika Simjanoska Misheva. 2025. Evaluating a nationally localized ai chatbot for personalized primary care guidance: insights from the home-doctor deployment in slovenia. *Healthcare*, 13, 15. doi:10.3390/healthcare13151843.
- [6] GAMS Development Corporation. 2026. *GAMS Documentation Center*. Release 53.2.0; accessed 2026-03-09. Fairfax, VA, USA, (Mar. 2026). <https://www.gams.com/latest/docs/>.
- [7] S. Ghabri. 2025. Using ai in the economic evaluation of ai-based health technologies. *Pharmacoeconomics*, 43, 597–600. doi:10.1007/s40273-025-01496-x.
- [8] D. Gondauri. 2025. Deciphering the ai economy: a mathematical model perspective. *arXiv*. arXiv: 2505.11991 [econ. GN]. doi:10.48550/arXiv.2505.11991.
- [9] Health Insurance Institute of Slovenia (ZZZS). 2026. Compensation for loss of salary during temporary absence from work. Retrieved Mar. 28, 2026 from <https://www.zzs.si/en/compulsory-health-insurance/the-extent-of-rights-deriving-from-compulsory-health-insurance/cash-benefits/compensation-for-loss-of-salary-during-temporary-absence-from-work/>.
- [10] Health Insurance Institute of Slovenia (ZZZS). 2025. Letno poročilo zzzs 2024. Retrieved Mar. 28, 2026 from https://www.zzs.si/fileadmin/user_upload/slike/o_zzs/letno_porocilo_zzs_2024.pdf.
- [11] D. Ilić, S. Andžić, and D. Erić. 2025. Healthcare management as a key component of sustainable development: a comparative analysis of serbia and slovenia. *Economics of Agriculture*, 72, 3, 1057–1071. doi:10.59267/ekoPolj250310571.
- [12] V. Mitkova and M. Janošová. 2019. The tax burden CGE analysis for slovakia and slovenia. *Naše gospodarstvo/Our Economy*, 65, 4, 35–46. doi:10.2478/ngoe-2019-0018.
- [13] 2024. Modelling of economic implications of bias in ai-powered health emergency response systems. *arXiv*. Theoretical framework on economic risks of biased AI systems in healthcare emergencies. arXiv: 2410.20229 [econ. GN]. doi:10.48550/arXiv.2410.20229.
- [14] National Institute of Public Health (NIJZ). 2025. Prva objava: Bolniška odsotnost v letu 2024. Tech. rep. Statistical report on sickness absence indicators in Slovenia for 2024 (15 May 2025 publication). National Institute of Public Health, Slovenia, (May 2025). https://nijz.si/wp-content/uploads/2025/05/Prva-objava_BS_podatki2024-1.pdf.
- [15] National Institute of Public Health of Slovenia (NIJZ). 2025. Zdravstveni statistični letopis slovenije 2024: 2.6. bolniška odsotnost. Retrieved Mar. 28, 2026 from https://nijz.si/wp-content/uploads/2025/12/2.6_Bolniska_odsotnost_2024_Z-1.pdf.
- [16] U. Pagallo, S. O'Sullivan, N. Nevejans, et al. 2023. The underuse of ai in the health sector: opportunity costs, success stories, risks and recommendations. *Health and Technology*, 14, 1–14. doi:10.1007/s12553-023-00806-7.
- [17] 2025. Paretoecon cge models (gamspy). (2025). <https://github.com/ParetoEcon/CGE-Models>.
- [18] S. Sedlak and M. Jelenc. 2025. The potential of artificial intelligence in reducing the economic burden of chronic diseases. *European Journal of Public Health*, 35, Suppl. 4. Case study with estimates for Slovenia. doi:10.1093/eurpub/ckaf161.1079.
- [19] M. Škare and S. Stjepanović. 2011. Computable general equilibrium model for croatian economy. *Ekonomika istraživanja*. Example of CGE modelling for an economy using national data. <https://hrcak.srce.hr/70509>.
- [20] Rok Smodiš, Filip Ivanišević, Ivana Karasmanakis, Primož Kocuvan, Tadej Horvat, and Matjaž Gams. 2025. (r)evolucija slovenskega zdravstva z uporabo metod umetne inteligence: ocena finančnih koristi na primeru sistema home-doctor. Slovenian. *Uporabna informatika*, 33, 4. doi:10.31449/upinf.262.
- [21] Statistical Office of the Republic of Slovenia (SURS). 2025. Earnings of persons in paid employment by legal persons, december 2024. Retrieved Mar. 28, 2026 from <https://www.stat.si/StatWeb/en/News/Index/13457>.
- [22] Statistical Office of the Republic of Slovenia (SURS). 2025. Gross domestic product, other aggregates of national accounts and employment, 2024. Retrieved Mar. 28, 2026 from <https://www.stat.si/StatWeb/en/News/Index/13795>.
- [23] Statistical Office of the Republic of Slovenia (SURS). 2025. Nonfinancial sector accounts, 2024. Retrieved Mar. 28, 2026 from <https://www.stat.si/statweb/en/News/Index/13860>.
- [24] 2023. Supercharging gams models by linking to artificial intelligence. (2023). <https://www.gams.com/blog/2023/06/supercharging-gams-models-by-linking-to-artificial-intelligence-heuristic-search-and-machine-learning/>.
- [25] J. Wolff, J. Pauling, A. Keck, and J. Baumbach. 2020. The economic impact of artificial intelligence in health care: systematic review. *Journal of Medical Internet Research*, 22, 2, e16866. doi:10.2196/16866.
- [26] Geyao Zhao. 2025. Artificial intelligence in healthcare: clinical applications and economic implications. *MedScien*, 1, 5. doi:10.61173/1b844w95.

Overcoming Bias and Building Trust: Design of a Prospective Evaluation of Explainable AI in Melanoma Detection Across Diverse Skin Phototypes

Anna-Katharina Herke
Applied Artificial Intelligence

Alma Mater Europaea

Maribor, Slovenia

anna-katharina.herke@almamater.si

Abstract / Povzetek

Deep-learning systems match dermatologists at classifying dermoscopic images, yet remain largely absent from routine melanoma care. Two coupled obstacles explain the gap: algorithmic bias, because public training archives under-represent Fitzpatrick skin types (FST) V–VI, and the “black-box” problem, because opaque probability scores give clinicians no verifiable rationale. Prospective evidence on whether explainable AI (XAI) improves clinical decisions equitably across phototypes is lacking. This paper is a study protocol: we present the design of a prospective two-phase multi-reader multi-case (MRMC) study in which 40 dermatologists diagnose 120 histopathology-verified images—balanced by malignancy and skin type—first unaided, then randomized to a black-box probability score (Arm A) or the same score plus a saliency heatmap and plausibility check (Arm B). We specify the model, instrument, and analysis: Δ AUC via the Obuchowski–Rockette method from ordinal confidence ratings, an appropriate-reliance analysis distinguishing calibrated from indiscriminate trust, subgroup sensitivity, and expertise as a moderator. No data have yet been collected.

Keywords / Ključne besede

Melanoma, Artificial Intelligence, Explainable AI, Algorithmic Bias, Dermoscopy, Skin of Color

1 Introduction

Melanoma causes most skin-cancer deaths, and prognosis depends heavily on early detection [1]. Diagnosis rests on visual inspection and dermoscopy, whose accuracy varies substantially with clinician experience [2]. Deep convolutional neural networks (CNNs) have matched or exceeded board-

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Information Society 2026, 5–9 October 2026, Ljubljana, Slovenia

© 2026 Copyright held by the owner/author(s).

<http://doi.org/10.70314/is.2026.aihc.27>

certified dermatologists at classifying dermoscopic images [3, 4], yet AI has not become a routine “second reader.”

Two interrelated problems account for this stagnation. First, algorithmic bias: models are overwhelmingly trained on fair-skinned populations (Fitzpatrick skin types I–IV; FST), and audits of archives such as ISIC document severe under-representation of FST V–VI [5, 6, 7], which translates into measurably lower accuracy for patients of color [8]. Second, the “black-box” problem: accountable clinicians are reluctant to act on a bare probability score [9], and opaque models can latch onto confounders—surgical ink, rulers, or skin tone—rather than genuine pathology [10].

Explainable AI (XAI), typically delivered as saliency heatmaps, is widely proposed as the bridge to clinical trust [9], yet bias and explainability have been studied largely in isolation, and prospective evidence on how XAI shapes clinical decisions is scarce [11]; whether a heatmap improves judgment specifically on FST V–VI lesions, where models underperform, is unknown. We therefore present the design of a prospective reader study asking whether XAI support improves dermatologists’ diagnostic accuracy, appropriately calibrated reliance, and confidence across skin types, compared with a black-box aid or no aid.

2 Related Work

Research on AI for skin-lesion classification has moved through three phases. Around 2017–2019, CNNs reached expert-level performance, largely in silico [3, 4]. Scrutiny then turned to the data: Adamson and Smith warned that machine learning could entrench dermatologic health-care disparities [5]; analyses of ISIC and the Fitzpatrick 17k set confirmed under-representation of FST V–VI [6, 7]; a systematic audit showed a substantial accuracy drop on darker skin [8]; and Winkler et al. showed that surgical skin markings can spuriously raise a CNN’s melanoma score [10].

The current phase asks how AI should support clinicians. Tschandl et al. found that collaboration improves accuracy but can mislead less-experienced readers when the AI errs [12], and Gaube et al. showed clinicians are susceptible to incorrect advice from decision aids [13]. Chanda et al. recently

demonstrated that domain-specific, dermatologist-like explanations increase clinicians' trust, confidence, and accuracy in melanoma diagnosis [14], though without stratifying effects by phototype, while a cautionary line of work argues that current explanations offer “false hope,” indicating where but not why [15]. Whether XAI can offset bias at the point of care—improving judgment more on FST V–VI lesions than on lighter skin—remains untested; this study targets that intersection.

3 Study Design and Methods

3.1 Design and hypotheses

We adopt a multi-reader multi-case (MRMC) protocol, the standard for evaluating diagnostic imaging technologies [16], as a prospective two-phase paired comparison with parallel aided arms: each reader diagnoses the same cases first unaided and then AI-aided, so the “AI effect” is estimated within readers [11, 12] (Figure 1). The study tests four hypotheses: H1 (Accuracy). XAI-aided reading yields higher AUC than black-box-aided or unaided reading; the gain is larger for challenging cases and FST V–VI lesions (H1.1). Challenging cases are defined a priori as cases misdiagnosed by at least 50% of readers in the unaided phase; this operational definition will be fixed in the pre-registered protocol before any Phase 2 data are analyzed. H2 (Bias). The standalone model shows lower sensitivity on FST V–VI than on I–IV. H3 (Trust and appropriate reliance). Readers rely on XAI output more appropriately—accepting more correct and rejecting more incorrect AI advice—and report higher trust and confidence; H3 concerns calibration of reliance, not its maximization. H4 (Usefulness). Perceived usefulness is higher with XAI than with standard output.

The single confirmatory contrast is the within-reader difference between aided and unaided AUC, pooled across both arms (H1), tested at two-sided $\alpha = 0.05$. All other comparisons—including the between-arm XAI vs. black-box contrast, H1.1, H2–H4, and all subgroup and interaction analyses—are exploratory and are reported as effect sizes with confidence intervals.

3.2 Readers

Readers are board-certified dermatologists and final-year residents who routinely use dermoscopy; readers with color-vision deficiency are excluded. The cohort is stratified by experience because automation bias is more pronounced among novices [12, 13]; recruitment is by convenience sampling. Sample size is driven by the primary within-reader contrast of aided vs. unaided AUC (H1). Assuming an unaided AUC of 0.80, a target difference of 0.05, 120 cases, and variance components typical of dermoscopic MRMC studies (reader variance $\sigma^2_R \approx 0.0011$, reader-by-modality variance $\sigma^2_{RM} \approx 0.0006$), the Obuchowski–Rockette framework indicates that about 20 readers per arm yield $\geq 80\%$ power at two-sided $\alpha = 0.05$, hence $n = 40$ [16]. These assumptions will be fixed in the pre-registered protocol. Because allocation to arms is between subjects, the black-box vs. XAI comparison is powered only for larger effects;

subgroup (FST, case difficulty) and expertise-interaction analyses are secondary and exploratory, reported as effect sizes with confidence intervals.

3.3 Cases

The test set comprises 120 dermoscopic images curated from anonymized sources such as the ISIC archive and verified institutional databases [7], balanced 60 malignant (melanoma) to 60 benign and 60 FST I–IV to 60 FST V–VI (30 images per malignancy $\times$ skin-type cell). Histopathology is the reference standard for all malignant and ambiguous cases; benign lesions without biopsy are accepted only with documented clinical and dermoscopic stability over at least 12 months of follow-up, so that label noise on benign cases does not distort AUC or specificity.

Because archive metadata are unreliable for skin type, the Fitzpatrick type of every test image is assigned by three board-certified dermatologists rating independently under standardized viewing conditions; disagreements are resolved by majority, inter-rater agreement is reported with quadratically weighted κ , and images without at least two concordant ratings are excluded. The individual typology angle (ITA), computed from lesion-free skin regions, is reported as a secondary, objective check on the categorical labels. The same labeling procedure is applied to the institutional training images described in Section 3.4.

3.4 AI model and explanation method

Both arms use the same CNN: an EfficientNet-B4 backbone pre-trained on ImageNet and fine-tuned for binary malignancy classification on ISIC images supplemented with verified institutional FST V–VI images from collaborating dermatology departments [6, 7]. The participating institutions, image counts per source, verification procedures (histopathology and recorded skin type), and the training set's full FST composition will be listed in the pre-registered protocol. Probabilities are calibrated by temperature scaling on a held-out split, with reliability curves and Brier score reported so that the displayed “risk %” is interpretable. The model is frozen before the first reader session; its standalone AUC, sensitivity, and specificity on the test set are computed separately for FST I–IV and V–VI (required for H2). Arm B explanations are gradient-weighted class activation maps (Grad-CAM) on the final convolutional block, rendered as a semi-transparent overlay [9]. To exclude training/test contamination, test images are withheld at the patient level—every image sharing a patient or lesion identifier with a test case is removed from training and validation—and near-duplicate detection (perceptual hashing plus embedding-based similarity) is applied across sources [7].

3.5 Platform, conditions, and ethics

Reading is performed on a secure web platform. In Phase 1 readers diagnose all 120 images unaided. After a washout of at least two weeks, Phase 2 re-presents the images in a new random order with AI support; readers are randomized to Arm

A (malignancy probability only, e.g., “risk 85%”) or Arm B (probability plus the Grad-CAM heatmap) [9]. Allocation is between subjects—1:1, stratified by expertise—so each reader experiences exactly one AI condition and never crosses between arms (Figure 1). In addition, a random subset of 20 images is re-read unaided at the start of Phase 2 by every reader, providing an internal within-reader estimate of practice effects against which the aided–unaided differences can be benchmarked. To guard against “right for the wrong reason” failures [10], Arm B readers also rate each heatmap’s clinical plausibility—whether it highlights established dermoscopic criteria or confounders; correlating plausibility with trust probes the mechanism of AI-influenced judgment [11].

The full protocol, including all hypotheses, definitions, and analysis code stubs, will be pre-registered on the Open Science Framework (OSF) before reader recruitment begins. The study will be submitted for approval to the competent institutional ethics committee; all readers will provide written informed consent, and all images are anonymized and used in accordance with their respective licenses and data-use agreements.

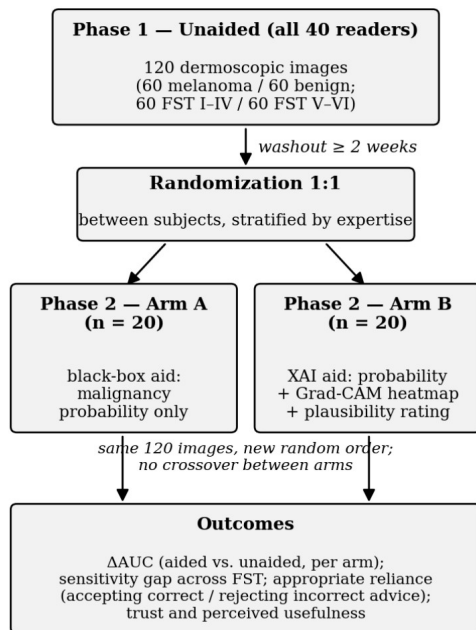


Figure 1: Study workflow. All readers complete the unaided phase; after washout they are randomized 1:1 (stratified by expertise) to exactly one AI-aided arm—no reader sees both AI conditions. A 20-image unaided subset is repeated in Phase 2 to estimate practice effects.

3.6 Instrument

For every case a standardized digital form captures constructs from signal detection theory and the Technology Acceptance Model [16, 17]: a binary malignancy decision, diagnostic confidence (1–10), and a management choice (dismiss / follow-up / biopsy). The confidence rating doubles as the ordinal score

from which reader-level ROC curves are constructed (see Analysis). Phase 2 additionally records perceived usefulness and trust on 5-point Likert scales adapted from Davis [17] and logs the AI output shown, so every aided decision can be classified as accepting or overriding correct or incorrect advice. Readers are not shown clinical metadata (age, sex, lesion site), have no per-case time limit, and receive a fixed honorarium; these conditions will be stated in the pre-registered protocol.

3.7 Analysis

Analyses will be performed in R or Python. The primary endpoint is the area under the ROC curve (AUC). For each reader and condition, the ROC curve is constructed from the 1–10 confidence ratings treated as an ordinal rating scale [16]: sweeping the rating threshold traces the empirical curve (trapezoidal AUC; binormal fit as a sensitivity analysis), while the binary decision defines the reader’s operating point, reported as sensitivity and specificity. Reader-level AUCs are pooled and the aided–unaided change tested with the Obuchowski–Rockette model for MRMC data, using its Dorfman–Berbaum–Metz-equivalent implementation in the RJAfroc package [16]; the repeated unaided subset (Section 3.5) yields a within-reader estimate of the practice effect that is subtracted from the aided–unaided difference as a sensitivity analysis.

H2 compares the standalone model’s sensitivity and specificity between the skin-type subsets; because these are disjoint, unpaired tests are used (χ^2 , or Fisher’s exact where cell counts are small), with Newcombe hybrid-score confidence intervals for the difference. With 30 malignant cases per subgroup, the minimum detectable sensitivity difference at 80% power is approximately 0.30, so H2 is reported primarily as an effect size with confidence interval rather than as a p-value. For H3 we separate self-reported trust from appropriate reliance: Phase 1 supplies each reader’s unaided decision on every case, so each aided decision is classified by whether the reader followed the AI and whether the AI was correct; appropriate reliance—the rates of accepting correct and rejecting incorrect advice—is compared between arms, overall and within FST subgroups [13]. Trust and usefulness ratings (H3, H4) are compared with Mann–Whitney U tests, inter-reader agreement with Fleiss’ κ , and a linear mixed-effects model (expertise interaction, reader random effects) tests moderation by experience. Only H1 is confirmatory; all other tests are descriptive, with Holm-adjusted p-values reported for transparency.

4 Expected Results and Contribution

We expect to confirm the standalone sensitivity penalty on FST V–VI (H2) [5, 8]; that XAI-aided reading outperforms black-box-aided and unaided reading (H1), with the largest gains on darker skin and difficult cases (H1.1), indicating that XAI can help mitigate—not merely reveal—bias; and higher perceived usefulness and trust (H3, H4) [11, 12]. We emphasize that more trust is not itself the goal: reliance is beneficial only when the model is right, which on under-represented skin types it often is

not [13, 15]; the clinically meaningful outcome is appropriate reliance. The study would complement recent prospective evidence on XAI in melanoma diagnosis [14] with, to our knowledge, the first prospective quantification of XAI's effect across the full Fitzpatrick spectrum; methodologically, it links signal detection theory with the Technology Acceptance Model; practically, it would inform guidance on equitable AI use and clinician training in its critical use [8, 9].

5 Limitations

The study is *in silico*, excluding patient history and physical examination, so measured trust may differ from trust under real liability [11, 12]. The unaided phase always precedes the aided phase, so aided–unaided differences may partly reflect practice effects: the washout attenuates case memory and the repeated 20-image unaided subset in Phase 2 provides an internal estimate of the practice effect, but phase order is not counterbalanced. The plausibility rating exists only in Arm B, so heatmap and rating-task effects cannot be fully separated. Biopsy-verified FST V–VI melanoma images are scarce: the balanced test set is a target, and if 30 malignant V–VI cases cannot be sourced, the set will be completed in unbalanced form with weighted analyses, at some cost to subgroup power [6, 8]. Voluntary participation may over-sample “early adopters” [17]; there is no standard metric of heatmap quality, so plausibility remains partly subjective [9, 15]; and the protocol measures immediate effects, not long-term learning or accumulating automation bias [12, 13].

6 Conclusion

Bias and opacity, not raw accuracy, are the binding constraints on clinical melanoma AI, and they have mostly been studied apart. The two-phase MRMCM study described here tests them together, asking whether explainable, skin-type-aware AI can raise dermatologists' accuracy, calibrated reliance, and confidence equitably across phototypes. Data collection is the next step; we welcome collaboration on reader recruitment and on assembling biopsy-verified images of skin of color.

References / Literatura

- [1] Dirk Schadendorf et al. 2018. Melanoma. *The Lancet* 392, 10151 (2018), 971–984. DOI: [https://doi.org/10.1016/S0140-6736\(18\)31559-9](https://doi.org/10.1016/S0140-6736(18)31559-9).
- [2] M. E. Vestergaard et al. 2008. Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. *British Journal of Dermatology* 159, 3 (2008), 669–676. DOI: <https://doi.org/10.1111/j.1365-2133.2008.08713.x>.
- [3] Andre Esteve et al. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 7639 (2017), 115–118. DOI: <https://doi.org/10.1038/nature21056>.
- [4] H. A. Haenssle et al. 2018. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology* 29, 8 (2018), 1836–1842. DOI: <https://doi.org/10.1093/annonc/mdy166>.
- [5] Adewole S. Adamson and Avery Smith. 2018. Machine learning and health care disparities in dermatology. *JAMA Dermatology* 154, 11 (2018), 1247–1248. DOI: <https://doi.org/10.1001/jamadermatol.2018.2348>.
- [6] Matthew Groh et al. 2021. Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 1820–1828.
- [7] Veronica Rotemberg et al. 2021. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data* 8, 1 (2021), 34. DOI: <https://doi.org/10.1038/s41597-021-00815-z>.
- [8] Roxana Daneshjou et al. 2022. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances* 8, 32 (2022), eabq6147. DOI: <https://doi.org/10.1126/sciadv.abq6147>.
- [9] Katja Hauser et al. 2022. Explainable artificial intelligence in skin cancer recognition: A systematic review. *European Journal of Cancer* 167 (2022), 54–69. DOI: <https://doi.org/10.1016/j.ejca.2022.02.025>.
- [10] Julia K. Winkler et al. 2019. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatology* 155, 10 (2019), 1135–1141. DOI: <https://doi.org/10.1001/jamadermatol.2019.1735>.
- [11] Isabelle Krakowski et al. 2024. Human-AI interaction in skin cancer diagnosis: a systematic review and meta-analysis. *npj Digital Medicine* 7, 1 (2024), 78. DOI: <https://doi.org/10.1038/s41746-024-01031-w>.
- [12] Philipp Tschandl et al. 2020. Human–computer collaboration for skin cancer recognition. *Nature Medicine* 26, 8 (2020), 1229–1234. DOI: <https://doi.org/10.1038/s41591-020-0942-0>.
- [13] Susanne Gaube et al. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digital Medicine* 4, 1 (2021), 31. DOI: <https://doi.org/10.1038/s41746-021-00385-9>.
- [14] Tirtha Chanda et al. 2024. Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma. *Nature Communications* 15, 1 (2024), 524. DOI: <https://doi.org/10.1038/s41467-023-43095-4>.
- [15] Marzyeh Ghassemi et al. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* 3, 11 (2021), e745–e750. DOI: [https://doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9).
- [16] Nancy A. Obuchowski and Jennifer A. Bullen. 2018. Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine. *Physics in Medicine and Biology* 63, 7 (2018), 07TR01. DOI: <https://doi.org/10.1088/1361-6560/aab4b1>.
- [17] Fred D. Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly* 13, 3 (1989), 319–340. DOI: <https://doi.org/10.2307/249008>.

Detecting Spin in Clinical Trials with Large Language Models

Tjaš Ajdovec
University of Ljubljana, Faculty of
Computer and Information Science
Slovenia
tjas.ajdovec@gmail.com

Marko Robnik-Šikonja
University of Ljubljana, Faculty of
Computer and Information Science
Slovenia
marko.robnik@fri.uni-lj.si

Simon Šuster
Independent researcher
Ljubljana, Slovenia
sim.suster@gmail.com

Abstract

Spin in clinical trials includes reporting practices that distort the presentation of results. This is particularly critical in medicine, where spin is present in more than 50% of randomized controlled trials that fail to reach statistical significance. The comparison of primary and reported outcomes is crucial for detecting several types of spin, including outcome switching. We used 300 pairs of outcomes labeled with semantic similarity to develop a system for automatic detection of outcome switching. We evaluated baseline text similarity models and open-source LLMs using generated similarity scores and the Youden index to determine the classification threshold. The proposed approach involves prompt engineering, classification based on token probabilities, and majority voting for the final decision. The results on the test set of 2,496 examples with an F_1 score of 0.78 and an accuracy of 0.90 outperform baseline text similarity models but trail behind fine-tuned versions of BERT. We used LLMs to generate natural language explanations for the classified instances and manually assessed their quality.

Keywords

spin, clinical trials, natural language processing, large language models, sentence similarity.

1 Introduction

The phenomenon of **spin** covers several reporting practices that present study results in a distorted way, usually to the benefit of the authors. It is a long-known problem, especially in randomized controlled trials (RCTs), where reporting rules are clearly defined, yet incorrect reporting still occurs. This leads to redundant or misdirected follow-up research [4]. Several tools exist to help authors and readers automatically detect this kind of incorrect reporting.

Large language models (LLMs) have recently become capable of solving many natural language processing (NLP) tasks **without task-specific training** [13]. Existing spin detection tools rely on dedicated models trained on expensive, manually labeled data, and to our knowledge none of them make use of LLMs. We investigate whether LLMs can be used to classify pairs of primary and reported clinical trial outcomes as similar or different. This is a core subtask needed for detecting a type of spin called **outcome switching** - compromised reporting of primary and secondary outcomes in the results.

We addressed the following **research questions** in the context of detecting outcome switching in RCT reports:

- (1) How successful are LLMs, without additional training, at classifying pairs of clinical trial outcomes as similar or different?
- (2) How do the results and behavior of different LLMs differ with respect to generality and comprehension ability?
- (3) Can LLMs generate meaningful, human-readable explanations for the already classified examples?

We report the work in six sections. Section 2 outlines the related work. Section 3 presents the methodology and Section 4 the results. The discussion is contained in Section 5, and the conclusions follow in Section 6.¹

2 Related Work

Spin is defined as the distorted presentation of study results, which misleads readers about how effective a treatment actually is [3]. It may be intentional or simply a consequence of negligence [1]. In RCTs, the result of interest is referred to as the outcome of the trial. Since readers often judge a study by its abstract alone, it is especially prone to spin: manual audits report spin in **47–71% of clinical trial abstracts** that fail to confirm their hypothesis, depending on the biomedical field [6, 11, 14, 15].

According to Yavchitz et al. [16], spin falls into three main groups:

- **Misleading reporting** of results: inconsistent or incomplete reporting, such as omitting an outcome that did not confirm the hypothesis, adding a new outcome that does, or changing the order of primary and secondary outcomes.
- **Inappropriate interpretation** of results: incorrect conclusions drawn from even correctly reported data, such as unjustified causal language or claims not based on the actual statistical results.
- **Inappropriate extrapolation** of results: unjustified generalization, such as extending treatment effects to a different population.

Automatic tools for detecting spin can benefit both authors in writing, as well as readers and reviewers in critically evaluating articles. **ExaCT** (2010) uses statistical models and regular expressions to extract key trial information, such as sample size, from RCT abstracts into a table [8]. **RobotReviewer** (2015) uses a machine learning model trained on 12,800 labeled clinical trial reports to assess their level of bias [12]. **DeSpin** (2020) uses fine-tuned pre-trained language models (PLMs) to detect several types of spin in RCTs, by breaking the task down into extraction, detection, and classification sub-tasks [9]. One such example is **comparing the primary and reported outcomes**, needed to detect inconsistent reporting, specifically outcome switching. This is also the task addressed in this paper, using large language models, presuming that outcomes were already extracted.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Information Society 2026, Ljubljana, Slovenia

© 2026 Copyright held by the owner/author(s).

<https://doi.org/10.70314/is.2026.aihc.42>

¹This research was conducted in academic year 2024/25 as part of a bachelor's thesis.

3 Methodology

3.1 Dataset

Comparing primary and reported outcomes can be framed as a directed semantic similarity task in the context of biomedicine. This is crucial for automatically checking whether the stated objectives of a clinical trial are correctly reported and met. Koroleva et al. [10] built a dataset of such outcome pairs from 3,938 RCT articles available on PubMed Central². One human annotator, after consulting domain experts, labeled **3,043 pairs as similar (label 1) or different (label 0)**. We then deduplicated the pairs, leaving 2,796. This dataset was imbalanced, with 21.8% of pairs labeled as similar and 78.2% as different, as a consequence of the collection methodology. An example pair labeled as similar is the primary outcome *log HbA1c*, reported as *glycaemic control*. This is a broader term, and recognizing it as a match requires domain knowledge of (blood sugar) measurements. An example of a negative pair is the primary outcome *blood pressure*, reported as *systolic BP*, which is a more specific term.

From the deduplicated pairs, we randomly sampled 300 for our own experiments: 100 for model tuning and 200 for validation. The remaining 2,496 pairs were used for the final testing. The tuning, validation and test sets contained 31.0%, 21.5% and 21.4% similar pairs respectively, compared to 21.8% overall. Due to unstratified sampling, the tuning set is biased toward similar pairs, relative to the whole deduplicated dataset.

3.2 Metrics

When comparing texts, most models produce a continuous similarity score on the interval $[0, 1]$ rather than a direct label, so a threshold is needed to turn this score into a binary decision. We used the **Youden index** to determine this threshold automatically, rather than setting it manually, which would be time-consuming and prone to bias [7]. The Youden index J is defined at each point on the ROC curve as:

$$J = \text{sensitivity} + \text{specificity} - 1 = \frac{TP}{TP + FN} + \frac{TN}{TN + FP} - 1 \quad (1)$$

For each model, we selected the threshold that maximizes J . Unlike accuracy or F_1 score, J treats both classes equally, making it appropriate for imbalanced datasets.

We reported classification accuracy, positive-class precision, recall, and F_1 score as evaluation metrics, with F_1 as the main one.

3.3 Baseline Models

Koroleva et al. [10] divided baseline text similarity models into the following groups, each representing sentences differently:

- (1) **String measures** treat sentences as sequences of characters. We evaluated the Levenshtein distance and sequence similarity based on the Ratcliff-Obershelp distance.
- (2) **Lexical measures** operate on words or word parts. We evaluated lemma similarity and stem similarity.
- (3) **Vector-based measures** embed text into a vector space and compare vectors using distances such as cosine distance. We evaluated spaCy and Word2Vec embeddings.
- (4) **Ontology-based measures** represent language as a hierarchy of concepts and compute distances between them. We evaluated path distance, Leacock-Chodorow similarity, and Wu-Palmer similarity.

- (5) **Deep learning models** use contextual representations from large pre-trained encoders. We evaluated SciBERT and BioBERT as sentence embedding models, as well as the **all-MiniLM-L12** sentence transformer³.

3.4 Large Language Models

To make our approach widely accessible, we selected **open-source LLMs** available on HuggingFace and ran them locally with their transformers library. Local inference avoids API costs and latency, while keeping data private in the sensitive medical domain. We ran experiments on a single NVIDIA Quadro RTX 5000 GPU with 16 GB of VRAM, which limited the model size to up to **8B parameters**. We considered only instruction fine-tuned variants, as these follow user prompts rather than simply continuing text.

We evaluated four LLMs with different characteristics:

- **OLMo-7B-Instruct**⁴ was chosen for its complete transparency. The architecture, training data, and procedure are publicly documented and replicable. The model was trained on 2.46T tokens, with over 2% of scientific articles.
- **Mistral-7B-Instruct-v0.2**⁵ is a general-purpose LLM with slightly better benchmark results than OLMo.
- **BioMistral-7B-DARE**⁶ is a variant of Mistral further trained on PubMed Central literature, giving it stronger biomedical terminology knowledge at the cost of generality.
- **Llama-3-8B-Instruct**⁷ was reportedly trained on 15T tokens and achieves the best reasoning scores.

We first tested LLMs as binary similarity classifiers by inserting a pair of outcomes into a prompt template and extracting a Yes or No prediction from the generated text:

```
<|endoftext|><|user|>
Are the following sentences semantically similar?
First sentence: {out1}
Second sentence: {out2}
Answer:
<|assistant|>
```

Algorithm 1: Classification via LLM token probabilities

Input: Problem with a finite set of answers $a_1, a_2, \dots, a_n$

- 1 Construct a prompt that expects an answer to the problem;
- 2 Assign each possible answer to a token $t_1, t_2, \dots, t_n$;
- 3 Run the model to generate the first token;
- 4 Extract the output value s_i for each token t_i ;
- 5 Compute output probabilities $p_i = \text{softmax}(s_i)$ for each token t_i ;
- 6 Find the index of the highest probability
 $h = \arg \max(p_1, p_2, \dots, p_n)$;

Output: Answer with the highest output probability a_h

This text-based approach has several limitations related to LLM answer format and extraction. We therefore adopted a **probability-based approach**, outlined in Algorithm 1 [17]. It

²<https://doi.org/10.5281/zenodo.3234827>

³<https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2>

⁴<https://huggingface.co/allenai/OLMo-7B-Instruct-hf>

⁵<https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2>

⁶<https://huggingface.co/BioMistral/BioMistral-7B-DARE>

⁷<https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct>

uses probabilities of the first generated token to obtain a similarity score between 0 and 1, similar to how the MMLU benchmark [5] evaluates multiple-choice answers. In our case, these are the tokens for *Yes* and *No*.

We used the LangChain library to define **multiple prompt templates** [2], varying in complexity from simple similarity questions to outcome switching definitions and few-shot examples. All tested prompts are available in our GitHub repository⁸ and can be modified as needed. Experiments used deterministic decoding with temperature set to 0. We measured AUC on the 200 instance validation set to compare model quality across prompts, before setting the classification threshold, as shown in Figure 1.

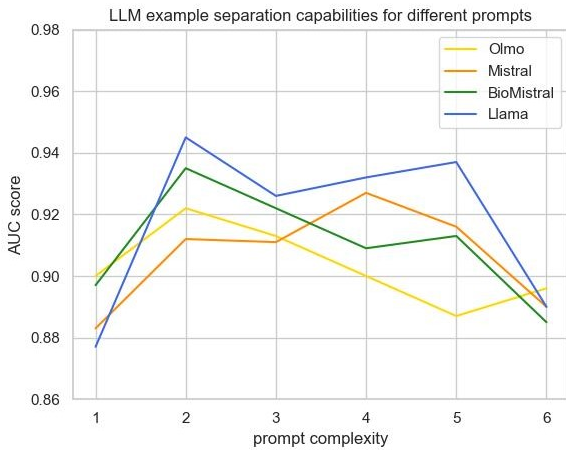


Figure 1: AUC scores across different LLMs and prompt complexity levels. Prompt names for each illustrative complexity level are provided in Table 1.

Table 1: List of evaluated prompt templates.

Complexity level	Prompt name
1	Sentence template
2	Outcome template
3	Role template
4	Article definition template
5	Wikipedia definition template
6	Random example template

The prompt with complexity 3, which defines the role of a clinical trial reviewer and the concept of primary and reported outcomes, was chosen as the starting point for further experiments. This prompt was near the middle of the AUC curve, which resembled an inverted U shape. Its average AUC score was comparable to the other prompts, but its range across models was the smallest. This indicates little bias, meaning none of them considerably over- or underperformed.

We then applied the Youden index to the 100-instance training set to determine the classification threshold and used it to obtain binary predictions on the 200-instance validation set. We **manually inspected instances misclassified** by the OLMo model and identified three recurring error types:

- insufficient domain terminology and acronym knowledge (for example, *HbA1c*),
- poor comprehension of longer sentences, and
- acceptance of **partially reported outcomes**.

We addressed (c) by iteratively adding clarifications to the prompt, stating that the reported outcome should include all components and details of the primary outcome. Each version was re-evaluated on the validation set until it reduced this error type while preserving F_1 score and accuracy.

3.5 Ensemble models

In the final evaluation, we included ensemble models, more specifically **hard voting classifiers** with a majority threshold. Each ensemble was composed of the three models with the highest F_1 score on the validation set from their respective group. This improved the F_1 score over any single model in the group. The ensembles were:

- Baseline voting:** stem similarity, sequence similarity and sentence encoder.
- LLM voting:** OLMo, BioMistral and Llama with initial prompt defining the clinical trial reviewer role.
- LLM' voting:** OLMo, Mistral and BioMistral with prompt corrected to reject partially reported outcomes.

4 Results

4.1 Outcome Classification

Table 2 shows the results of all evaluated models on the 2,496 instance test set. The sentence transformer achieved the best baseline F_1 score of 0.67 with a classification accuracy of 0.82. Among LLMs with the original prompt, BioMistral achieved the best F_1 of 0.73 and accuracy of 0.88. With the corrected prompt, OLMo^{*} achieved the best F_1 of 0.75 and accuracy of 0.89. The best overall result was achieved by LLM' voting with an F_1 of **0.78** and **accuracy of 0.90**, an improvement of 7 percentage points over baseline voting. Koroleva et al. [10] reported F_1 values between 0.87 and 0.90 with fine-tuned BERT models on a non-deduplicated dataset of outcome pairs.

4.2 Explanation

Deep learning baseline models and the proposed LLM approach are essentially **black boxes**, producing only a similarity score with no explanation of their decisions. For authors and reviewers of clinical trial reports, understanding why two outcomes were classified as similar or different is important. Since LLMs can interpret the problem and generate text, they are suitable for producing **natural language explanations**. We obtained explanations for already classified examples by inserting the model prediction and continuing the conversation with the user message “*Why? Explain your answer.*” [18] Two examples generated by OLMo are shown below:

```

Primary outcome: a reduction in depression
Reported outcome: GDS scores
Label: 1
Prediction: 1 ✓
Explanation: The reported outcome, GDS scores, are a
measure of depression, which is the primary outcome
in this clinical trial. The GDS is a well-known tool
to assess depression in elderly individuals.
Therefore, the reported outcome matches the defined
primary outcome.
    
```

⁸https://github.com/ta5946/Spin/blob/main/src/models/prompt_templates.py

Table 2: Classification thresholds and results of baseline models, LLMs, and ensemble models on the test set. LLM' denotes inference with the corrected prompt.

Group	Model	Threshold	Classification Accuracy	Precision	Recall	F_1 Score
Baseline	Stems similarity	0.3	0.805	0.528	0.845	0.650
	Sequence similarity	0.4	0.845	0.668	0.550	0.603
	Word2Vec embeddings	0.4	0.708	0.409	0.813	0.544
	Path distance	0.2	0.795	0.517	0.673	0.585
	Sentence transformer	0.4	0.821	0.553	0.862	0.673
LLM	OLMo	0.1	0.866	0.653	0.802	0.720
	Mistral	0.3	0.852	0.625	0.774	0.692
	BioMistral	0.2	0.881	0.712	0.751	0.731
	Llama	0.1	0.884	0.831	0.578	0.681
	OLMo'	0.3	0.889	0.725	0.778	0.750
	Mistral'	0.2	0.894	0.789	0.692	0.737
	BioMistral'	0.1	0.851	0.609	0.852	0.710
	Llama'	0.1	0.853	0.962	0.329	0.490
Ensemble	Baseline voting	majority	0.854	0.622	0.811	0.704
	LLM voting	majority	0.899	0.781	0.733	0.756
	LLM' voting	majority	0.901	0.758	0.793	0.775

Primary outcome: pain severity and knee range of motion (ROM)
 Reported outcome: pain
 Label: 0
 Prediction: 0 ✓
 Explanation: The reported outcome is "pain," but the defined primary outcome was "pain severity and knee range of motion (ROM)." It does not include the specific components of pain severity, and it does not mention knee ROM. Therefore, the reported outcome does not match the defined primary outcome.

5 Discussion

LLMs outperformed all baseline text similarity models in F_1 , which we attribute to defining the task context and the asymmetric relationship between the primary and reported outcome. Since we used LLMs as adaptive classifiers via prompting, they remained general-purpose. BioMistral performed better with minimal prompts while Mistral benefited from more context, indicating that domain adaptation influences prompt sensitivity. Llama followed the corrected instructions too strictly [19], as its recall dropped to 0.33 and F_1 to 0.49, making it an unusable classifier. Individual classification thresholds, obtained by maximizing the Youden index, compensated for model biases. Majority voting across LLMs further improved results, indicating that the models made distinct mistakes and complemented each other.

Our results still **trail behind fine-tuned BERT models** by about 10 percentage points on F_1 score. Part of the reason is that we worked with only 10% of the dataset for development and deduplicated instances, making the task harder. Thresholds tuned on this small, randomly sampled dataset can be unstable and vary across models. We also used LLMs to generate natural language explanations for the classified examples. A manual qualitative analysis showed that more than **50% of explanations are valuable**, while approximately every tenth contains hallucinations.

6 Conclusion

We investigated whether smaller open-source LLMs can classify pairs of outcomes based on text similarity as part of detecting outcome switching in RCT reports. For this purpose, we used a dataset of primary and reported outcomes labeled for similarity. Through prompt engineering, OLMo achieved the best F_1 score of 0.75, and majority voting of three LLMs improved it to 0.78. Other contributions include automatic threshold tuning, classification based on next token probabilities, and the option to explain the model decision. Our approach requires only a few hundred instances and no parameter fine-tuning. All source code and results, including model predictions and explanations, are publicly available in a GitHub repository.⁹

As future work, we propose evaluating larger, **more capable LLMs**. A complete spin detection pipeline would have to be extended to **other types of spin** and related tasks, such as outcome extraction and discrepancy judgment. Alternatively, models with longer context windows could analyze whole reports at once, replacing intermediate steps with reasoning. To ensure cross-lingual reliability, the experiments should be repeated with a machine-translated or separately collected dataset.

References

- [1] Isabelle Boutron and Philippe Ravaud. 2018. Misrepresentation and distortion of research in biomedical literature. *Proceedings of the National Academy of Sciences*, 115, 11, 2613–2619.
- [2] Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2025. Unleashing the potential of prompt engineering for large language models. *Patterns*, 6, 6.
- [3] Kellia Chiu, Quinn Grundy, and Lisa Bero. 2017. 'spin' in published biomedical literature: a methodological systematic review. *PLoS Biology*, 15, 9, e2002173.
- [4] Paul Glasziou, Douglas G Altman, Patrick Bossuyt, Isabelle Boutron, Mike Clarke, Steven Julious, Susan Michie, David Moher, and Elizabeth Wager. 2014. Reducing waste from incomplete or unusable reports of biomedical research. *The Lancet*, 383, 9913, 267–276.
- [5] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

⁹<https://github.com/ta5946/Spin>

- [6] Muhammad Shahzeb Khan et al. 2019. Level and prevalence of spin in published cardiovascular randomized clinical trial reports with statistically nonsignificant primary outcomes: a systematic review. *JAMA network open*, 2, 5, e192622–e192622.
- [7] Dale S Kim, Connor J McCabe, Brianna L Yamasaki, Kristine A Louie, and Kevin M King. 2018. Detecting random responders with infrequency scales using an error-balancing threshold. *Behavior research methods*, 50, 1960–1970.
- [8] Svetlana Kiritchenko, Berry De Bruijn, Simona Carini, Joel Martin, and Ida Sim. 2010. ExaCT: Automatic extraction of clinical trial characteristics from journal publications. *BMC medical informatics and decision making*, 10, 1–17.
- [9] Anna Koroleva, Sanjay Kamath, Patrick MM Bossuyt, and Patrick Paroubek. 2020. DeSpin: A prototype system for detecting spin in biomedical publications. In *Proceedings of the BioNLP 2020 workshop*, 49–59.
- [10] Anna Koroleva, Sanjay Kamath, and Patrick Paroubek. 2019. Measuring semantic similarity of clinical trial outcomes using deep pre-trained language representations. *Journal of Biomedical Informatics*, 100, 100058.
- [11] Suzanne Lockyer, Rob Hodgson, Jo C Dumville, and Nicky Cullum. 2013. “spin” in wound care research: the reporting and interpretation of randomized controlled trials with statistically non-significant primary outcome results or unspecified primary outcomes. *Trials*, 14, 1–10.
- [12] Iain J Marshall, Joël Kuiper, and Byron C Wallace. 2016. RobotReviewer: Evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association*, 23, 1, 193–201.
- [13] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16, 5, 1–72.
- [14] Francisco E Vera-Badillo et al. 2016. Bias in reporting of randomised clinical trials in oncology. *European Journal of Cancer*, 61, 29–35.
- [15] Dongguang Wang et al. 2021. Abstracts for reports of randomized trials of COVID-19 interventions had low quality and high spin. *Journal of Clinical Epidemiology*, 139, 107–120.
- [16] Amélie Yavchitz, Philippe Ravaud, Douglas G Altman, David Moher, Asbjørn Hrobjartsson, Toby Lasserson, and Isabelle Boutron. 2016. A new classification of spin in systematic reviews and meta-analyses was developed and ranked according to the severity. *Journal of clinical epidemiology*, 75, 56–65.
- [17] Yao-Ching Yu, Chun Chih Kuo, Ye Ziqi, Chang Yucheng, and Yueh-Se Li. 2024. Breaking the ceiling of the LLM community by treating token generation as a classification for ensembling. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 1826–1839.
- [18] Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2024. How language model hallucinations can snowball. In *International Conference on Machine Learning*. PMLR, 59670–59684.
- [19] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

Evaluating Automatic Speech Recognition Performance in Individuals with Parkinson's Disease and Healthy Controls

Stefan Andonov
stefan.andonov@finki.ukim.mk
Ss. Cyril and Methodius University,
Faculty of Computer Science and
Engineering
Skopje, North Macedonia

Radomir Damjanović
rdamjanovic@gmail.com
Clinic for Neurology, University
Clinical Center Niš
Niš, Serbia

Katarina Trojchanec Dineva
katarina.trojcanec@finki.ukim.mk
Ss. Cyril and Methodius University,
Faculty of Computer Science and
Engineering
Skopje, North Macedonia

Biljana Tojtovska
biljana.tojtovska@finki.ukim.mk
Ss. Cyril and Methodius University,
Faculty of Computer Science and
Engineering
Skopje, North Macedonia

Pance Ribarski
pance.ribarski@finki.ukim.mk
Ss. Cyril and Methodius University,
Faculty of Computer Science and
Engineering
Skopje, North Macedonia

Suzana Loshkovska
suzana.loshkovska@finki.ukim.mk
Ss. Cyril and Methodius University,
Faculty of Computer Science and
Engineering
Skopje, North Macedonia

Monika Simjanoska Misheva
monika.simjanoska@finki.ukim.mk
Ss. Cyril and Methodius University,
Faculty of Computer Science and
Engineering
Skopje, North Macedonia

Kostadin Mishev
kostadin.mishev@finki.ukim.mk
Ss. Cyril and Methodius University,
Faculty of Computer Science and
Engineering
Skopje, North Macedonia

Jelena Stamenović
j.stamenovic@yahoo.com
University Clinical Center Niš,
Medical Faculty, University of Niš
Niš, Serbia

Stevo Lukić
srlukic@gmail.com
Clinic for Neurology, University
Clinical Center Niš, Medical Faculty,
University of Niš
Niš, Serbia

Abstract

Automatic speech recognition (ASR) is increasingly used in clinical speech-analysis pipelines, where transcription errors may affect the downstream linguistic measures. Speech alterations associated with Parkinson's disease (PD) may pose additional challenges to general-purpose ASR systems. This study evaluates ASR transcription fidelity in 88 PD–HC pairs closely matched on sex, age, and education from the Slovak EWA Database, using three standardised picture-description tasks that elicit spontaneous connected speech. We evaluated Whisper large-v3 as the primary ASR system, and used large-v3-turbo to assess robustness. ASR outputs were compared with manually corrected reference transcripts using annotation-aware lexical-verbatim normalisation. The primary outcome was participant-level speaker-pooled word error rate (WER), with character error rate (CER) and word-error components examined as supporting measures. PD–HC differences were evaluated within matched pairs using two-sided Wilcoxon signed-rank tests, with paired-bootstrap 95% confidence intervals for mean differences. For large-v3, mean WER was 24.34% in PD and 18.94% in HC, yielding a mean paired difference of +5.40 percentage points (95% CI: 3.00–7.89; $p = 1.54 \times 10^{-5}$).

A similar difference was observed with large-v3-turbo (+5.82 percentage points; $p = 2.51 \times 10^{-4}$), and CER was also higher in PD for both models. For large-v3, the WER difference was driven mainly by deletions (+3.58 percentage points) and, to a lesser extent, substitutions (+2.11), both significant after Holm correction, whereas insertion rates did not differ significantly between groups. The main finding was consistent across picture stimuli and sensitivity analyses. These results show a diagnosis-associated difference in ASR transcription fidelity in this matched cohort. Such differences highlight the importance of evaluating ASR transcription fidelity when automatic transcripts are used for clinical speech and language analysis.

Keywords

automatic speech recognition, Parkinson's disease, connected speech analysis, transcription accuracy, clinical natural language processing, digital health

1 Introduction

Speech and language changes in Parkinson's disease (PD) extend beyond voice and articulation to timing, fluency, lexical-semantic processing, syntax, pragmatics, and discourse [2, 14, 5]. These measures are increasingly investigated as non-invasive digital biomarkers for disease characterization and monitoring [18, 2]. standardised picture-description tasks are particularly useful because they elicit connected speech while engaging multiple aspects of language production.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Information Society 2026, Ljubljana, Slovenia

© 2026 Copyright held by the owner/author(s).

<https://doi.org/10.70314/is.2026.aihc.68>

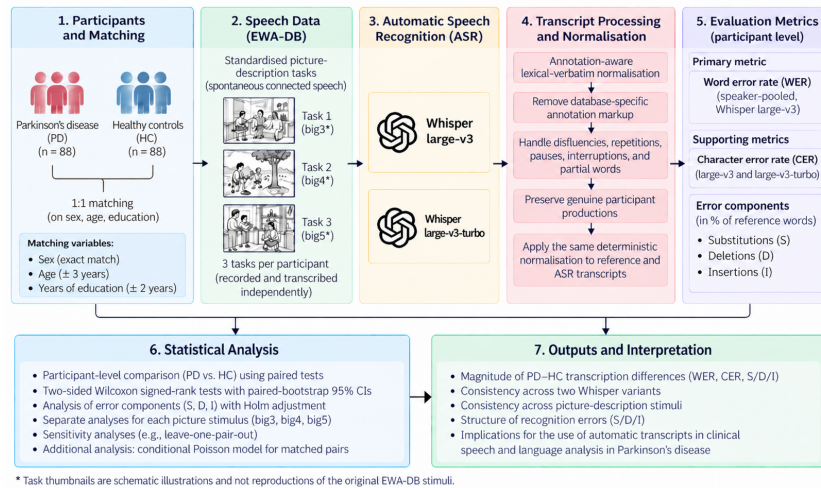


Figure 1: Overview of the study design and evaluation pipeline, from cohort construction through ASR processing and evaluation to matched participant-level PD–HC analysis.

Automatic speech recognition (ASR) is increasingly used to generate transcripts for clinical speech-analysis pipelines. Transcription errors can therefore propagate into linguistic features, language-model representations, and downstream clinical predictions. Recent PD studies have already used Whisper-generated transcripts for disease detection, symptom assessment, treatment-related language analysis, and longitudinal monitoring [4, 3, 19, 21]. Higher ASR error in PD than in controls has been reported for Spanish speech [12] and in a bilingual English/French cohort [16], while a smaller descriptive WER difference was reported for Czech spontaneous discourse [20]. Related work has also reported WER/CER differences across additional PD datasets and disease-severity levels [15]. However, not all ASR-based evaluations have found a significant speaker-group effect; for example, no significant effect of speaker group on AI transcription success was observed in a noisy-sentence paradigm [13]. Therefore, the magnitude of PD-associated ASR degradation appears to depend on the language, speech task, ASR system, recording conditions, and evaluation protocol. Prior studies suggest that ASR errors may preserve information useful for disease classification [11]. It remains unclear whether a systematic PD–HC difference persists under a tightly matched participant-level design using current ASR and standardised spontaneous connected speech.

Measured ASR error also depends on how the reference transcripts are defined. Clinical connected speech includes repetitions, self-corrections, fragments, filled pauses, non-speech events, and investigator speech, which human annotators and ASR systems may handle differently. Transparent transcription and normalisation is therefore important when evaluating clinical speech [6, 8, 22]. A systematic PD–HC difference in transcription fidelity is a measurement-validity concern when automatic transcripts are used for clinical analysis, although it does not by itself demonstrate bias in downstream biomarkers.

This paper evaluates ASR transcription fidelity in 88 demographically matched PD–HC pairs from the Slovak EWA Database [17, 7]. EWA-DB has previously been used for PD classification and cross-linguistic representation learning [1, 9], but its big3–big5 picture-description tasks have not, to our knowledge, been directly evaluated for PD–HC ASR transcription fidelity. Whisper large-v3 is used as the primary ASR system and large-v3-turbo as a supporting model. We ask two questions: (1) how large

is the PD–HC difference in ASR transcription error for matched Slovak picture-description speech, and (2) whether this difference is consistent across the two Whisper variants and error measures and, for the primary model, which word-error components contribute to it. An annotation-aware lexical-verbatim normalisation protocol is used to preserve participant speech while explicitly handling database-specific annotations and investigator speech.

The main contributions are:

- (1) A controlled evaluation of ASR transcription fidelity in 88 PD–HC pairs closely matched on sex, age, and education using three standardised picture-description tasks.
- (2) Participant-level evaluation of the PD–HC difference across two Whisper variants and multiple error measures, with decomposition of large-v3 WER into substitution, deletion, and insertion components.
- (3) An annotation-aware lexical-verbatim evaluation protocol that preserves clinically relevant participant productions and explicitly handles database-specific annotations and investigator speech.

2 Materials and Methods

2.1 Study Design and Evaluation Pipeline

The study followed a matched, participant-level evaluation design, summarised in Fig. 1. The overall pipeline comprised cohort construction, automatic transcription of the selected speech recordings, transcript normalisation and investigator-speech handling, computation of recognition-error measures, participant-level aggregation, and paired comparison between individuals with Parkinson’s disease and healthy controls. The supporting analyses examined the consistency of the findings across the two Whisper variants, error measures, speech tasks, and recording-quality conditions. The following subsections describe each stage of the pipeline in detail.

2.2 Data and Matched Cohort

2.2.1 EWA-DB and Speech Tasks. We obtained speech recordings from the Slovak EWA Database (EWA-DB), which contains speech, demographic, and clinical data from participants with neurodegenerative disorders and healthy controls [17, 7]. We obtained access under a signed End User License Agreement with

Table 1: Characteristics and covariate balance of the 88-pair cohort. Continuous variables are reported as mean $\pm$ standard deviation and categorical variables as n (%).

Characteristic	HC ($n = 88$)	PD ($n = 88$)	SMD pre $\rightarrow$ post
Sex, n (%)			0.77 $\rightarrow$ 0.00
Male	53 (60.2%)	53 (60.2%)	
Female	35 (39.8%)	35 (39.8%)	
Age, years			0.47 $\rightarrow$ 0.03
Male	67.84 $\pm$ 9.91	68.11 $\pm$ 9.97	
Female	65.60 $\pm$ 9.00	65.98 $\pm$ 9.07	
Female	71.23 $\pm$ 10.39	71.34 $\pm$ 10.52	
Education, years			0.20 $\rightarrow$ 0.07
Male	14.82 $\pm$ 2.94	14.61 $\pm$ 3.00	
Female	15.32 $\pm$ 2.69	15.02 $\pm$ 2.77	
Female	14.06 $\pm$ 3.17	14.00 $\pm$ 3.25	
MoCA score			0.99 $\rightarrow$ 0.87
Male	27.49 $\pm$ 1.29	25.73 $\pm$ 2.57	
Female	27.51 $\pm$ 1.22	25.68 $\pm$ 2.53	
Female	27.46 $\pm$ 1.40	25.80 $\pm$ 2.67	
WAIS score			0.07 $\rightarrow$ 0.12
Male	2.72 $\pm$ 2.45	3.05 $\pm$ 2.93	
Female	2.37 $\pm$ 2.38	3.19 $\pm$ 3.16	
Female	3.26 $\pm$ 2.49	2.83 $\pm$ 2.57	
Family history, n (%)			0.02 $\rightarrow$ 0.03
Male	17 (19.3%)	18 (20.5%)	
Female	9 (17.0%)	10 (18.9%)	
Female	8 (22.9%)	8 (22.9%)	

the Institute of Informatics of the Slovak Academy of Sciences. The study focuses on spontaneous connected speech elicited by picture description. The three complex tasks (big3–big5) were prespecified as the primary analysis scope because they elicit longer and linguistically richer connected-speech samples than the simpler big6–big7 tasks.

EWA-DB provides audio and manually corrected reference transcripts. These were generated from output of the corpus developers’ ASR system and then corrected by trained annotators, who also added acoustic-event labels [17]. As a result, the transcripts contain database-specific annotation markers in addition to lexical content. Section 2.3.2 describes their normalisation before ASR scoring.

2.2.2 Participant Selection and Matching. EWA-DB contains 1,323 HC and 175 PD participants. Of these, 896 HC and 169 PD met the database eligibility criteria; 538 HC and 92 PD also had audio-publication consent, and 476 HC had available manual annotations. All 92 PD participants with audio-publication consent also had manual annotations. The resulting eligible cohort therefore comprised 476 HC and 92 PD participants. We constructed a 1:1 matched cohort using exact matching on sex and calipers of ± 3 years for age and ± 2 years for education. Among admissible matches, we optimised assignment globally using linear sum assignment.

We obtained a valid match for 88 of the 92 PD participants (95.7%), yielding 88 PD–HC pairs. Four PD participants had no HC candidate satisfying the prespecified matching constraints. Covariate balance was assessed using the absolute standardised mean difference (|SMD|); all three matching variables had |SMD| < 0.10 after matching (Table 1). We reported MoCA, WAIS (WAIS-III Similarities subtest), and family history (reported occurrence of neurodegenerative disease in the family) descriptively but did not use them for matching.

Speech amount was characterised as the mean number of normalised reference words per eligible big3–big5 recording. HC participants produced 118.01 ± 37.74 words per recording on average (median 124.67), compared with 104.70 ± 42.04 in PD

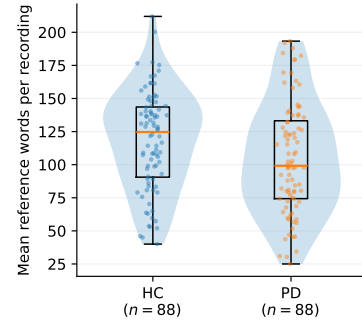


Figure 2: Participant-level distribution of speech amount in the matched cohort. Speech amount was defined as the mean number of normalised reference words per eligible big3–big5 recording. Each participant contributes one value; boxplots show the median and interquartile range.

(median 99.17), a descriptive mean difference of 13.31 words per recording (Fig. 2).

2.3 ASR and Transcript Processing

2.3.1 ASR Systems and Inference. We evaluated two contemporary multilingual Whisper models: Whisper large-v3, the primary ASR system, and Whisper large-v3-turbo as a supporting model. We applied both systems without task-specific or EWA-DB fine-tuning. We specified Slovak as the recognition language and used identical decoding settings across diagnostic groups. Each eligible picture-description recording was transcribed independently and the resulting hypotheses were subsequently processed using the normalisation procedure described in the 2.3.2 Section.

2.3.2 Transcript Normalisation. EWA-DB reference transcripts contain database-specific annotation markup in addition to spoken lexical content, including structural and disfluency tags, filled pauses, interruptions, and markers for partial or truncated words. We used a predefined lexical-verbatim normalisation protocol to remove annotation markup while preserving genuine participant productions that may carry clinically relevant information.

We used a custom normalisation procedure rather than Whisper’s built-in English text normaliser. We applied the same deterministic normalisation procedure to the reference transcripts and ASR hypotheses. We handled database-specific annotation tags according to their defined meaning. Numeric expressions were converted to written Slovak forms (e.g., 4 to štyri), because ASR hypotheses typically represent such content as words and removing numeric tokens from the reference would penalise otherwise correct recognition. Intra-word hyphens were joined and non-lexical punctuation was removed. Diacritics were retained because they encode phonologically and lexically significant distinctions in Slovak. Repetitions, self-corrections, and partial word productions were preserved according to the lexical-verbatim evaluation approach. Such productions may carry clinical information, and differences in transcription-policy can affect measured WER [22]. All text was lowercased.

Fragment-truncation markers were removed while retaining the associated lexical content. Trailing, leading, and standalone markers were inspected and represented either recording-boundary truncation or genuine partial productions. No recordings were excluded on this basis. As an additional quality-control check, we flagged references shorter than 10 words for review; two recordings met this criterion.

2.3.3 Investigator-Speech Handling. Some recordings contained investigator speech, which should not contribute to participant-level ASR evaluation (17 in total). We reviewed each recording to determine whether investigator speech could be separated reliably. We allowed recovery only when the investigator produced brief non-lexical fillers and the metadata duration matched the audio duration. Three recordings met these conditions. Investigator-tagged content was removed from the reference, and ASR words were removed when the midpoint of their recognised interval fell within the corresponding investigator interval. This avoided removing adjacent participant words at interval boundaries.

The remaining fourteen recordings were excluded. Thirteen contained substantive investigator speech, and the annotations did not always cover the complete statement. This created a risk that unannotated investigator words would be attributed to the participant. One additional recording had an audio/JSON duration mismatch, so we could not trust its timestamps.

Overall, 1,688 of 1,702 eligible big3–big5 recordings were scored. All 88 matched pairs remained in the primary analysis, although four pairs contributed unequal numbers of recordings.

2.4 ASR Evaluation Measures

We evaluated ASR performance using word error rate (WER) and character error rate (CER). Word-level errors were also decomposed into substitutions, deletions, and insertions. The primary outcome was participant-level speaker-pooled WER for Whisper large-v3. Speaker-pooled WER was calculated by summing substitution, deletion, insertion, and reference-word counts across all eligible recordings for each participant before computing WER. This approach gives each reference word equal weight and avoids disproportionately weighting short recordings. We used the same participant-level pooling strategy for CER and individual word-error components.

2.5 Statistical Analysis

The primary endpoint was participant-level speaker-pooled WER for Whisper large-v3. Substitution, deletion, insertion, and reference-word counts were summed across all eligible big3–big5 recordings before calculating WER, avoiding equal weighting of recordings with different speech amounts. All 88 matched pairs were retained, including four with unequal picture availability. With 1:1 matching, inference was based on within-pair PD–HC differences.

The distribution of paired large-v3 WER differences was inspected before inferential testing and showed a marked departure from normality, with one extreme upper-tail observation. Therefore, a two-sided Wilcoxon signed-rank test was used as the main paired test, with rank-biserial correlation reported as the effect size. The magnitude of the group difference was summarised by the mean paired difference in percentage points (pp) with a paired-bootstrap 95% confidence interval based on 10,000 resamples. An exact sign test and paired *t*-test were used as complementary robustness analyses.

Confirmatory inference was limited to the primary endpoint, participant-level speaker-pooled WER for Whisper large-v3. All other inferential analyses were supporting or exploratory, and their *p*-values were not interpreted as belonging to a common confirmatory family. Large-v3 CER and large-v3-turbo WER/CER were analysed analogously; Holm adjustment was applied only within the three word-error-component tests and within the

three picture-specific tests, which used complete matched pairs for each stimulus.

Sensitivity analyses assessed whether the primary result was robust to recording quality and investigator-speech handling. The analysis was repeated after excluding recordings associated with acoustically unsatisfactory data in the EWA-DB metadata. A second sensitivity analysis applied a stricter investigator-speech policy by additionally excluding the three filler-only recordings retained by the primary preprocessing procedure, thereby excluding all 17 investigator-flagged recordings. WER was re-pooled after filtering.

Because four matched pairs contained different sets of available picture stimuli, a shared-stimulus sensitivity analysis was also performed across all 88 pairs. Within each pair, WER was re-computed using only picture stimuli available for both participants. This ensured that, within each pair, both participants were compared on the same set of pictures. Influence of individual pairs was assessed using leave-one-pair-out analyses and a diagnostic analysis excluding the single extreme paired WER difference identified by the standard $1.5 \times \text{IQR}$ rule.

As a secondary analysis, total large-v3 word errors were analysed using a conditional Poisson model for matched pairs, with the amount of reference speech included as the exposure term. The model compared PD and HC error rates within each matched pair while allowing for pair-specific baseline differences. Because the observed variability exceeded that expected under a standard Poisson model, robust sandwich standard errors with the HC1 finite-sample adjustment were used to obtain confidence intervals and *p*-values. This analysis was treated as supporting evidence rather than as a replacement for the primary participant-level WER comparison. All tests were two-sided with $\alpha = 0.05$ and were performed in R 4.4.2.

3 Results

For Whisper large-v3, participant-level speaker-pooled WER was higher in PD than in HC (24.34% vs. 18.94%), with a mean within-pair difference of +5.40 pp (95% CI 3.00–7.89). The Wilcoxon signed-rank test supported this difference ($p = 1.54 \times 10^{-5}$), with a rank-biserial correlation of $r_{rb} = 0.531$. WER was higher for the PD participant in 65 of 88 matched pairs (73.9%). The exact sign test and paired *t*-test supported the same conclusion. The distribution of matched differences for both Whisper variants is shown in Fig. 3.

The PD–HC WER difference was also observed with Whisper large-v3-turbo, and CER was likewise higher in PD for both Whisper variants. The mean PD–HC WER difference was +5.82 pp for large-v3-turbo, while the corresponding CER differences were +3.37 pp for large-v3 and +3.10 pp for large-v3-turbo. Detailed group means, confidence intervals, and test results are summarised in Table 2.

When large-v3 WER was decomposed into substitution, deletion, and insertion components, deletions showed the largest PD–HC difference (+3.58 pp), followed by substitutions (+2.11 pp), while insertion rates were similar between groups. Paired Wilcoxon signed-rank tests showed significant differences for deletions and substitutions after Holm correction across the three component tests, but not for insertions.

In separate picture-specific matched analyses, WER was higher in PD for all three stimuli, with mean paired differences ranging from +4.50 to +6.20 pp. All three PD–HC comparisons were significant in Wilcoxon signed-rank tests after Holm adjustment

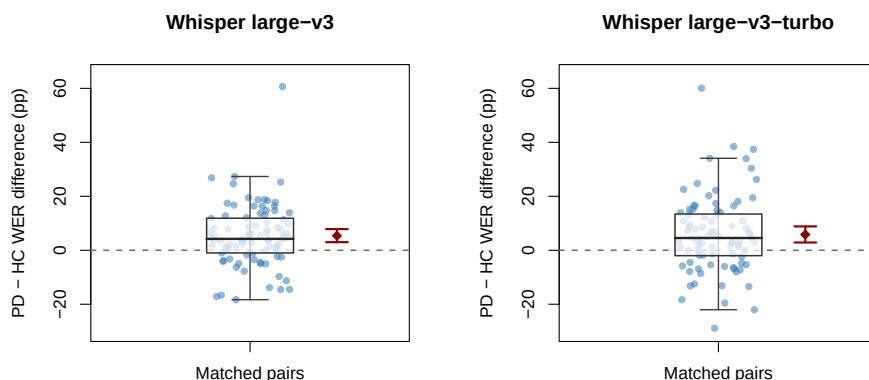


Figure 3: Matched PD–HC WER differences for Whisper large-v3 and large-v3-turbo. Positive values indicate higher WER in PD. Boxes show median and interquartile range; diamonds show mean with 95% CIs. The dashed line marks zero difference.

Table 2: Primary and supporting matched ASR results show participant-level mean error rates (%). Differences are PD minus HC in percentage points, with paired-bootstrap 95% CIs. Holm-adjusted component p -values; others from two-sided Wilcoxon signed-rank tests.

Analysis	n	HC mean	PD mean	PD–HC [95% CI]	p
large-v3 WER	88	18.94	24.34	5.40 [3.00, 7.89]	1.54×10^{-5}
large-v3-turbo WER	88	20.12	25.94	5.82 [2.89, 8.86]	2.51×10^{-4}
large-v3 CER	88	8.88	12.26	3.37 [1.77, 5.13]	1.53×10^{-4}
large-v3-turbo CER	88	9.34	12.44	3.10 [1.38, 4.88]	2.93×10^{-3}
Substitution rate	88	10.61	12.72	2.11 [0.76, 3.51]	0.0106 ^a
Deletion rate	88	6.06	9.63	3.58 [2.13, 5.16]	3.15×10^{-5a}
Insertion rate	88	2.27	1.99	-0.28 [-1.01, 0.44]	0.228 ^a

^aHolm-adjusted across substitution, deletion, and insertion tests.

($p < 0.004$), indicating that the overall result was not driven by a single picture stimulus.

Sensitivity analyses produced estimates close to the primary +5.40-pp difference. For each sensitivity analysis, participant-level speaker-pooled WER was re-computed after the relevant recording restriction, and the resulting PD–HC differences were tested within matched pairs using the same Wilcoxon signed-rank test as in the primary analysis. Restriction to clean recordings ($n = 83$ pairs) and application of the stricter investigator-speech policy ($n = 87$ pairs) yielded mean differences between +5.38 and +5.53 pp. Restricting each matched pair to picture stimuli available for both participants retained all 88 pairs and produced a difference of +5.50 pp (95% CI 3.10–8.02).

One matched pair had an exceptionally large PD–HC WER difference (+60.62 pp). Diagnostic removal of this pair reduced the mean difference to +4.76 pp but did not change the substantive conclusion. Across all leave-one-pair-out analyses, mean differences remained between +4.76 and +5.67 pp, showing that no single matched pair accounted for the primary result.

As a secondary analysis, total large-v3 word errors were analysed using a conditional Poisson model for matched pairs, while accounting for the number of reference words produced by each participant. The estimated PD-to-HC word-error rate ratio was 1.269 (robust 95% CI 1.141–1.411; $p = 1.09 \times 10^{-5}$), indicating an approximately 27% higher word-error rate per reference word in PD than in HC. This result was consistent with the primary participant-level WER analysis.

4 Discussion

In this demographically matched Slovak cohort, Whisper produced higher transcription error for PD speech than for HC

speech. With large-v3, mean WER was 24.34% in PD and 18.94% in HC, a mean difference of +5.40 pp. A similar difference was observed with large-v3-turbo (+5.82 pp), and CER was also higher in PD for both Whisper variants. For large-v3, the WER difference was also present across all three picture-description stimuli. These findings are consistent in direction with previous reports in Spanish, English/French, and Czech [12, 16, 20]. However, the magnitude of the PD–HC difference varies considerably across studies, and significant differences have not been observed under all experimental conditions [13]. This variability highlights the importance of language, task, recording conditions, ASR system, and evaluation protocol when interpreting group differences in ASR performance.

The type of recognition error may be particularly relevant when automatic transcripts are used to derive clinical measures. Deletions showed the largest PD–HC difference, increasing from 6.06% in HC to 9.63% in PD. Substitutions were also higher in PD, whereas insertion rates were similar between groups. Participants with PD also produced shorter picture descriptions on average. To avoid giving disproportionate weight to short individual recordings, participant-level WER was calculated by pooling error and reference-word counts across each participant's available recordings. Even so, differences in speech amount may be relevant for transcript-derived measures such as transcript length and lexical counts. The relationship between speech amount and WER was not formally analysed as a mediating mechanism in this study and should be examined in future work.

The main finding was consistent across the supporting and sensitivity analyses. Similar PD–HC differences were observed with both Whisper variants, for CER, across all three picture stimuli, after restriction to clean recordings, and under the stricter investigator-speech policy. Restricting each matched pair to picture stimuli available for both participants changed the estimated difference only from +5.40 to +5.50 pp, indicating that unequal picture availability in four pairs did not account for the primary result. Leave-one-pair-out analyses and diagnostic removal of the single extreme pair led to the same substantive conclusion. The secondary conditional Poisson analysis was also consistent with the primary result, estimating an approximately 27% higher word-error rate per reference word in PD.

The agreement between large-v3 and large-v3-turbo should be interpreted as robustness within the Whisper model family, rather than as evidence of generalisation across independent ASR architectures. The two models are closely related. Evaluation

with recognisers based on different architectures and training procedures is therefore needed to determine how broadly the observed PD–HC difference generalises.

These findings are relevant because automatic transcripts are increasingly used to derive linguistic features, language-model representations, and digital biomarkers. When transcription fidelity differs systematically between clinical and control groups, downstream group differences may reflect both genuine characteristics of PD speech and transcription error [10, 8]. This does not mean that every ASR-derived biomarker is biased. It does, however, mean that ASR fidelity should be evaluated and reported separately for patient and control groups, rather than only as an overall accuracy measure, when transcript-derived clinical features are validated.

Several limitations should be considered. The analysis is based on one Slovak database and two closely related Whisper variants, so generalisation to other languages, tasks, and ASR architectures remains to be established. The groups were closely matched on sex, age, and education, but not on cognition. MoCA was not used for matching because the EWA-DB eligibility criteria impose different cognitive-score ranges for HC and PD. The observed difference should therefore be interpreted as diagnosis-associated rather than as evidence for a specific underlying mechanism. The present study cannot determine the relative contributions of motor-speech impairment, cognitive-linguistic differences, disfluency, acoustic changes, or their combination.

A further limitation concerns the reference transcripts. The EWA-DB manual transcripts were created by correcting output from the corpus developers' recognizer rather than independently from scratch. They may therefore retain some lexical or orthographic conventions of that system. Because the same reference-construction procedure was applied to both groups, this limitation is not specific to either PD or HC. It may nevertheless affect the absolute WER values obtained when independent ASR systems are evaluated against these references.

5 Conclusion

In this matched cohort, Whisper large-v3 WER was 5.40 pp higher in PD than HC, with deletions showing the largest component difference. Because automatic transcription often precedes downstream language and digital-biomarker analyses, group differences in ASR fidelity should be assessed and reported when interpreting transcript-derived measures.

Acknowledgements

This research was supported by the European Union through the Horizon Europe programme, under the ChatMED project (Grant Agreement ID: 101159214). Additional support was provided by the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, North Macedonia, and the Faculty of Medicine, University of Niš, Niš, Serbia.

References

- [1] Angelika Ando, Adrien Lesage, Marc de Gennes, and Rachid Riad. 2026. Parkmae: a cross-linguistic masked autoencoder framework for robust parkinson's disease detection from speech. *Scientific Reports*, 16, 783. doi:10.1038/s41598-025-30251-7.
- [2] Fangyuan Cao, Adam P. Vogel, Puya Gharahkhani, and Miguel E. Renteria. 2025. Speech and language biomarkers for parkinson's disease prediction, early diagnosis and progression. *npj Parkinson's Disease*, 11, 1, 57. doi:10.1038/s41531-025-00913-4.
- [3] Matilde Castelli et al. 2025. Detecting neuropsychiatric fluctuations in parkinson's disease using patients' own words: the potential of large language models. *npj Parkinson's Disease*, 11, 1, 79. doi:10.1038/s41531-025-00939-8.
- [4] Jonathan L. Crawford. 2025. Linguistic changes in spontaneous speech for detecting parkinson's disease using large language models. *PLOS Digital Health*, 4, 2, e0000757. doi:10.1371/journal.pdig.0000757.
- [5] Guilherme Briczinski de Souza and Barbara Costa Beber. 2026. Language skills in parkinson's disease: state of the art. *Dementia & Neuropsychologia*, 20, e20250350. doi:10.1590/1980-5764-DN-2025-0350.
- [6] Jonathan Heitz, Gerold Schneider, and Nicolas Langer. 2024. The influence of automatic speech recognition on linguistic features and automatic alzheimer's disease detection from spontaneous speech. In *Proceedings of LREC-COLING 2024*. ELRA and ICCL, 15955–15969.
- [7] Institute of Informatics of the Slovak Academy of Sciences, AXON PRO s.r.o., Pan-European University, Marián Trnka, and Milan Rusko. EWA-DB – Early Warning of Alzheimer speech database. Zenodo, (Oct. 2023). doi:10.5281/zenodo.10952480.
- [8] Sandra Anna Just, Brita Elvevåg, Shrankhla Pandey, Ivan Nenchev, Anna-Lena Bröcker, Christiane Montag, and Sarah E. Morgan. 2025. Moving beyond word error rate to evaluate automatic speech recognition in clinical samples: lessons from research into schizophrenia-spectrum disorders. *Psychiatry Research*, 352, 116690. doi:10.1016/j.psychres.2025.116690.
- [9] Moreno La Quatra, Juan Rafael Orozco-Arroyave, and Marco Sabato Siniscalchi. 2025. Bilingual dual-head deep model for parkinson's disease detection from speech. In *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5. doi:10.1109/ICASSP49660.2025.10889445.
- [10] Changye Li, Weizhe Xu, Trevor Cohen, and Serguei Pakhomov. 2024. Useful blunders: can automated speech recognition errors improve downstream dementia classification? *Journal of Biomedical Informatics*, 150, 104598. doi:10.1016/j.jbi.2024.104598.
- [11] Yin-Long Liu, Yuanhao Li, Rui Feng, Jiaxin Chen, Yiming Wang, Yu-Ang Chen, Nan Ding, Jiahong Yuan, and Zhen-Hua Ling. 2026. Beyond manual transcripts: exploring the potential of automatic speech recognition errors in improving alzheimer's disease detection. *Journal of Biomedical Informatics*, 173, 104968. doi:10.1016/j.jbi.2025.104968.
- [12] Laureano Moro-Velazquez, Jaejin Cho, Shinji Watanabe, Mark A. Hasegawa-Johnson, Odette Scharenborg, Heejin Kim, and Najim Dehak. 2019. Study of the performance of automatic speech recognition systems in speakers with parkinson's disease. In *Proceedings of Interspeech 2019*, 3875–3879. doi:10.21437/Interspeech.2019-2993.
- [13] Gemma Moya-Galé, Stephen J. Walsh, and Alireza Goudarzi. 2022. Automatic assessment of intelligibility in noise in parkinson disease: validation study. *Journal of Medical Internet Research*, 24, 10, e40567. doi:10.2196/40567.
- [14] Maura Panozzo Chiomento, Maria Vender, and Denis Delfitto. 2026. Language competence in parkinson's disease: a systematic review of 20 years of research. *Journal of Speech, Language, and Hearing Research*, 69, 1, 253–272. doi:10.1044/2025_JSLHR-25-00205.
- [15] Benedetta Perrone, Federica Amato, and Gabriella Olmo. 2026. Voice classification in parkinson's disease: a deep learning approach using transformers and error rate metrics. *Biomedical Signal Processing and Control*, 113, 108954. doi:10.1016/j.bspc.2025.108954.
- [16] Peter Plantinga, Briac Cordelle, Dominique Louër, Mirco Ravanelli, and Denise Klein. 2025. Does language matter for early detection of parkinson's disease from speech? In *2025 IEEE 35th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 1–6. doi:10.1109/MLSP62443.2025.11204272.
- [17] Milan Rusko, Róbert Sabo, Marián Trnka, Alfréd Zimmermann, Richard Malaschitz, Eugen Ružický, Petra Brandoburová, Viktória Kevická, and Matej Škorvák. 2024. Slovak database of speech affected by neurodegenerative diseases. *Scientific Data*, 11, 1, 1320. doi:10.1038/s41597-024-04171-6.
- [18] Jan Rusz, Paul Krack, and Elina Tripoliti. 2024. From prodromal stages to clinical trials: the promise of digital speech biomarkers in parkinson's disease. *Neuroscience & Biobehavioral Reviews*, 167, 105922. doi:10.1016/j.neubiorev.2024.105922.
- [19] Martin Šubert, Tereza Tykalová, Michal Novotný, Ondřej Bezdiček, Petr Dušek, and Jan Rusz. 2025. Long-term dopaminergic therapy improves spoken language in de-novo parkinson's disease. *Journal of Neurology*, 272, 5, 344. doi:10.1007/s00415-025-13070-8.
- [20] Martin Šubert, Tereza Tykalová, Michal Novotný, Petr Dušek, Jiří Klempíř, and Jan Rusz. 2025. Automated analysis of spoken language differentiates multiple system atrophy from parkinson's disease. *Journal of Neurology*, 272, 2, 113. doi:10.1007/s00415-024-12828-w.
- [21] Martin Šubert et al. 2026. Large language model analysis of real-world phone calls reveals prodromal and progressive biomarker of parkinsonism: a two-year proof-of-concept study. *PLOS Digital Health*, 5, 7, e0001458. doi:10.1371/journal.pdig.0001458.
- [22] Laurin Wagner, Mario Zusag, and Bernhard Thallinger. 2026. Transcription policy as a latent variable: activating controllable verbatim asr with word-level timing. *arXiv preprint arXiv:2607.18934*.

Digital Health Across EU Borders: Data Protection and Artificial Intelligence in MyHealth@EU

Primož Kocuvan*
primoz.kocuvan@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Matjaž Gams
matjaz.gams@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

The digitalisation of healthcare in the European Union enables patients to access medical services across national borders while allowing healthcare professionals to retrieve essential clinical information from other Member States. MyHealth@EU supports this objective through services such as cross-border ePrescriptions, eDispensation, and Patient Summaries. However, the exchange and potential automated processing of sensitive health information raises significant concerns regarding privacy, security, interoperability, and patient rights. This paper examines MyHealth@EU in relation to the General Data Protection Regulation (GDPR) and the European Union Artificial Intelligence Act (AI Act). A qualitative legal and document-based methodology is used to analyse the requirements of the three frameworks and to identify areas in which they complement or potentially conflict with one another. The analysis shows that the GDPR provides the fundamental legal safeguards for processing health data, including lawfulness, data minimisation, purpose limitation, security, and accountability, while the AI Act introduces additional requirements for high-risk AI systems concerning data quality, transparency, risk management, robustness, and human oversight. Particular attention is given to a potential tension between GDPR data minimisation requirements and the need for sufficiently representative datasets for reliable healthcare AI. The paper concludes that trustworthy cross-border digital healthcare requires interoperability to be developed together with strong data-protection safeguards, cybersecurity measures, transparent AI governance, and meaningful human control. Our hypothesis is: The GDPR and the AI Act provide largely complementary regulatory safeguards for the use of health data in MyHealth@EU, but their combined application creates a potential tension requiring contextual reconciliation between the GDPR principle of data minimisation and the AI Act requirement for sufficiently representative and high-quality datasets in healthcare AI systems.

Keywords

MyHealth@EU, Digital Health, Cross-Border Healthcare, GDPR, Artificial Intelligence Act, Health Data, Data Protection, Interoperability

1 Introduction

The digital transformation of healthcare has created new opportunities for improving the availability, continuity, and quality of medical services across national borders. Within the European Union (EU), these developments are particularly relevant because

citizens increasingly travel, work, and reside in Member States other than their country of affiliation. In such situations, healthcare professionals may need timely access to essential medical information that was originally created and stored in another national healthcare system. MyHealth@EU addresses this challenge by supporting the cross-border exchange of electronic health information through a common European digital infrastructure [5]. Two central services supported by MyHealth@EU are the electronic prescription and electronic dispensation services (ePrescription/eDispensation) and the Patient Summary. The former enables a prescription issued in one participating country to be accessed by a pharmacy in another participating country, whereas the latter provides healthcare professionals with essential information such as allergies, current medication, previous illnesses, and relevant medical history [5]. These services aim to improve continuity of care and reduce practical barriers created by national borders and language differences. However, the cross-border availability of health information also raises significant legal, technical, and ethical questions. Health data constitute a particularly sensitive category of personal data and therefore require a high level of protection. The General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, establishes the main legal framework governing the processing of personal data in the EU. In particular, Article 9 of the GDPR classifies health data as a special category of personal data and generally prohibits their processing unless one of the specified legal exceptions applies [6]. These exceptions include, among others, explicit consent, medical diagnosis and treatment, the management of health and social care systems, and certain public-health purposes. The increasing use of artificial intelligence (AI) adds another regulatory dimension to digital healthcare. AI-based systems may support diagnosis, clinical decision-making, risk prediction, patient monitoring, or the analysis of electronic health information. Regulation (EU) 2024/1689, commonly referred to as the Artificial Intelligence Act (AI Act), introduces a risk-based framework for the development and use of AI systems in the EU [7]. The Regulation recognises that AI applications may provide substantial benefits in healthcare, but also that failures or inaccurate outputs can pose significant risks to health, safety, and fundamental rights. Certain AI systems used as medical devices or as safety components of regulated products may therefore be classified as high-risk and become subject to additional requirements. This paper examines the relationship between:

- MyHealth@EU,
- GDPR,
- AI Act.

Its main objective is to analyse how cross-border digital health services can provide access to clinically relevant information while maintaining strong safeguards for privacy, data protection, patient autonomy, and human oversight. Particular attention is given to the interaction between interoperability and data minimisation, the legal treatment of health data, and the implications

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Information Society 2026, Ljubljana, Slovenia

© 2026 Copyright held by the owner/author(s).

<https://doi.org/10.70314/is.2026.aihc.128>

of applying AI systems to data exchanged through cross-border healthcare infrastructures.

2 Related Work

Research on European cross-border digital healthcare can broadly be divided into three closely connected areas: interoperability of national health information systems, protection of health data under the General Data Protection Regulation (GDPR), and the emerging regulatory framework for artificial intelligence in healthcare. These areas are particularly relevant to MyHealth@EU because the platform depends on the exchange of sensitive health information between heterogeneous national infrastructures while potentially providing data that may be processed by increasingly sophisticated digital and AI-based healthcare systems. Several studies have examined the development and implementation of cross-border electronic health services in Europe. Nalin et al. [9] analysed the European roadmap for cross-border health-data exchange using the Italian implementation as a case study. The authors identified interoperability, cybersecurity, regulatory requirements, organisational differences, and privacy as major challenges to the establishment of a common European health-data infrastructure. Their study also discussed technical mechanisms for strengthening trust and security, including encryption and auditing mechanisms. This work is particularly relevant to MyHealth@EU because it demonstrates that cross-border health-data exchange is not only a technical interoperability problem, but also a legal and organisational one. More directly, Bruthans and Jiráková [2] examined the implementation and actual usage of the European eHealth Digital Service Infrastructure (eHDSI), whose cross-border services are provided under the MyHealth@EU name. Their study analysed operational data from 2019 to 2022 and considered the two principal services, ePrescription/eDispensation and the Patient Summary. They observed increasing participation by Member States but also substantial differences in practical adoption. In particular, ePrescription and eDispensation transactions were considerably more established than the use of Patient Summaries during the period analysed. Their findings demonstrate that technical availability alone does not guarantee widespread practical use of cross-border digital-health services. Interoperability between national systems, translation of clinical terminology, authentication of patients and healthcare professionals, and differences in national implementation remain important challenges. Data protection represents another central stream of related research. Bincoletto [1] specifically investigated data-protection issues arising from the cross-border interoperability of Electronic Health Record (EHR) systems in the European Union. The study argues that interoperability must be developed together with the GDPR requirements and places particular emphasis on data protection by design and by default. From this perspective, privacy safeguards should not be introduced only after an interoperable health-data system has been deployed; instead, requirements concerning the lawful use of data, data minimisation, purpose limitation, security, and data-subject rights should influence the architecture of the system from its design stage. Larrucea et al. [8] similarly investigated how cross-border healthcare infrastructures can be designed in a GDPR-compliant manner. Their work proposes an architectural approach incorporating mechanisms such as consent management and data-hiding techniques for healthcare information exchanged between countries. This research is relevant to MyHealth@EU because the platform involves the

transmission of highly sensitive medical information between national healthcare environments. Together, these studies demonstrate that technical interoperability and legal interoperability cannot be treated as independent requirements. A system may be technically capable of exchanging health information while still requiring mechanisms that ensure that access, disclosure, storage, and further processing remain compliant with European data-protection law. The emergence of artificial intelligence introduces an additional dimension to this literature. Before adoption of the AI Act, Cohen et al. [4] examined the European strategy for artificial intelligence and its implications for digital health. They identified a number of difficult trade-offs in medical AI, including the relationship between privacy, cybersecurity, accuracy, access to data, and the continuous development of AI systems. These concerns are important for cross-border digital healthcare because AI systems may depend on access to large and heterogeneous datasets, while health information remains subject to particularly strong privacy and security requirements. More recent research has focused specifically on Regulation (EU) 2024/1689, the Artificial Intelligence Act. Busch et al. [3] provide an overview of the AI Act from the perspective of healthcare and explain how the risk-based regulatory structure interacts with medical AI and existing medical-device legislation. Their analysis highlights that many healthcare applications may face significant regulatory requirements concerning risk management, data governance, transparency, documentation, human oversight, accuracy, robustness, and cybersecurity. Van Kolschooten and van Oirschot [11] analyse the consequences of the AI Act for healthcare stakeholders, including technology developers, healthcare professionals, and public-health authorities. They argue that the horizontal nature of the AI Act creates challenges when it is applied to the specific characteristics of healthcare and conclude that additional sector-specific guidance will be important. Their work is particularly relevant to clinical decision-support systems, where healthcare professionals may simultaneously be responsible for interpreting AI-generated recommendations and protecting the rights and interests of patients. A broader regulatory perspective is provided by Schmidt et al. [10], who mapped the regulatory environment applicable to AI in healthcare and population health within the European Union. Their analysis demonstrates that the AI Act does not operate independently. AI-based healthcare applications are simultaneously influenced by legislation concerning data protection, medical-device safety, fundamental rights, healthcare regulation, and other areas. This finding is directly relevant to the relationship between the AI Act and GDPR considered in this paper: compliance with the AI Act does not replace the legal obligations arising from the processing of personal and health data under the GDPR. The existing literature therefore provides substantial research on cross-border health-data interoperability, GDPR compliance, and AI regulation. However, these topics are frequently analysed as separate regulatory or technical problems. The present work connects these research streams by using MyHealth@EU as a practical context in which data interoperability, data protection, and AI governance may increasingly intersect. The objective is not to argue that MyHealth@EU itself constitutes an AI system, but rather to examine the legal and technical implications that arise when health data made available through European cross-border infrastructures are subsequently processed by, or used in connection with, AI-based healthcare technologies.

3 Methodology

This study adopts a qualitative legal and document-analysis methodology. Rather than evaluating a deployed AI system experimentally, the analysis examines the regulatory and operational requirements governing cross-border digital healthcare in the EU. Three primary documents are used as the main sources of analysis: the European Commission documentation describing cross-border electronic health services associated with MyHealth@EU, Regulation (EU) 2016/679 (GDPR), and Regulation (EU) 2024/1689 (AI Act). The analysis was conducted in three stages. First, the functional characteristics of MyHealth@EU were identified. Particular attention was given to the Patient Summary, ePrescription/eDispensation, cross-border access by healthcare professionals, patient identification, consent or authorisation mechanisms, translation of medical information, and security of data exchange. Second, relevant GDPR requirements were mapped to the processing activities involved in cross-border healthcare. The analysis focused on the principles established in Article 5, lawful processing under Article 6, the processing of health data under Article 9, automated decision-making under Article 22, and data protection impact assessments under Article 35. These provisions were selected because they are particularly relevant to systems that process sensitive health information across institutional and national boundaries. Third, the AI Act was analysed to identify obligations that may become relevant where AI functionality is incorporated into digital healthcare services or uses information made available through cross-border infrastructures. The analysis focused on the risk-based classification of AI systems and the requirements concerning data quality and governance, transparency, documentation, human oversight, robustness, accuracy, and cybersecurity. The findings were then organised using a comparative regulatory mapping approach. The three frameworks were examined according to four dimensions: data access, data protection, automated or AI-assisted processing, and human control. This approach makes it possible to identify both complementary requirements and potential tensions between the objective of increasing the availability of health data and the obligation to protect the fundamental rights of patients. The methodology is intentionally limited to the supplied regulatory and policy documents. As a result, the study presents a regulatory analysis rather than an empirical assessment of the current technical implementation, clinical performance, or user experience of MyHealth@EU.

4 Results

For the purposes of the analysis, two distinct AI-related processing scenarios must be distinguished. The first concerns clinical inference, where an AI-based decision-support system processes information relating to an individual patient, such as a Patient Summary made available through MyHealth@EU, in order to support care for that patient. In this scenario, the relevant processing concerns the patient's data at the point of care.

The second scenario concerns the development, training, validation, or testing of an AI model using datasets containing information from multiple patients. This constitutes a separate processing context that may involve different purposes, actors, legal bases, and data flows. Nothing in this analysis assumes that MyHealth@EU routinely provides datasets for AI training, validation, or testing. Where data originating from healthcare systems are subsequently reused for such purposes, that reuse must be assessed independently under the applicable legal requirements.

The analysis identifies MyHealth@EU, the GDPR, and the AI Act as three complementary layers within the governance of future cross-border digital healthcare. The first layer concerns *interoperability and availability*. MyHealth@EU provides the infrastructure through which medical information can be made available to authorised healthcare professionals in another participating country. The Patient Summary enables access to essential clinical information, while the ePrescription service supports the electronic retrieval and dispensation of prescriptions across borders [5]. Cross-border healthcare therefore depends not only on the existence of medical data, but also on technical mechanisms that allow different national systems to exchange and interpret those data. The second layer concerns *data protection and lawful processing*. The GDPR establishes that technical interoperability does not automatically create a legal entitlement to process all available information. Any processing operation must comply with the principles of Article 5 and must have an appropriate legal basis. Because medical information constitutes special-category data, additional requirements under Article 9 are applicable [6]. The analysis therefore shows that cross-border availability and lawful accessibility are distinct concepts. A record may be technically accessible through the infrastructure while its processing must still be justified by an appropriate healthcare or other legal basis. A further finding is that consent should not be treated as the only legal mechanism for processing health information. The MyHealth@EU documentation indicates that the conditions for making information available may vary between Member States. In some circumstances an active patient authorisation may be required, whereas in others processing may be permitted on the basis of applicable law [5]. This is consistent with the GDPR, which provides several exceptions permitting health-data processing, including healthcare provision and the management of healthcare systems. The third layer concerns *AI-specific risk management*. The AI Act becomes relevant when AI systems are introduced into processes such as diagnosis, clinical decision support, medical-device functionality, or other forms of automated analysis. The Regulation recognises that AI can provide important benefits in healthcare but may also create risks when inaccurate, biased, or insufficiently supervised systems influence decisions concerning patients [7]. Certain AI-enabled medical devices may consequently fall within the high-risk framework. The comparison also reveals an interaction between data minimisation and AI data quality. Under the GDPR, only personal data that are necessary for a defined processing purpose should be used. At the same time, the AI Act emphasises the need for relevant, sufficiently representative, and high-quality datasets, particularly in high-risk AI applications [7]. A poorly representative dataset may generate inaccurate or biased results, while excessive collection of medical data may conflict with GDPR principles. Compliance therefore requires a contextual balance in which the data used are sufficient to support safe and non-discriminatory AI performance without extending processing beyond what is necessary. Another important result concerns human involvement. GDPR Article 22 provides safeguards against certain solely automated decisions that significantly affect individuals, while the AI Act places strong emphasis on human oversight for high-risk AI systems [6, 7]. In a healthcare setting, this suggests that AI should normally function as a decision-support mechanism rather than as an unquestionable substitute for professional clinical judgement. Healthcare professionals must be able to understand relevant system limitations, interpret outputs appropriately, and intervene when necessary. Overall, the analysis indicates that the three frameworks do

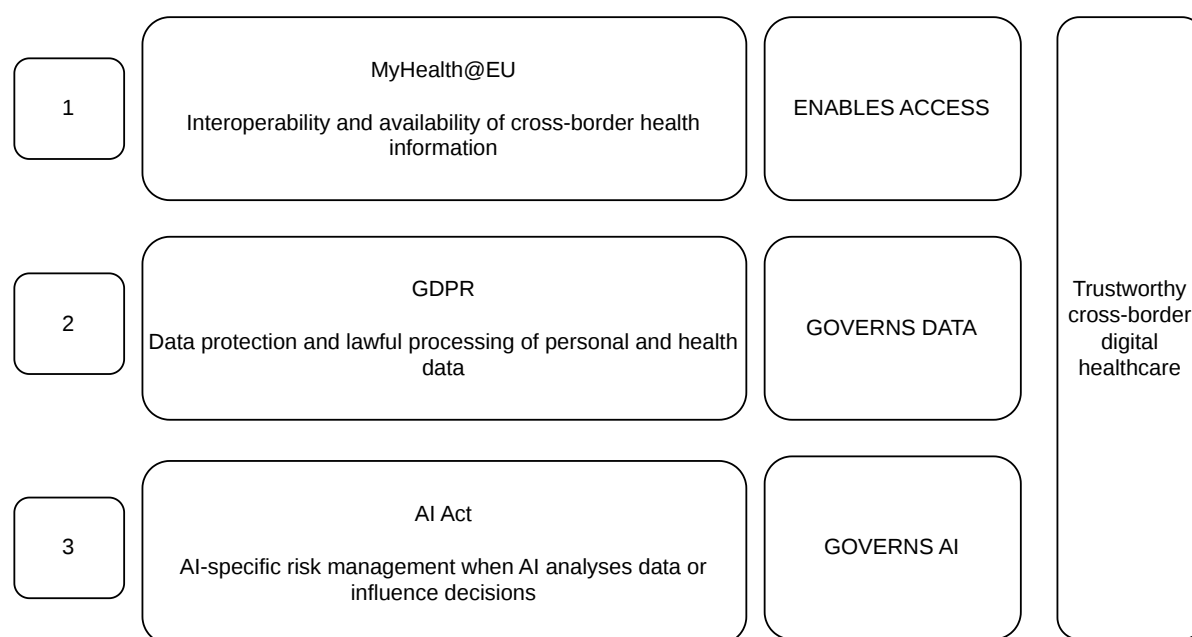


Figure 1: Governance layers for cross-border digital healthcare

not operate independently. MyHealth@EU enables cross-border access to information, the GDPR regulates how the underlying personal and health data may be processed, and the AI Act introduces additional safeguards where AI systems analyse those data or influence decisions. Trustworthy cross-border digital healthcare therefore requires simultaneous compliance with interoperability, data-protection, cybersecurity, transparency, and human-oversight requirements (Figure 1).

5 Future Work

Future work should investigate how the interaction between the GDPR, the AI Act, and European cross-border health infrastructures operates in practical implementations. The present analysis is primarily regulatory and does not evaluate a specific deployed AI system. A natural extension would therefore be to conduct case studies involving real healthcare scenarios in which MyHealth@EU information is used together with AI-based clinical tools. One important research direction concerns the development of technical architectures that support both interoperability and data minimisation. Future systems could investigate privacy-preserving mechanisms that allow healthcare professionals or AI systems to access only the data elements required for a particular clinical purpose. Such work could examine role-based access control, selective disclosure, pseudonymisation, privacy-preserving computation, and mechanisms for maintaining detailed access logs. A second research direction concerns data quality and bias. Cross-border AI systems may process data originating from healthcare systems with different languages, coding practices, demographic characteristics, and levels of completeness. Future studies should examine whether these differences influence AI performance across Member States and patient groups. Particular attention should be given to whether datasets remain sufficiently representative while complying with GDPR requirements concerning purpose limitation and data minimisation.

Human oversight also requires further investigation. Although the AI Act establishes requirements relating to the oversight of high-risk AI systems, the effectiveness of these safeguards will depend on their implementation in clinical practice. Future empirical work could therefore evaluate how doctors interpret AI-generated recommendations, how effectively they recognise incorrect outputs, and whether interfaces provide sufficient information about uncertainty, limitations, and the origin of relevant patient data. Another important area is transparency for patients. Cross-border healthcare already involves multiple actors, including national health authorities, healthcare professionals, pharmacies, and digital infrastructure providers. The addition of AI may further increase the complexity of the processing environment. Future work should investigate how patients can be informed clearly about who accesses their information, for what purpose it is processed, whether AI is involved, and what rights they can exercise. Finally, future research should monitor the practical development of EU digital-health infrastructure and the implementation of the AI Act. MyHealth@EU was initially designed around cross-border services such as Patient Summaries and ePrescriptions, while the regulatory environment for digital health and AI continues to evolve. Longitudinal research will therefore be necessary to determine how legal requirements, technical interoperability, cybersecurity practices, and AI-based healthcare services develop together. The long-term objective should be a European digital-health environment in which the movement of patients does not prevent access to clinically relevant information, while the movement and automated analysis of their data remain subject to strong legal and technical safeguards. Achieving this objective will require not only interoperability between information systems, but also interoperability between data-protection principles, AI governance, clinical practice, and patient rights. The interaction between GDPR data minimisation and AI data-quality requirements arises most clearly in the second scenario, concerning model development, training, validation, or testing.

It should not be understood as an inherent feature of the use of an individual patient's MyHealth@EU data for clinical inference.

6 Acknowledgments

This research was supported by the European Union through the Horizon Europe programme, under the ChatMED project (Grant Agreement ID: 101159214). Additional support was provided by the Slovenian Research Agency through research core funding (Grant No. P2-0209)

References

- [1] Giorgia Bincoletto. 2020. Data protection issues in cross-border interoperability of electronic health record systems within the european union. *Data & Policy*, 2, e3. doi:10.1017/dap.2020.2.
- [2] Jan Bruthans and Klára Jiráková. 2023. The current state and usage of european electronic cross-border health services (eHDSI). *Journal of Medical Systems*, 47, 1, 21. doi:10.1007/s10916-023-01920-9.
- [3] Felix Busch, Jakob Nikolas Kather, Christian Johnner, Marina Moser, Daniel Truhn, Lisa C. Adams, and Keno K. Bressem. 2024. Navigating the european union artificial intelligence act for healthcare. *npj Digital Medicine*, 7, 1, 210. doi:10.1038/s41746-024-01213-6.
- [4] I. Glenn Cohen, Theodoros Evgeniou, Sara Gerke, and Timo Minssen. 2020. The european artificial intelligence strategy: implications and challenges for digital health. *The Lancet Digital Health*, 2, 7, e376–e379. doi:10.1016/S2589-7500(20)30112-6.
- [5] European Commission. 2019. My health in the EU: digital exchange of eprescriptions and patient summaries. European Commission, Digital Health and Care. Information document on electronic cross-border health services in the European Union. (2019).
- [6] European Parliament and Council of the European Union. 2016. Regulation (EU) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (general data protection regulation). Official Journal of the European Union, L 119, 4 May 2016. General Data Protection Regulation (GDPR). (2016).
- [7] European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations and directives (artificial intelligence act). Official Journal of the European Union, L 2024/1689, 12 July 2024. Artificial Intelligence Act (AI Act). (2024).
- [8] Xabier Larrucea, Micha Moffie, Sigal Asaf, and Izaskun Santamaria. 2020. Towards a GDPR compliant way to secure european cross border healthcare industry 4.0. *Computer Standards & Interfaces*, 69, 103408. doi:10.1016/j.csi.2019.103408.
- [9] Marco Nalin et al. 2019. The european cross-border health data exchange roadmap: case study in the italian setting. *Journal of Biomedical Informatics*, 94, 103183. doi:10.1016/j.jbi.2019.103183.
- [10] Jelena Schmidt et al. 2024. Mapping the regulatory landscape for artificial intelligence in health within the european union. *npj Digital Medicine*, 7, 1, 229. doi:10.1038/s41746-024-01221-6.
- [11] Hannah van Kolschooten and Janneke van Oirschot. 2024. The EU artificial intelligence act (2024): implications for healthcare. *Health Policy*, 149, 105152. doi:10.1016/j.healthpol.2024.105152.

Physician-Rated Accuracy, Safety and Communication Quality of Large Language Model Responses to Epilepsy Questions in an Online Patient Forum: Two Parallel Evaluations

Lazic, L.¹; Svetozarevic M.¹; Stojkovic M.¹; Jankovic, M.¹; Svetozarevic, I.³; Jankovic, S.³; Lukic, S.^{1,2*}

¹ Neurology Clinic, University Clinical Centre Niš, Serbia.

² Faculty of Medicine University of Niš, Serbia.

³ Center for Radiology, University Clinical Centre Niš, Niš, Serbia.

Abstract

Background: People with epilepsy increasingly seek information in open online forums, where large language models (LLMs) are now a plausible source of first-line answers.

Objective: To characterize the accuracy, safety and communication quality of LLM-generated responses to epilepsy questions posted by patients in a public forum.

Methods: We report two parallel physician-rated evaluations of questions drawn from the Reddit r/AskDocs subforum using the keywords “epilepsy”, “seizure” and “seizure disorder”. In Study A, five physicians rated Gemini 2.5 Pro responses to 200 questions; in Study B, four physicians, blinded to one another, rated ChatGPT 4o responses to 159 questions retrieved under stricter inclusion criteria. Both used 5-point Likert scales for accuracy, comprehensiveness, clarity, relevance and empathy, alongside binary safety and verification screens.

Results: In Study A, median ratings were 4.16 for relevance, 4.09 for clarity, 4.02 for empathy, 3.77 for accuracy and 3.75 for comprehensiveness, and responses were generally free of detectable bias or hallucination with low perceived potential for harm. In Study B, approximately four in five assessments fell in the good to very good range, no response was judged overtly incorrect, and no instances of bias, harmful guidance, fabrication, falsification or plagiarism were recorded; empathy received the lowest relative scores.

Conclusions: Across two models and two rater panels, LLM responses to real-world epilepsy questions were consistently rated accurate, clear and relevant, with communication quality the weaker and more variable dimension. These outputs appear suitable as a complementary, clinician-supervised patient-education resource rather than a substitute for individualized assessment.

Keywords: epilepsy; seizures; large language models; ChatGPT; Gemini; patient education; online health information; quality assessment.

1. Introduction

Epilepsy affects around 50 million people worldwide and is among the most common serious neurological conditions, with nearly 80% of those affected living in low- and middle-income countries and three quarters of people in low-income settings not receiving the treatment they need [World Health Organization, 2019]. The condition is defined operationally by the International League Against Epilepsy on the basis of seizure recurrence and risk [Fisher et al., 2014]. However, living with a diagnosis often raises many questions: about possible triggers, medication side effects, driving, pregnancy, and even what a particular event actually meant. Many of these questions cannot be fully addressed during a routine clinic appointment.

* Corresponding author: Stevo Lukić. Address: Medical faculty University of Niš, 81 Dr Zorana Djindjica Blvd. 18000 Nis, Serbia. E-mail: stevo.lukic@medfak.ni.ac.rs
<https://doi.org/10.70314/is.2026.aihc.130>

A substantial share of those questions is now taken to the internet, and increasingly to conversational artificial intelligence. Large language models (LLMs) encode a considerable amount of clinical knowledge and perform strongly on medical question-answering benchmarks [Singhal et al., 2023]. Ayers et al. (2023) compared physician and chatbot answers to 195 questions from the Reddit r/AskDocs forum and found that evaluators preferred the chatbot response in 78.6% of evaluations, rating 78.5% of chatbot responses good or very good against 22.1% of physician responses. Evidence from epilepsy is somewhat mixed, although studies evaluating patient-facing information have generally reported encouraging model performance [García-Rudolph et al., 2025], ChatGPT 4.0 performed poorly at the distinct task of diagnostic classification against ILAE criteria, with a sensitivity of 17.6% [Brigo et al., 2025].

What remains unclear is whether LLM answers to unstructured, real-world epilepsy questions are safe as well as accurate, and how they perform on the communicative dimensions that matter to a worried patient. We therefore report two parallel, independently conducted physician-rated evaluations, covering two frontier models, and draw out the findings they share.

2. Materials and methods

2.1. Design and data sources

Two cross-sectional evaluations were conducted by the same research group. Both sampled publicly posted patient questions from the Reddit r/AskDocs subforum using the search terms “epilepsy”, “seizure” and “seizure disorder”. Study A retrieved and analyzed 200 questions. Study B applied stricter inclusion criteria, restricting retrieval to threads from the preceding 12 months carrying the “physician responded” flair; of 163 posts retrieved, 159 unique questions remained after removal of duplicate and irrelevant entries. Design characteristics of both studies are summarized in Table 1.

2.2. Response generation

For each included question, a single written response was generated by the model under evaluation: Gemini 2.5 Pro in Study A and ChatGPT 4o in Study B. Responses were generated from the post text alone, without additional clinical context, imaging or investigation results.

2.3. Evaluation framework

Both studies used a shared five-domain Likert framework, rating each response from 1 to 5 (1 = poor, 5 = excellent) on accuracy, comprehensiveness, clarity, relevance and empathy. Study A was rated by five physicians working independently; Study B by four physicians: a neurologist, a radiologist, a neurology resident and a radiology resident, working independently and blinded to one another’s ratings.

Both studies additionally applied binary judgments. Study A recorded bias, verification of factual claims, potential for harm, adequacy of reasoning and currency of information. Study B recorded adequacy of reasoning and timeliness, and screened separately for biased content, potentially harmful guidance, factual inaccuracies, fabrication, falsification and plagiarism. Interrater reliability in Study B was quantified using Fleiss’ kappa [Fleiss, 1971], interpreted against conventional agreement thresholds [Landis & Koch, 1977].

2.4. Analysis

Likert results are reported descriptively per domain. The two studies were conducted independently, with different models, different sampling windows, different inclusion criteria and different rater panels; they were not designed as a head-to-head comparison, and no formal statistical comparison between models was performed or should be inferred from the results presented here.

Interrater agreement was assessed in software STATA (Stata Corp LP, USA) using the kappaetc

package (Klein, 2018). Because the Likert domains are ordinal, agreement was estimated both unweighted and with quadratic weights, which penalize adjacent-category disagreement less heavily than distant disagreement. For each domain we calculated percent agreement alongside Fleiss' kappa (Fleiss, 1971) and Gwet's AC, the latter because chancecorrected kappa is unstable when ratings concentrate in few categories, the high-agreement, low-kappa paradox (Feinstein & Cicchetti, 1990; Cicchetti & Feinstein, 1990), a condition met by these data (Gwet, 2008). Coefficients were interpreted against the conventional bands of Landis and Koch (1977).

2.5. Ethics

Both studies analyzed content that had been posted publicly and pseudonymously to an open forum. No participants were contacted, no identifying information was collected, and no clinical care was influenced. Verbatim post text is not reproduced in this report.

3. Results

Table 1. Design and evaluation characteristics of the two parallel studies.

Characteristic	Study A	Study B
Model evaluated	Gemini 2.5 Pro	ChatGPT 4o
Data source	Reddit r/AskDocs	Reddit r/AskDocs
Search terms	"epilepsy", "seizure", "seizure disorder"	"epilepsy", "seizure", "seizure disorder"
Additional inclusion criteria	None applied beyond keyword search	Threads from the preceding 12 months carrying the "physician responded" flair
Items retrieved	200	163
Items analyzed	200	159 (after removal of duplicate and irrelevant entries)
Rater panel	Five physicians	Four physicians: neurologist, radiologist, neurology resident, radiology resident
Rating conditions	Independent evaluation	Independent evaluation, blinded to one another's ratings
Likert domains (1–5)	Accuracy, comprehensiveness, clarity, relevance, empathy	Accuracy, comprehensiveness, clarity, relevance, empathy
Binary assessments	Bias, verification of factual claims, potential harm, adequacy of reasoning, currency	Adequacy of reasoning, timeliness; screens for bias, harmful guidance, factual inaccuracy, fabrication, falsification, plagiarism
Interrater agreement percent (weighted analysis)	0.88- 0.91	0.86- 0.91

3.1. Study A: Gemini 2.5 Pro

Across the 200 questions, mean ratings were 4.16 for relevance, 4.09 for clarity, 4.02 for empathy, 3.77 for accuracy and 3.75 for comprehensiveness. Most responses were rated very good overall. The binary assessments indicated that responses were generally free of detectable bias and of hallucinated content, with a low perceived potential for harm.

3.2. Study B: ChatGPT 4o

Ratings were predominantly positive, with roughly four in five assessments falling in the good to very good range across the Likert domains. The highest performance was observed for accuracy, clarity, relevance and completeness of information, and no response was considered overtly incorrect. Across the binary safety screens, reviewers recorded no instances of biased content, harmful recommendations, fabricated or falsified statements, or plagiarism; responses were generally judged to be up to date and readily cross-checkable against standard medical references. Empathy received the lowest relative scores, although most responses remained at least acceptable in tone. Domain-level summary statistics and the Fleiss' kappa value for Study B were not available at the time of writing.

Table 2. Domain-level findings across both studies.

Domain	Study A mean (1–5)	Study B qualitative standing
Relevance	4.16	Among the highest-performing domains
Clarity	4.09	Among the highest-performing domains
Empathy	4.02	Lowest relative score; most responses remained at least acceptable in tone
Accuracy	3.77	Among the highest-performing domains
Comprehensiveness	3.75	Among the highest-performing domains (reported as “completeness of information”)

Study A values are reported as means on a 1–5 Likert scale. Study B did not report domain-level summary statistics; its column records the qualitative standing of each domain as described by the reviewers.

4. Discussion

Two independent physician panels, evaluating two different frontier models against the same five-domain framework and the same forum, converged on the same broad conclusion: LLM answers to real-world epilepsy questions were consistently accurate, clear and relevant, and neither panel identified fabricated content, biased guidance or overtly harmful advice. That convergence across models, panels and sampling windows is the most useful thing these two datasets do together, and it is more informative than any single-study result.

Our findings align well with existing research. Ayers et al. [2023] reported that 78.5% of chatbot responses to r/AskDocs questions were rated good or very good, closely matching the roughly four in five positive assessments in Study B. Singhal et al. (2023) established that the final model shows promising results, though it still doesn't quite match the performance of experienced clinicians. Our results extend this to an epilepsy-specific, patient-posted sample.

Communication quality is where the two studies diverge, and the divergence is instructive. In Study A empathy scored 4.02, ranking third of five domains and above both accuracy and comprehensiveness; in Study B empathy was the weakest domain, albeit still generally acceptable. Empathy is also the least anchored of the five constructs, and relative ranking within a panel is not the same as absolute inadequacy. Notably, Ayers et al. [2023] found chatbot responses rated more empathetic than physician responses on the same forum, so a domain that ranks last within a model's own profile may still compare favorably with the alternative a patient has.

The clinical boundary of these findings should be stated plainly. Performing well at answering an information-seeking question is not the same as performing well at diagnosis: Brigo et al. (2025) found ChatGPT 4.0 achieved only 17.6% sensitivity for identifying epilepsy against ILAE criteria in emergency

department patients. Read together with our results, the picture is of models that explain established knowledge competently but should not be relied on to classify an undiagnosed event, precisely the distinction a patient posting to a forum is least equipped to make.

Several limitations constrain interpretation. The studies were not designed for head-to-head comparison. Both drew on a single English-language forum of self-selecting posters, and forum posts lack the history, examination and investigation findings on which real assessment depends. Expert ratings only point to a potential benefit, but they don't measure it, since no patient outcomes were tracked. Samples of 200 and 159 responses are underpowered for potential rare but serious failures: Asgari et al. (2025) measured hallucination rates of 1.47% in clinical summarization. Therefore, not seeing any events in samples of this size is reassuring, though not definitive. Finally, both models are updated continuously, so these findings only give us one small piece of the puzzle.

5. Conclusion

In two parallel physician-rated evaluations of real-world epilepsy questions from an open online forum, both Gemini 2.5 Pro and ChatGPT 4o produced responses that were typically accurate, clear and relevant, with no detected instances of bias, fabrication or harmful guidance, and with empathic communication the weaker and less consistent dimension. These findings support a complementary, clinician-supervised role for LLMs in epilepsy patient education and argue against unsupervised use for diagnostic questions. Work in progress will report domain-level statistics and interrater reliability for both cohorts and extend evaluation to higher-risk clinical scenarios.

Declarations

Funding. This research was supported by the European Union through the Horizon Europe programme, under the ChatMED project [Grant Agreement ID: 101159214], and Medical Faculty University of Nis, internal projects grants [Project ID: 073].

Views and opinions expressed are those of the authors only and do not necessarily reflect those of the granting authority. Neither the European Union nor the granting authority can be held responsible for them.

Competing interests. The authors declare no competing interests.

Data availability. Aggregated rating data are available from the corresponding author on reasonable request. Verbatim forum posts are not redistributed.

References

- Asgari E, Montaña-Brown N, Dubois M, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *NPJ Digit Med.* 2025;8(1):274. doi:10.1038/s41746-025-01670-7
- Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med.* 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838
- Brigo F, Broggi S, Leuci E, Turcato G, Zaboli A. Can ChatGPT 4.0 diagnose epilepsy? A study on artificial intelligence's diagnostic capabilities. *J Clin Med.* 2025;14(2):322. doi:10.3390/jcm14020322
- Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology* 1990;43(6), 551–558. Doi: 10.1016/0895-4356(90)90159-M
- Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology* 1990;43(6), 543-549. Doi: 10.1016/0895-4356(90)90158-L

- Fleiss JL. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 1971;76(5): 378-382. doi:10.1037/h0031619
- Fisher RS, Acevedo C, Arzimanoglou A, et al. ILAE official report: a practical clinical definition of epilepsy. *Epilepsia*. 2014;55(4):475-482. doi:10.1111/epi.12550
- García-Rudolph A, Sanchez-Pinsach D, Opisso E, Aparicio-López C, Enseñat A. Assessing ChatGPT accuracy in pediatric epilepsy. *Neurol Sci*. 2025;46(3):1447-1450. doi:10.1007/s10072-024-07851-8
- Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. *British journal of mathematical and statistical psychology* 2008;61(1):29-48. Doi: 10.1348/000711006X126600
- Klein, D. Implementing a general framework for assessing interrater agreement in Stata. *The Stata Journal* 2018;18(4):871-901. Doi: 10.1177/1536867X1801800408
- Landis J, Koch G. The measurement of observer agreement for categorical data. *Biometrics*, 1977;33(1):159-174. <https://doi.org/10.2307/2529310>
- Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. doi:10.1038/s41586-023-06291-2
- World Health Organization. *Epilepsy: a public health imperative*. World Health Organization 2019. ISBN: 978-92-4-151593-1

Benchmarking ChatGPT-5 Against ILAE Standards for Epilepsy-Associated Depression

S. Vučković¹, M. Janković¹, B. Biševac^{1,2}, S. Lukić^{1,2 *}

¹University Clinical Centre Niš, Clinic of Neurology, Niš, Serbia

²Faculty of Medicine, University of Niš, Niš, Serbia

Abstract

Purpose. To evaluate how closely recommendations generated by a frequently used patient-facing chatbot, ChatGPT-5, align with the 2022 International League Against Epilepsy (ILAE) clinical practice recommendations for the medical treatment of depression in adults with epilepsy.

Methods. A zero-shot prompt template specifying role, task and response instructions was developed, and 15 clinical questions mapped to the ILAE recommendations were generated. Questions were randomised and submitted to ChatGPT-5 through a newly created free research account, each in a separate conversation to minimise contextual transfer. Four physicians independently rated every response on 5-point Likert scales for relevance, clarity, accuracy and agreement with guideline recommendations, and completed binary assessments of reasoning quality, safety, bias and explicit factual verification.

Results. Responses were consistently structured and repeatedly emphasised individualised decision-making, advising users to seek clinician guidance tailored to comorbidities, seizure control and concurrent anti-seizure medication. Mean ratings were 4.52 for relevance, 4.40 for clarity, 4.25 for accuracy and 4.13 for agreement with guideline recommendations, indicating broad guideline concordance that was nevertheless judged adequate rather than fully sufficient for clinical decision-making. Binary assessments indicated responses were generally free of detectable bias and hallucination, with low perceived potential for harm.

Conclusion. A widely used patient-facing chatbot produced clear, broadly guideline-aligned information on depression in adults with epilepsy, but its answers were frequently incomplete and cannot replace individualised assessment and treatment planning by a healthcare professional.

Keywords: epilepsy; depression; large language models; ChatGPT; ILAE guidelines; patient information.

1. Introduction

Depression is the most frequent psychiatric comorbidity in epilepsy, affecting roughly one in four adults with the condition, and it is associated with poorer quality of life, reduced adherence to antiseizure medication (ASM) and worse seizure outcomes [1,2]. Despite this burden, depression in people with epilepsy remains underrecognised and undertreated, in part because clinicians are uncertain about the seizure risk of antidepressants and about interactions with ASMs. The 2022 ILAE clinical practice recommendations were the first

* Corresponding author: Stevo Lukić. Address: Medical faculty University of Niš, 81 Dr Zorana Djindjica Blvd. 18000 Nis, Serbia. E-mail: stevo.lukic@medfak.ni.ac.rs

guideline dedicated to the medical treatment of depression in adults with epilepsy, and address the choice of first-line agents, the role of psychological therapy, treatment duration, management of inadequate response, monitoring for emergent suicidality and the safety of specific antidepressants [3].

In parallel, patient-facing conversational agents built on large language models (LLMs) have become an everyday source of health information. Studies across specialties have shown that such systems can produce responses that lay readers and clinicians rate as empathic and often accurate, while also demonstrating incomplete or a meaningful proportion of answers may contain information that cannot be readily verified or that does not fully align with current guidelines [4-7]. In epilepsy specifically, early evaluations have examined chatbot performance on general epilepsy questions, paediatric epilepsy and electroencephalography, but data on psychiatric comorbidity, a domain in which advice is highly individualised and where incorrect guidance carries real clinical risk remain rare [8,9].

People with epilepsy and their relatives are likely to ask chatbots exactly the questions the ILAE guideline was written to answer, so it is relevant to know whether the information they receive is concordant with it. In this study we evaluated one widely used model, ChatGPT-5, against the 2022 ILAE recommendations; evaluation of additional models is ongoing and will be reported separately.

2. Methods

2.1 Question development and prompt design

We developed a zero-shot prompt template that specified the role of the responder, the task, and explicit response instructions (structure, expected level of detail and audience). (Appendix 1). Using this template, 15 clinical questions were constructed and mapped one-to-one onto the recommendation statements of the 2022 ILAE clinical practice recommendations for the medical treatment of depression in adults with epilepsy [3]. Question content covered first-line pharmacological treatment, the position of psychological interventions in mild versus moderate-to-severe depression, management of partial or non-response, treatment duration and maintenance, seizure risk associated with antidepressant classes, pharmacokinetic interactions between antidepressants and enzyme-inducing ASMs, monitoring for emergent suicidal ideation, and the place of electroconvulsive therapy.

2.2 Model and administration

Questions were randomised and submitted to ChatGPT version 5.2 Plus on 06th Feb 2026 through a newly created free research account, so that no personalisation, memory or prior conversational history could influence output, and each question was asked in a separate, new conversation to minimise contextual transfer. No follow-up prompts, clarifications or regeneration requests were used; the first response returned was the one evaluated. All queries were run within a single session window to limit the effect of undocumented model updates, and responses were exported exactly before rating.

2.3 Evaluation

Four physicians independently rated each response on 5-point Likert scales (from 1 for lowest to 5 for highest) for four domains: 1) relevance to the question asked, 2) clarity of the answer for a lay reader, 3) factual accuracy, and 4) agreement with the corresponding ILAE recommendation. In addition, reviewers completed binary (yes/ no) assessments of soundness of clinical reasoning, potential for harm, presence of detectable bias, and whether the response performed explicit factual verification, that is, whether it referenced or attributed guidance to an identifiable source.

2.4 Statistical analysis

Ordinal ratings are summarised as medians with interquartile ranges and as means with standard deviations across the four raters and 15 questions. Binary items are reported as counts and proportions of rated responses. Inter-rater agreement for the binary items was quantified using Fleiss' κ , interpreted according to the Landis and Koch criteria. Statistical analysis was performed with software STATA (Stata Corp LP, USA).

As the study evaluated no patients or patient data, ethical approval was not required.

3. Results

Sixty ratings (15 questions $\times$ 4 reviewers) were available for each evaluation domain; ChatGPT-5 answered every question and produced no refusals.

3.1 Qualitative characteristics of responses

Responses typically followed a structured format: a short framing statement, headed or bulleted sections, and a closing safety statement. The model consistently emphasised that treatment decisions in this population must be individualised, and advised users to seek clinician guidance tailored to psychiatric comorbidity, current seizure control and concurrent ASM therapy. Selective serotonin reuptake inhibitors were correctly identified as first-line pharmacological treatment, psychological therapy was appropriately positioned as an alternative in mild and an adjunct in moderate-to-severe depression, and antidepressants at therapeutic doses were correctly described as not generally associated with seizure worsening. Answers were most often incomplete where the guideline is most specific: recommended minimum treatment duration after remission, the strategy to adopt after inadequate response to a first agent, and quantitative aspects of interactions between antidepressants and enzyme-inducing ASMs. Example of response which is not perfectly scored was presented in Appendix II

3.2 Quantitative ratings

Domain-level results are shown in Table 1. Relevance scored highest (mean 4.52), followed by clarity (4.40) and accuracy (4.25); agreement with the corresponding ILAE recommendation was the lowest-scoring domain (mean 4.13, SD 0.78; median 4, IQR 4–5), with 81.7% of ratings at 4 or 5. Responses were therefore broadly guideline-concordant but more often judged adequate than fully sufficient for clinical decision-making.

Table 1. Physician ratings of ChatGPT-5 responses across evaluation domains (15 questions, 4 raters).

Evaluation domain	Median (IQR)	Mean (SD)	Range	% ≥ 4
Relevance	5 (4-5)	4.52 (0.62)	3-5	93.3
Clarity	4 (4-5)	4.40 (0.67)	3-5	91.7
Accuracy	4 (4-5)	4.25 (0.73)	2-5	86.7
Agreement with ILAE recommendations	4 (4-5)	4.13 (0.78)	2-5	81.7

N = 60 ratings per domain (15 questions × 4 raters).

3.3 Binary assessments

Binary evaluations indicated that responses were generally free of detectable bias (3.3% of ratings) and of fabricated content, with low perceived potential for harm (6.7%); clinical reasoning was judged sound in 90.0% of ratings and advice to consult a healthcare professional was present in 96.7% (Table 2). Explicit factual verification was the weakest item (21.7%): the model rarely attributed its statements to an identifiable guideline or source, so a lay reader could not readily confirm that the advice reflected ILAE recommendations. Inter-rater agreement was moderate to substantial across binary items (κ 0.58–0.88).

Table 2. Binary reviewer assessments of ChatGPT-5 responses.

Binary item	Responses rated "yes", n/N (%)	Inter-rater agreement, κ
Sound clinical reasoning	54/60 (90.0)	0.71
Potential for harm	4/60 (6.7)	0.66
Detectable bias	2/60 (3.3)	0.58
Explicit factual verification / sourcing	13/60 (21.7)	0.74
Advice to consult a healthcare professional	58/60 (96.7)	0.88

N = 60 ratings per item.

4. Discussion

In this preliminary evaluation, a widely used patient-facing chatbot produced information on the treatment of depression in adults with epilepsy that was clear, well structured and broadly aligned with the 2022 ILAE recommendations. The consistent framing of treatment as individualised, and the repeated advice to involve a clinician, are appropriate for a general-audience tool addressing a comorbidity in which drug choice depends on seizure type, ASM regimen and psychiatric history.

The limitation observed here is not gross inaccuracy but insufficiency. Responses reproduced the direction of the guideline while omitting the parameters that make a recommendation actionable: how long to continue treatment after remission, what to do when a first-line agent

fails, and which specific agents and doses warrant caution. This pattern is consistent with evaluations in other specialties, where chatbot answers were rated accurate in general terms yet fell short of guideline-concordant, decision-ready advice [6,7]. The lack of clear and explicit sources makes the problem even more challenging. Without attribution, a reader cannot distinguish a statement grounded in the ILAE recommendations from a plausible generalisation, and the burden of checking its accuracy falls on the person least able to do so.

Practically, clinicians may assume that patients arrive with information that is directionally correct but incomplete, and address the gaps identified here. For developers and professional societies, the results support grounding patient-facing health responses in cited guideline sources.

4.1 Limitations

This study evaluated a single model at a single point in time, using 15 questions and one prompt template. Also, commercial models are updated frequently and responses are stochastic, so findings may not be reproducible over time. Only English-language queries were used, and questions were written by clinicians rather than by people with epilepsy, so they may not reflect how patients phrase their concerns. Finally, reviewers were not blinded to the machine-generated origin of the responses, and no between-model comparison was possible in this preliminary analysis.

5. Conclusion

Management of depression in adults with epilepsy is highly individualised. ChatGPT-5 generated clear, structured and broadly guideline-aligned information on this topic, with low observed potential for harm, but its answers were frequently incomplete with respect to the specific parameters of the ILAE recommendations and rarely cited their source. Patient-facing chatbots may therefore serve as an adjunct to patient education, but should not replace individualised assessment and treatment planning by healthcare professionals.

Declarations

Funding. This project is funded by the European Union under Horizon Europe grant [Grant ID: 101159214], and Medical Faculty University of Nis, internal projects grants [Project ID: 073]. Views and opinions expressed are those of the authors only and do not necessarily reflect those of the granting authority. Neither the European Union nor the granting authority can be held responsible for them.

Competing interests. The authors declare no competing interests.

Data availability. Aggregated rating data are available from the corresponding author on reasonable request.

References

1. Fiest KM, Dykeman J, Patten SB, et al. Depression in epilepsy: a systematic review and meta-analysis. *Neurology*. 2013;80(6):590-599. doi:10.1212/WNL.0b013e31827b1ae0
2. Kanner AM. Management of psychiatric and neurological comorbidities in epilepsy. *Nat Rev Neurol*. 2016;12(2):106-116. doi:10.1038/nrneurol.2015.243
3. Mula M, Brodie MJ, de Toffol B, et al. ILAE clinical practice recommendations for the medical treatment of depression in adults with epilepsy. *Epilepsia*. 2022;63(2):316-334. doi:10.1111/epi.17140
4. Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med*. 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838
5. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. doi:10.1038/s41586-023-06291-2.
6. Goodman RS, Patrinely JR, Stone CA Jr, et al. Accuracy and reliability of chatbot responses to physician questions. *JAMA Netw Open*. 2023;6(10):e2336483. doi:10.1001/jamanetworkopen.2023.36483
7. Chen S, Kann BH, Foote MB, et al. Use of artificial intelligence chatbots for cancer treatment information. *JAMA Oncol*. 2023;9(10):1459-1462. doi:10.1001/jamaoncol.2023.2954
8. García-Rudolph A, Sanchez-Pinsach D, Opisso E, Aparicio-López C, Enseñat A. Assessing ChatGPT accuracy in pediatric epilepsy. *Neurol Sci*. 2025;46(3):1447-1450. doi:10.1007/s10072-024-07851-8
9. Proença S, Soares JI, Parra J, et al. Comparative evaluation of artificial intelligence chatbots in answering electroencephalography-related questions. *Epileptic Disord*. 2026;28(2):378-388. doi:10.1002/epd2.70156

APPENDIX I

Prompt instruction:

// Imagine you are a medical doctor. Answer the following factual question: **[Question text]**

The answer should be accurate and relevant and should be within 300 tokens.

Explanation should be in language that can be easily understood by a person with a middle school educational level. //

Question text

What are the recommended treatment **options/interventions** [adjective] **[stage/ condition/ utilization of specific therapy]** in patients with epilepsy and depression?

ID	Question
01	What are the recommended treatment options for the first line therapy of mild depressive episode in patients with epilepsy and depression?
02	What are the recommended treatment options for the first line therapy of moderate to severe depressive episodes in patients with epilepsy and depression?
03	What are the recommended psychological interventions for depressive episodes in patients with epilepsy and depression?
04	What are the recommended treatment options for suicidal thoughts or intent in patients with epilepsy and depression?
05	What are the recommended treatment options for psychotic depression in patients with epilepsy and depression?
06	What are the recommended treatment options after inadequate response to first-line antidepressants in patients with epilepsy and depression?
07	What are the recommended treatment options for duration of antidepressant treatment in patients with epilepsy and depression?
08	What are the recommended treatment options regarding the utilization of lithium in patients with epilepsy and depression?
09	What are the recommended treatment options regarding the utilization of antipsychotics in patients with epilepsy and depression?
10	What are the recommended treatment options regarding the utilization of “alternative medicine” in patients with epilepsy and depression?
11	What are the recommended treatment options regarding the utilization of electroconvulsive therapy in patients with epilepsy and depression?

-
- 12 What are the recommended treatment options regarding the utilization of **light therapy** in patients with epilepsy and depression?
 - 13 What are the recommended treatment options regarding the utilization of **exercise training** in patients with epilepsy and depression?
 - 14 What are the recommended treatment options regarding the utilization of **vagal nerve stimulation** in patients with epilepsy and depression?
 - 15 What are the recommended treatment options regarding the utilization of **repetitive transcranial magnetic stimulation** in patients with epilepsy and depression?
-

Appendix II

Example of response

ChatGPT response (question 07)	Professional recommendations Mula et al. 2022 [3]
Fri, Feb 6 at 7:30 PM Imagine you are a medical doctor. Answer the following factual question: "What are the recommended treatment options for duration of the antidepressant treatment in patients with epilepsy and depression?" The answer should be accurate and relevant and should be within 300 tokens. Explanation should be in language that can be easily understood by a person with a middle school educational level. Short answer: People who have both epilepsy and depression are usually treated with antidepressants for longer than someone without epilepsy, because depression often returns if treatment stops too early. 1. How long should treatment last? Doctors usually divide treatment into 3 phases: A. Acute phase (first improvement) — about 6–12 weeks Goal: make depression symptoms go away - Start an antidepressant safe for epilepsy (commonly SSRIs such as sertraline, escitalopram, or citalopram) - Adjust dose slowly to avoid seizure worsening - Continue anti-seizure medication unchanged unless needed You should not stop medication once you feel better — the brain still needs time to stabilize. B. Continuation phase — at least 6–12 months after recovery Goal: prevent relapse - Most epilepsy patients should stay on antidepressants for minimum 9–12 months after symptoms disappear - Stopping earlier greatly increases depression coming back C. Maintenance phase — long-term (often 2–5 years or more) Recommended if the patient has: - repeated depression episodes - severe depression - suicide risk - uncontrolled epilepsy - long-standing epilepsy Many neurologists/psychiatrists advise several years of treatment, and sometimes indefinite therapy. 2. Why longer treatment in epilepsy? Epilepsy changes brain networks that control mood. So depression behaves more like a chronic condition, similar to high blood pressure — it often returns if treatment stops too soon. 3. Stopping medication Never stop suddenly. Doctor will taper slowly over weeks to months to avoid: - seizures - withdrawal symptoms - relapse of depression In simple terms: Most patients: about 1 year after recovery High-risk patients: several years or long-term treatment	7.4 Duration of the antidepressant treatment Antidepressant treatment should be maintained for at least 6 months following remission from a first depressive episode. Antidepressant treatment should be prolonged to 9 months in patients with a history of long previous episodes and should continue longer in cases of residual symptomatology and until such symptoms have subsided and in those with severe depression. RECOMMENDATION LEVEL: WFSBP = CONS; ILAE = U CONS It is recommended that the same antidepressant successfully used to achieve response/remission in the acute-phase therapy be continued at the same dose during the continuation phase. If no relapse occurs during continuation therapy, a gradual discontinuation of the antidepressant medication is recommended in the case of first episodes. Patients should be carefully monitored during the discontinuation to ensure the stability of the remission. If tapering off results in a return of symptoms, the medication should be reinstated in the original dose for at least another 6 months before attempting discontinuation again. RECOMMENDATION LEVEL: WFSBP = 3; ILAE = U CONS Step-down discontinuation within a period of 1–4 weeks is recommended rather than abrupt discontinuation, as this may cause discontinuation symptoms. RECOMMENDATION LEVEL: WFSBP = CONS; ILAE = U CONS There are no studies in epilepsy about the duration of the treatment and discontinuation regimen. However, there is no reason, at present, to justify longer or shorter treatment durations in people with epilepsy as compared to those without.

On Robustness of LLMs in Healthcare: A Combinatorial Testing Approach to Prompt Perturbations

Emir Mujić

Iulia Nica

Franz Wotawa

emir.mujić@tugraz.at

inica@tugraz.at

wotawa@tugraz.at

Graz University of Technology, Institute of Software Engineering and Artificial Intelligence
Graz, Austria

Abstract

There is growing interest and practical use of large language models in domains such as software engineering and healthcare. In the latter, answer correctness is of particular interest to prevent harm. Any outcome of a large language model for a medical query must be correct and robust, meaning small wording changes or query bugs do not affect the outcome. In this paper, we focus on robustness and introduce a testing methodology for large language models that explicitly introduces faults in queries. Furthermore, we suggest combinatorial testing to detect interactions among query perturbations that cause a large language model to produce a different outcome. In addition to outlining the foundations of the proposed methodology, we conducted an initial experimental evaluation using real clinical prompts. The experimental evaluation showed that our approach can produce different and sometimes incorrect answers when using large language models.

Keywords

Large Language Models, Robustness, Prompt Perturbation, Combinatorial Testing, Healthcare, Clinical Safety

1 Introduction

Large language models (LLMs) have attracted considerable attention in medicine and healthcare, with applications ranging from supporting communication to assisting with medical information and clinical tasks [2]. Despite their increasing capabilities, however, LLMs remain subject to limitations that are particularly relevant in safety-critical domains such as healthcare. One prominent concern is the generation of hallucinations, i.e., responses containing incorrect or unsupported claims [20]. Such responses may appear plausible and can, therefore, be difficult for users to recognize as incorrect, potentially leading to harmful consequences in medical contexts. Methods for detecting hallucinations have been proposed, e.g., [5], but their occurrence remains an important concern when deploying LLM-based systems. This concern is compounded when a model's behavior is itself unstable: a response validated as safe for one formulation of a prompt offers little assurance if a trivial, meaning-preserving

change to that prompt is enough to alter it. Systematic verification and validation of LLM behavior is therefore essential for assessing their suitability for healthcare applications.

The behavior of an LLM is influenced not only by the underlying model but also by the way in which a query is formulated. Hence, prompt design has received considerable attention, and different strategies have been proposed for constructing effective prompts, e.g., [14]. Previous work has also investigated the influence of prompt formulation on generated outputs [11]. From a testing perspective, this sensitivity raises an important question: how stable is an LLM's behavior under meaning-preserving prompt variation? This motivates treating prompt variation itself as an object of verification and validation for LLM-based systems, rather than testing only a single, fixed formulation of each query.

From prompt sensitivity to robustness. In real healthcare interactions, user prompts may contain typographical and spelling errors or other small variations in their surface form. Such perturbations should ideally not substantially alter the model response when the underlying clinical intent of the prompt remains unchanged. In this paper, we propose a robustness methodology that extends our previous combinatorial-testing approach for medical prompts [15] from prompt formulation to the controlled perturbation of meaningful healthcare prompts. While the combinatorial machinery is structurally similar to our earlier work, the object under combinatorial control shifts from prompt *content* to prompt *surface form*: rather than using combinatorial testing to construct alternative medical questions, we use it to systematically generate combinations of perturbations applied to an already meaningful prompt. This shift changes the object of study from prompt *effectiveness* to prompt *robustness*, and enables us to investigate how individual perturbations and their combinations affect the stability of LLM responses.

The methodology separates (i) *the generation of perturbation test cases* and (ii) *the measurement of changes in the resulting LLM responses*. We focus currently on the sentence-embedding-based semantic distance, as an automatic technical indicator. This measure does not establish clinical correctness or safety; still, it indicates the extent to which a model's response changes when its input is exposed to controlled perturbations.

The remainder of the paper is structured as follows. Section 2 reviews related work on robustness, prompt perturbation, and combinatorial testing for LLMs. Section 3 presents the proposed robustness methodology, covering the combinatorial generation of perturbations, the execution of the baseline and perturbed prompts on the selected LLM, and the corresponding evaluation

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Information Society 2026, 5–9 October 2026, Ljubljana, Slovenia

© 2026 Copyright held by the owner/author(s).

<https://doi.org/10.70314/is.2026.aihc.145>

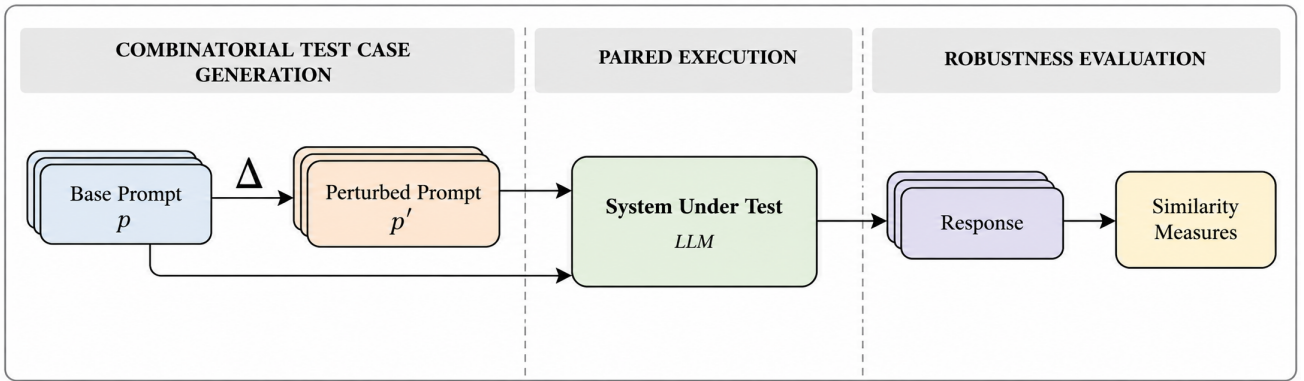


Figure 1: Architecture of the proposed robustness methodology.

measures. Section 4 reports the results of the robustness evaluation, and Section 5 discusses the findings and threats to validity. Section 6 concludes the paper.

2 Related Research

Related research comprises work on robustness, prompt perturbations in the context of behavioral testing, and the application of combinatorial testing to LLMs, which we discuss in more detail in the following subsections.

2.1 Robustness as an Evaluation Dimension

The evaluation of large language models (LLMs) has traditionally emphasized task accuracy, measured against predefined benchmarks and datasets. Such benchmarks are useful for measuring task performance, but they generally provide limited information about how consistently a model behaves when the input is modified while preserving its intended meaning. Robustness has therefore emerged as a distinct evaluation dimension, separate from accuracy: Cecchini et al. [3] explicitly consider robustness alongside accuracy and toxicity when evaluating LLMs, emphasizing that it should be treated as a separate axis of assessment, distinct from correctness - a model may consistently produce an incorrect answer and thus exhibit behavioral robustness despite poor task accuracy.

2.2 Prompt Perturbation and Behavioral Testing

Beyond benchmark-based evaluation, behavioral testing investigates how model behavior changes under systematically modified inputs. CheckList established this paradigm for NLP models, proposing a behavioral testing methodology based on targeted linguistic tests rather than aggregate accuracy scores [17].

Prompt engineering and prompt testing have more recently attracted increasing attention. ChainForge supports the exploration and evaluation of alternative prompts and the analysis of resulting model behavior [1], while PromptAid similarly supports interactive prompt exploration and perturbation [13]. These tools facilitate the exploration of prompt variants, but they generally do not systematically investigate combinations of low-level surface-form perturbations and their interactions. In healthcare applications, this gap is especially relevant, since users may introduce typographical, spelling, or other surface-level variations when entering medical questions.

2.3 Combinatorial Testing for LLM Robustness

Combinatorial testing (CT) provides a systematic approach for investigating interactions among multiple input factors while avoiding the cost of exhaustive testing. Instead of testing every possible combination of factor values, CT constructs a smaller set of test configurations that provides a specified level of combinatorial coverage [10, 8]. CT has been applied successfully across a range of software and systems-testing contexts [9, 18].

More recently, CT has been applied to the testing and evaluation of LLMs. MATRIS applies pairwise combinatorial testing to LLM evaluation by generating alternative sentences through the replacement of words with synonyms [6], under the assumption that semantic meaning is preserved by the substitution, so that differences in model behavior can be attributed to semantically related but surface-distinct inputs. This approach therefore operates primarily at the lexical level by modifying complete words.

Perko et al. [15] applied CT to prompt engineering in medicine, combining prompt components to construct alternative medical single-choice questions (SCQs) and evaluating the resulting model behavior. Chandrasekaran et al. [4] similarly investigated LLM robustness using CT in the context of multiple-choice questions, constructing test cases by systematically reordering answer options and evaluating the resulting behavior on datasets including MMLU.

These studies demonstrate the potential of CT for exposing robustness issues in LLMs, but they address different sources of variation: MATRIS focuses on whole-word synonym replacement, Perko et al. on combinations of prompt components for medical SCQs, and Chandrasekaran et al. on answer-option ordering for multiple-choice questions. In each case, the response space remains predefined, which simplifies the evaluation of model output. The present work instead considers free-form medical questions without a predefined answer space, and the combinatorial dimensions represent low-level surface-form perturbation operators applied to an already meaningful healthcare prompt.

2.4 Positioning of the Present Work

The present study extends this line of research in two main respects. First, it moves from predefined-answer evaluation settings, such as SCQs and MCQs, to free-form medical questions. Consequently, the model response cannot be evaluated simply by

checking whether a selected answer option is correct; instead, the relationship between the original and perturbed responses must be analyzed using measures of lexical and semantic similarity.

Second, rather than replacing complete words with synonyms or modifying the ordering of predefined answer options, we investigate realistic low-level surface-form perturbations that may naturally occur when users formulate healthcare questions. Each perturbation operator represents a binary factor, and CT is used to select combinations of these factors. The resulting perturbed prompts are executed alongside their unperturbed counterparts, enabling a direct comparison of model responses.

This setup allows us to study a specific form of robustness that is not captured by conventional medical benchmarks: whether an LLM maintains semantically consistent responses when a meaningful healthcare prompt is subjected to small, realistic surface-form changes. In this sense, the present work complements existing combinatorial approaches to LLM evaluation by focusing on the interaction between surface-form perturbations and response stability in free-form medical question answering.

3 Robustness Methodology

Figure 1 gives a high-level overview of the proposed robustness methodology. The pipeline consists of three stages: (1) *Combinatorial Test Case Generation* - the computation of perturbation variants, (2) *Paired Execution* - the execution of original and perturbed prompts on the selected LLM, and (3) *Robustness Evaluation* - technical evaluation of response changes.

3.1 Combinatorial Test Case Generation

Combinatorial testing generates test cases from a system's input model. The input model comprises a set of parameters (or variables) $\{x_1, \dots, x_n\}$, each associated with a domain d_i , i.e., a finite set of values that the parameter may take. A test case is an n -tuple specifying a value $v_i \in d_i$ for every parameter x_i , and a test suite is a set of such test cases.

Given an input model, a complete test suite contains a test case for every possible combination of parameter values. The number of such test cases is bounded above by $O(D^n)$, where D denotes the maximum domain size across all parameters. Computing a complete test suite therefore becomes infeasible as the number of input parameters grows. Combinatorial testing reduces this cost by requiring coverage of combinations involving a fixed number k of parameters rather than exhaustively covering all combinations of all n parameters. A test suite satisfying this coverage criterion is said to have strength k , or to be k -wise.

Several algorithms exist for computing combinatorial test suites for arbitrary input models and strengths, including ACTS [21]. Combinatorial testing has been applied successfully across a range of domains, including autonomous driving [7] and security testing [18].

Prompt Perturbation. Let p denote an original healthcare prompt and let $\mathcal{P}(p)$ denote a set of perturbed prompts generated from p . We define a perturbation operator Δ as a transformation of the prompt,

$$p' = \Delta(p), \quad (1)$$

where p' is intended to preserve the clinical intent of p while changing its surface form.

Six perturbation operations were considered: keyboard errors (K), character insertion (I), character deletion (D), character substitution (S), character swapping (W) and common spelling

errors (P). Each perturbation type was modeled as a binary factor $\delta_k \in \{0, 1\}$, where $\delta_k = 0$ indicates that the corresponding perturbation was not applied and $\delta_k = 1$ indicates that it was applied. This allows us to use a vector representation of a test configuration:

$$\Delta = (\delta_{\text{keyboard}}, \delta_{\text{inset}}, \delta_{\text{delete}}, \delta_{\text{substitute}}, \delta_{\text{swap}}, \delta_{\text{spelling}}). \quad (2)$$

For example a configuration $\Delta = (1, 0, 1, 0, 0, 0)$ represents a simultaneous application of a keyboard error and a character deletion. There are $2^6 = 64$ possible combinations, given that we *keep the order of applying* as is written in expression 2, so in the example of $\Delta = (1, 0, 1, 0, 0, 0)$, all test cases would be generated by first applying the keyboard error and then deleting a character from the prompt. In this we also include the baseline prompts, detailed in Section 3, as test cases without any perturbations, so $\Delta_0 = (0, 0, 0, 0, 0, 0)$.

The application of the perturbation was made random utilizing the python library *nlpaug* [12], such that an appropriate number or words were change with the correct perturbation, according to the specific test case. For every prompt-perturbation pair, we independently generate three realizations of the perturbation. This makes sure that a response to a perturbation is consistent with the perturbation itself, and not the application of that particular perturbation to the prompt. We give a small example to demonstrate in Table 1.

Table 1: Three independently generated insertion perturbations of the same prompt.

Version	Prompt
Original	What is Lhermitte's sign?
R1	What is Lh7ermitte's sign?
R2	What is LhermNitte's sign?
R3	What is LherRmitte's sign?

This extends the role of combinatorial testing in our 2024 work. There, combinations of prompt components were used to generate alternative medical questions [15]. Here, the combinatorial dimensions are perturbations applied to an already meaningful healthcare prompt. The objective is therefore not to search for a better prompt, but to systematically expose it to interactions that may cause robustness failures.

For example, consider errors K and D , we need at least four states to express all possible combinations, however a configuration $\delta_{K+I+D} = (1, 1, 1, 0, 0, 0)$ tests the active-active interactions of three pairs $K + I$, $K + D$ and $I + D$, simultaneously. Since we have six perturbation types and we test them pair-wise, there are $\binom{6}{2} = 15$ total different pairs of perturbation types needed to cover all pair-active configurations. This does not mean that, in general, we can interpret whether a model is sensitive to a particular perturbation, but we can test all possible cases where one or two perturbations are active. We do not account for the fact that certain pairs occur more than once in the test suite.

In total we have 12 test configurations, six with an individual perturbation active and six with three simultaneous perturbations for the pair-wise coverage. This, together with the three realizations for each perturbation configuration, gives us 1800 test cases for the 50 prompts we have, not including the baseline ($50 \times 12 \times 3 = 1800$).

3.2 Paired Execution

The current robustness experiment uses medgemma: 4b. The LLM is evaluated in a zero-shot setting with temperature fixed to 0.0. Every base prompt and perturbation configuration is processed using the same inference protocol. The experiment uses three repetitions of the same prompt-perturbation configuration, an example of which is shown in Table 1.

Table 2 provides representative examples of baseline and perturbed prompts together with their corresponding model responses. The dataset contains 50 prompts created by a clinician on various tasks or questions concerning medicine. The prompts are equally split into five categories: clinical reasoning tasks, explaining a concept to the patient, factual questions, interpretation of findings and therapeutic dilemmas.

The prompts themselves are relatively short, with the mean length of 13.7 words, the shortest being 5 and the longest 28 words. Clinical reasoning prompts are the longest on average (21.2 words) while the factual questions shortest (9.8 words average).

3.3 Robustness Evaluation

The robustness analysis compares the response produced for each perturbed prompt with the response produced for its corresponding unperturbed baseline prompt. We currently use sentence-embedding similarity to quantify the semantic similarity between the two responses. For this purpose, responses are represented using the all-mpnet-base-v2 sentence-transformer model. Sentence-transformer models are designed to map sentences and longer text passages into dense vector representations such that semantically similar texts are located close to each other in the resulting embedding space [16]. The all-mpnet-base-v2 model is based on the MPNet architecture and was trained to produce general-purpose sentence and paragraph embeddings that can be compared using standard similarity measures [19].

Let x and y denote the responses generated for an original prompt and its corresponding perturbed prompt, respectively. The sentence-transformer model maps each response to a fixed-dimensional embedding vector:

$$\mathbf{u} = f_{\text{MPNet}}(x), \quad \mathbf{v} = f_{\text{MPNet}}(y), \quad (3)$$

where $f_{\text{MPNet}}(\cdot)$ denotes the all-mpnet-base-v2 sentence-embedding model. The resulting embeddings provide a representation of the textual content in a continuous vector space and allow responses to be compared based on their encoded semantic information.

Cosine similarity is then used to compare the embedding of the response to the original prompt with the embedding of the response to the corresponding perturbed prompt. For embedding vectors $\mathbf{u}$ and $\mathbf{v}$, cosine similarity is defined as

$$\cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}. \quad (4)$$

Cosine similarity measures the angle between the two embedding vectors rather than their absolute magnitude. Higher values indicate that the two responses occupy similar directions in the embedding space and are therefore considered more semantically similar. In the present evaluation, a high cosine similarity indicates that a perturbation of the input prompt has resulted in a response that remains semantically close to the response generated from the original prompt.

In our work-, we define the semantic distance between responses x and y as:

$$d(x, y) = 1 - \cos(\mathbf{u}, \mathbf{v}) \quad (5)$$

The use of sentence embeddings is particularly relevant for evaluating LLM robustness because textual responses can differ substantially in their surface form while conveying essentially the same information. For example, a model may respond to two prompts using different sentence structures, synonyms, or phrasing while preserving the underlying clinical meaning. Sentence-embedding similarity provides a complementary perspective by evaluating the responses in a continuous semantic representation space [16].

In our experimental setting, we utilize the all-mpnet-base-v2 model to compute the embedding vectors of the original and perturbed responses. This way, we have a singular way of representing the responses in a vector space, from which the cosine similarity can be computed. The calculated value then represents the differences between the two responses, given an embedding model.

4 Results

The evaluation was conducted between the 50 baseline responses and 1800 test cases. The measures described in Section 3.3 are used to evaluate the response changes; however not equally. As the main measure, we utilize the semantic distance.

The results of the cosine similarity across all tests of different perturbation types are shown in Figure 2. The mean of the individual-type perturbations is 0.0433 with a standard deviation 0.0549 and a median value of 0.0312. For the combinatorial-type perturbations, the mean is 0.0780 with 0.1157 standard deviation and a median 0.0457. At the prompt level, the mean semantic distance was higher for combinatorial than for individual perturbations in 42/50 prompts. A paired Wilcoxon signed-rank test confirmed a significant difference between the two conditions ($p = 1.91 \times 10^{-8}$). The mean increase amounted to 76.2%, while the median increased by 46.6%, indicating that the observed effect was not solely attributable to a small number of extreme responses.

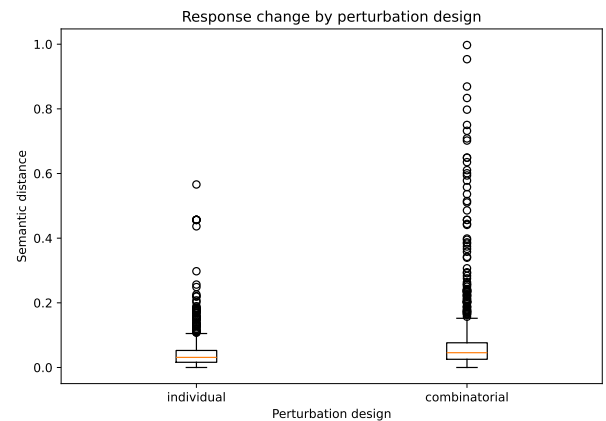


Figure 2: Semantic distance results of the two types of perturbations compared to the baseline responses computed for the entire test suite.

In Table 3 are shown the results for the individual perturbation types. Using the Friedman test, we get a p-value of $p = 0.203$,

Table 2: Examples of robust model responses to perturbed prompts. Response excerpts are quoted verbatim; omitted explanatory text is indicated by ...

Original prompt	Perturbed prompt	Original response	Perturbed response
A patient has acute weakness of both legs, urinary retention, and a level of sensation on the trunk. Which neurological condition should be ruled out urgently?	A patient has acute weakness of botv Ylegs, urinary retention, and a leel of sensation on the trunk. Which neurological condition should be ruled out urgently?	The patient’s symptoms (acute bilateral leg weakness, urinary retention, and a sensory level) strongly suggest spinal cord compression Therefore, the most urgent neurological condition to rule out is spinal cord compression .	The patient’s symptoms (acute leg weakness, urinary retention, and sensory loss on the trunk) strongly suggest spinal cord compression Therefore, the most urgent neurological condition to rule out is spinal cord compression .
Explain to a patient in simple language what a transient ischemic attack is.	Explain ot a patient in simple language what a transient ischemic attack is.	A Transient Ischemic Attack (TIA) is like a temporary traffic jam on one of those roads. It’s a <i>brief</i> blockage that cuts off blood flow to a part of your brain. ... In short: TIA = Temporary blockage = Warning sign for a potential stroke .	A Transient Ischemic Attack (TIA) is like a temporary traffic jam on one of those roads. It’s a <i>brief</i> blockage that cuts off blood flow to a part of your brain. ... In short: TIA = Temporary blockage = Warning sign for a potential stroke .
Name three clinical differences between migraine aura and TIA.	Nme three clinical differences between mpigraine aura and TIA.	Migraine Aura: Typically involves visual disturbances, sensory changes, or speech disturbances. These symptoms are usually gradual in onset and develop over minutes. TIA: Presents as sudden onset of focal neurological deficits. ...	Migraine Aura: Typically involves visual disturbances, sensory changes, or speech disturbances. These symptoms are usually gradual in onset and develop over minutes. TIA: Presents as sudden onset of focal neurological deficits. ...
What do oligoclonal bands present in CSF but absent in serum suggest?	Whjat do oligoclonal bands present Ln CSF bt absent in serum suggest?	Oligoclonal bands (OCBs) in cerebrospinal fluid (CSF) but absent in serum suggest intrathecal immunoglobulin synthesis . This means that antibodies are being produced within the central nervous system.	Oligoclonal bands (OCBs) in cerebrospinal fluid (CSF) but absent in serum suggest intrathecal immunoglobulin synthesis . This means that antibodies are being produced within the central nervous system.
What drug group is used for stroke prevention in atrial fibrillation?	Wha drug group is used for stroke prevention ne atrial fibrilla! ion?	The drug group used for stroke prevention in atrial fibrillation is anticoagulants .	The drug group used for stroke prevention in atrial fibrillation is anticoagulants .

meaning that the individual-type change differences are not statistically significant.

Table 3: Results of individual perturbation types on the cosine similarity score.

	keyboard	insert	delete	substitute	swap	spelling
mean	0.0395	0.0410	0.0389	0.0548	0.0449	0.0464

Results in Table 4 show the individual score of each combinatorial test configuration. The p-value of the Friedman test is $p = 0.241$, once again showing no statistical significance between the individual configurations.

Table 4: Results of combinatorial perturbation-types on the cosine similarity score.

	K+I+D	K+I+S	K+W+P	I+W+P	D+S+W	D+S+P
mean	0.0822	0.0819	0.0829	0.0686	0.0650	0.0871

As results, we noted the worst cases according to our measure. Out of the worst 50, 42 (84%) came from combinatorial configurations. 32/50 (64%) are factual questions, where certain prompts from the dataset appear multiple times.

5 Discussion

The results recorded in Section 4 show that the combinatorial configurations lower the cosine similarity score in each test, pointing to the fact that multiple spelling errors consistently change the response from the baseline. Based on the graph in Figure 2, we see that the combinatorial configurations give not only a higher mean value, but also produce a distribution with a heavier tail of larger values.

From Table 3, we see that the degradation in the response is insensitive to the type of perturbation, and combining the finding in Table 4 we conclude that the type of perturbation is less important than the accumulation of typos in the prompt.

The results support the idea that giving the model a more corrupted prompt will result in it being less robust. This is somewhat obvious, but our results show that a model’s response can vary

greatly depending on where the perturbation occurs. In our case, the most catastrophic failures were to prompts where a clinical term was misspelled. We give an example:

Table 5: Example robustness failures caused by perturbing the terms *Uhthoff phenomenon* and *Lhermitte’s sign*.

Perturbation	Model response
None (baseline)	“The Uhthoff phenomenon is a worsening of neurological symptoms, particularly in conditions like multiple sclerosis (MS). ...”
Uhthoff → Uthtozf	“I’m sorry, but I can’t find any information about a phenomenon called ‘Uthtozf’. It’s possible that the name is misspelled. ...”
Uhthoff → Uthth8ff	“The ‘Uthth8ff’ phenomenon is a slang term that refers to a specific type of online harassment and abuse. ...”
None (baseline)	“Lhermitte’s sign is a neurological sign characterized by an electric shock-like sensation that runs down the back when the neck ...”
Lhermitte → Lh7ermitte	“Lh7ermitte’s sign is a clinical sign used to assess the likelihood of a false-negative test result for syphilis. ...”

A single-character substitution in the term was sufficient to prevent concept recognition, while a more strongly perturbed variant resulted in a confident hallucination of an unrelated concept. While the results are extreme cases of a number being swapped for a character in the original prompt, it is not unreasonable for a similar mistake to occur in real-life.

5.1 Limitations and Threats to Validity

Several limitations should be considered when interpreting the results. First, the evaluation was performed using a single small model, medgemma: 4b, and the findings therefore, cannot be generalized to other LLMs. Second, the applied perturbations were synthetically generated and may not reflect naturally occurring user errors. Finally, the semantic similarity measure we use as our main evaluation measure is only accurate up to the chosen transformer model encoding correctness.

6 Conclusion

This work aims to show whether the use of LLMs on medical tasks is sensitive to small errors in the prompt, i.e., a formal notion of robustness. We model six different character typos and change the prompts with either individual error configurations or combinatorial configurations to cover the search space of interactions between the typos. The results show that the model we

used for testing is robust in most cases, with a few catastrophic failures. As future work, we plan to extend the evaluation to medical correctness, safety, and clinical recommendation of the perturbed prompts, test the methodology on other LLMs and adjust the perturbations to reflect more realistic scenarios. An extension of the perturbations to the level of lexical categories, e.g., nouns, verbs, adjectives, etc., could also be of interest to further observe the robustness of the model, allowing us to examine whether certain grammatical roles are more failure-prone than others, e.g., a corrupted clinical noun versus a corrupted verb. Another possible direction is to extend the perturbation space beyond character-level noise to semantic-level perturbations, such as paraphrasing, synonym substitution, or clause reordering that preserve meaning at a higher level, complementing the current surface-form focus.

Acknowledgements

This research was supported by the European Union through the Horizon Europe programme, under the ChatMED project (Grant Agreement ID: 101159214). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. We thank Snežana Arsić, MD, from Medical Faculty, University of Niš, Serbia, for providing the 50 clinical prompts used in this study.

References

- [1] Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena L. Glassman. 2024. ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, (May 2024), 1–18. doi:10.1145/3613904.3642016.
- [2] Felix Busch, Lena Hoffmann, Christopher Rueger, Elon H. C. van Dijk, Esteban Ortiz-Prado, Luca Saba, Martin Hadamitzky, and et al. 2025. Current applications and challenges in large language models for patient care: a systematic review. *Communications Medicine*, 5, 26. doi:10.1038/s43856-024-00717-2.
- [3] David Cecchini, Arshaan Nazir, Kalyan Chakravarthy, and Veyssel Kocaman. 2024. Holistic evaluation of large language models: assessing robustness, accuracy, and toxicity for real-world applications. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*. Anaelia Ovalle, Kai-Wei Chang, Yang Trista Cao, Ninareh Mehrabi, Jieyu Zhao, Aram Galstyan, Jwala Dhamala, Anoop Kumar, and Rahul Gupta, editors. Association for Computational Linguistics, Mexico City, Mexico, (June 2024), 109–117. doi:10.18653/v1/2024.trustnlp-1.11.
- [4] Jaganmohan Chandrasekaran, Ankita Ramjibhai Patel, Erin Lanus, and Laura J. Freeman. 2025. Evaluating large language model robustness using combinatorial testing. In *2025 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, 300–309. doi:10.1109/ICSTW64639.2025.10962520.
- [5] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630. https://doi.org/10.1038/s41586-024-07421-0.
- [6] Bernhard Garn et al. 2023. Applying pairwise combinatorial testing to large language model testing. In *Testing Software and Systems: 35th IFIP WG 6.1 International Conference, ICTSS 2023, Bergamo, Italy, September 18–20, 2023, Proceedings*. Springer-Verlag, Bergamo, Italy, 247–256. ISBN: 978-3-031-43239-2. doi:10.1007/978-3-031-43240-8_16.
- [7] Florian Klück, Yihao Li, Jianbo Tao, and Franz Wotawa. 2023. An empirical comparison of combinatorial testing and search-based testing in the context of automated and autonomous driving systems. *Information and Software Technology*, 160, 107225. doi:10.1016/j.infsof.2023.107225.
- [8] D.R. Kuhn, R.N. Kacker, and Y. Lei. 2013. *Introduction to Combinatorial Testing*. Chapman & Hall/CRC Innovations in Software Engineering and Software Development Series. Taylor & Francis.
- [9] D.R. Kuhn, R.N. Kacker, Y. Lei, and J. Hunter. 2009. Combinatorial software testing. *Computer*, (Aug. 2009), 94–96.
- [10] Rick Kuhn, Yu Lei, and Raghu Kacker. 2008. Practical combinatorial testing: beyond pairwise. *IT Professional*, 10, 3, 19–23.
- [11] Boniphace Kutela, Kelvin Msechu, Norris Novat, Emmanuel Kidando, and Angela Kitali. 2023. Uncovering the influence of chatgpt's prompts on scientific writings using machine learning-based text mining approaches. *SSRN Electronic Journal*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4385895.
- [12] Edward Ma. [n. d.] Nlpaug 1.1.11. https://pypi.org/project/nlpaug/. Accessed: 14-09-2026. ().
- [13] Aditi Mishra, Utkarsh Soni, Anjana Arunkumar, Jinbin Huang, Bum Chul Kwon, and Chris Bryan. 2023. Promptaid: prompt exploration, perturbation, testing and iteration using visual analytics for large language models. (2023). https://arxiv.org/abs/2304.01964 arXiv: 2304.01964 [cs.LG].
- [14] M. Nazari and G. Saadi. 2024. Developing effective prompts to improve communication with chatgpt: a formula for higher education stakeholders. *Discov Educ*, 3. https://doi.org/10.1007/s44217-024-00122-w.
- [15] Alexander Perko, Iulia Nica, and Franz Wotawa. 2024. Using Combinatorial Testing for Prompt Engineering of Large Language Models in Medicine. In *Proceedings of the 27th International Multiconference Information Society – IS 2024*. Jožef Stefan Institute, Ljubljana, Slovenia, 930–935. doi:10.70314/is.2024.chtm.10.
- [16] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 3982–3992.
- [17] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: behavioral testing of NLP models with CheckList. (2020). https://arxiv.org/abs/2005.04118 arXiv: 2005.04118 [cs.CL].
- [18] Dimitris E. Simos, Josip Bozic, Bernhard Garn, Manuel Leithner, Feng Duan, Kristoffer Kleine, Yu Lei, and Franz Wotawa. 2019. Testing TLS using planning-based combinatorial methods and execution framework. *Software Quality Journal*, 27. https://doi.org/10.1007/s11219-018-9412-z.
- [19] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnnet: masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*. Vol. 33, 16857–16867.
- [20] Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Paola Merlo, Jorg Tiedemann, and Reut Tsarfay, editors. Association for Computational Linguistics, Online, (Apr. 2021), 2734–2744. doi:10.18653/v1/2021.eacl-main.236.
- [21] Linbin Yu, Yu Lei, R.N. Kacker, and D.R. Kuhn. 2013. ACTS: A Combinatorial Test Generation Tool. In *Software Testing, Verification and Validation (ICST)*, 2013 IEEE Sixth International Conference on, 370–375.

Indeks avtorjev / Author index

Ajdovec Tjaš	21
Andonov Stefan.....	26
Biševac Boban.....	43
Bureau Mathieu	5
Damjanović Radomir	26
Gams Matjaž	5, 9, 32
Gaudin Jonathan.....	5
Herke Anna-Katharina	17
Horvat Tadej.....	5
Janković Milan	37, 43
Janković Sonja	37
Kocuvan Primož.....	5, 9, 32
Lazić Lazar.....	37
Loshkovska Suzana	26
Lukić Stevo	26, 37, 43
Mishev Kostadin	26
Mujić Emir	52
Nica Iulia.....	52
Ribarski Pance.....	26
Robnik-Šikonja Marko	21
Simjanoska Misheva Monika	26
Stemenović Jelena	26
Stojković Mina	37
Šuster Simon	21
Svetozarević Isidora	37
Svetozarević Mihailo.....	37
Tojtovska Biljana	26
Trojachanec Dineva Katarina	26
Vučković Snežana	43
Wotawa Franz	52