Zbornik 28. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2025

Zvezek K

Proceedings of the 28th International Multiconference

INFORMATION SOCIETY – IS 2025

Volume K

Uporaba AI v zdravstvu
AI in Healthcare

Urednika / Editors

Matjaž Gams, Tjaša Kolar

http://is.ijs.si

8 October 2025 / 8 October 2025
Ljubljana, Slovenia

Urednika:

Matjaž Gams
Odsek za inteligentne sisteme, Institut »Jožef Stefan«, Ljubljana, Slovenija

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

Založnik: Institut »Jožef Stefan«, Ljubljana
Priprava zbornika: Mitja Lasič, Vesna Lasič, Mitja Lenič
Oblikovanje naslovnice: Vesna Lasič

Dostop do e-publikacije:
http://library.ijs.si/Stacks/Proceedings/InformationSociety

Ljubljana, oktober 2025

Informacijska družba
ISSN 2630-371X

PREDGOVOR MULTIKONFERENCI INFORMACIJSKA DRUŽBA 2025

28. mednarodna multikonferenca *Informacijska družba* se odvija v času izjemne rasti umetne inteligence, njenih aplikacij in vplivov na človeštvo. Vsako leto 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, kjer 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 **Niko Schlamberger**, priznanje za raziskovalni dosežek leta pa **Tome Eftimov**. »Informacijsko limono« za najmanj primerno informacijsko tematiko je prejela odsotnost obveznega pouka računalništva v osnovnih šolah. »Informacijsko jagodo« za najboljši sistem ali storitev v letih 2024/2025 pa so prejeli Marko Robnik Šikonja, Damir Vreš in Simon Krek s skupino za slovenski veliki jezikovni model GAMS. 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 prihodnje dosežke, ki jih bo soustvarjala ta konferenca.

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

FOREWORD TO THE MULTICONFERENCE INFORMATION SOCIETY 2025

The 28th International Multiconference on the Information Society takes place at a time of remarkable growth in artificial intelligence, its applications, and its impact on humanity. Each year we enter a new era in which generative AI and other innovative approaches shape the path toward superintelligence and singularity — phenomena that will shape the future of human civilization. The conference is both a traditional scientific forum and an academically open incubator for new, bold ideas and perspectives.

In addition to artificial intelligence, this year's conference addresses other pressing issues of our time: environmental preservation, demographic challenges, healthcare, and the transformation of social structures. The rapid development of AI offers potential solutions to many of today's challenges and highlights the importance of collaboration among researchers, experts, and policymakers in designing sustainable strategies. We are acutely aware that we live in an era of profound change, where innovative approaches and deep knowledge are essential to creating an information society that is safe, inclusive, and sustainable.

This year's multiconference brings together twelve thematically diverse meetings reflecting the breadth and depth of the information sciences: from artificial intelligence in healthcare, demographic and family studies, and the digital transformation of nursing and digital inclusion, to research in cognitive science, healthy longevity, and education in the information society. Additional conferences include 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 scientific papers, round tables and workshops will provide opportunities for in-depth exchanges of views, making an important contribution to shaping the future information society. *Legends of Computing and Informatics* serves as a national »Hall of Fame« honoring outstanding individuals in the field. We will continue to promote research and development, excellence, and collaboration. Extended papers will be published in the journal *Informatica*, supported by a long-standing tradition and in cooperation with academic institutions and professional associations such as ACM Slovenia, SLAIS, the Slovenian Society Informatika, and the Slovenian Academy of Engineering.

Each year we recognize the most distinguished achievements. In 2025, the Michie-Turing Award for lifetime contribution to the development and promotion of the information society was awarded to **Niko Schlamberger**, while the Award for Research Achievement of the Year went to **Tome Eftimov**. The »Information Lemon« for the least appropriate information-related topic was awarded to the absence of compulsory computer science education in primary schools. The »Information Strawberry« for the best system or service in 2024/2025 was awarded to Marko Robnik Šikonja, Damir Vreš and Simon Krek together with their team, for developing the Slovenian large language model GAMS. We extend our warmest congratulations to all awardees.

Our vision remains clear: to identify, seize, and shape the opportunities offered by digital transformation, and to create an information society that benefits all its members. We sincerely thank all participants for their contributions and look forward to jointly shaping the future achievements that this conference will help bring about.

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 Ciglarič, 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 Mladenič
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
Aleš Ude
Bojan Blažica
Matjaž Kljun
Robert Blatnik
Erik Dovgan
Špela Stres
Anton Gradišek

KAZALO / TABLE OF CONTENTS

Zbornik 28. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2025

Zvezek K

Proceedings of the 28th International Multiconference

INFORMATION SOCIETY – IS 2025

Volume K

Uporaba AI v zdravstvu
AI in Healthcare

Urednika / Editors

Matjaž Gams, Žiga Kolar

http://is.ijs.si

8 October 2025 / 8 October 2025
Ljubljana, Slovenia

PREDGOVOR

Umetna inteligenca, zlasti generativna umetna inteligenca, kot je ChatGPT, je spremenila številne panoge. V zdravstvu je njen vpliv še posebej velik, saj ne vpliva le na informacije, ampak tudi na človeška življenja. Z izboljšanjem izidov zdravljenja pacientov, poenostavitvijo delovnih tokov in podporo kliničnemu odločanju ima umetna inteligenca potencial, da preoblikuje prihodnost medicine.

Vloga umetne inteligence presega pomoč strokovnjakom; neposredno izboljšuje oskrbo pacientov. Virtualne konzultacije, preverjanje simptomov in izobraževanje pacientov širijo dostop do zdravstvenega varstva za tiste, ki se soočajo z geografskimi ali časovnimi ovirami. Hkrati avtomatizacija rutinskih nalog zmanjšuje administrativno breme zdravnikov, kar jim omogoča, da se osredotočijo na to, kar je najpomembnejše – oskrbo pacientov. Ta premik je ključen tudi pri reševanju izčrpanosti zdravnikov, ki je vse bolj pereča tema v sodobnem zdravstvu.

Vendar pa je treba obljube umetne inteligence uravnotežiti z odgovornostjo. Etični in varnostni izzivi ostajajo: zaščita zasebnosti pacientov, zmanjšanje pristranskosti algoritmov in zagotavljanje točnosti medicinskih nasvetov. Umetna inteligenca bi morala dopolnjevati in ne nadomeščati človeško strokovno znanje, zlasti pri kritičnih odločitvah. Pregledni, odgovorni in varnostno usmerjeni sistemi so bistveni za gradnjo trajnega zaupanja.

V prihodnosti bodo ChatGPT in sorodne tehnologije morda imele osrednjo vlogo v personalizirani medicini, zgodnjem odkrivanju bolezni, odkrivanju zdravil in globalnih pobudah na področju zdravstvene iniciative. Z analizo ogromnih količin podatkov – od genetike do trendov na ravni prebivalstva – bi umetna inteligenca lahko odprla nove možnosti za natančno zdravljenje in preprečevanje bolezni.

Ta konferenca temelji na prispevkih projekta ChatMED, zlasti na osebni medicinski platformi HomeDOCtor, ki temelji na LLM in se v Sloveniji redno uporablja že devet mesecev, v Makedoniji in Srbiji pa kot prototip. Po ocenah bi taki sistemi lahko prinesli 100 milijonov evrov koristi, če bi se intenzivno uporabljali na nacionalni ravni. Projekt ponuja edinstveno priložnost za preučitev najnovejših raziskav, nastajajočih aplikacij in etičnih vidikov ChatGPT v zdravstvu. Skupaj bomo razmislili o trenutnih zmogljivostih, obravnavali ključne izzive in raziskali prihodnji potencial umetne inteligence pri ustvarjanju varnejšega in učinkovitejšega zdravstvenega sistema.

Franz Wotawa
Monika Smiljanovska
Stevo Lukić
Matjaž Gams

FOREWORD

Artificial Intelligence, and particularly conversational AI such as ChatGPT, has transformed many industries. In healthcare, its impact is especially profound, as it touches not only information but human lives. By improving patient outcomes, streamlining workflows, and supporting clinical decision-making, AI has the potential to reshape the future of medicine.

AI's role goes beyond assisting professionals; it directly enhances patient care. Virtual consultations, symptom checks, and patient education expand access to healthcare for those facing geographic or time barriers. At the same time, automation of routine tasks reduces clinicians' administrative burden, enabling them to focus on what matters most—caring for patients. This shift is also crucial in addressing physician burnout, an increasingly urgent issue in modern healthcare.

Yet the promise of AI must be balanced with responsibility. Ethical and safety challenges remain: protecting patient privacy, minimizing algorithmic bias, and ensuring the accuracy of medical advice. AI should augment and not replace human expertise, particularly in critical decisions. Transparent, accountable, and safety-first systems are essential to building lasting trust.

Looking ahead, ChatGPT and related technologies may play a central role in personalized medicine, early disease detection, drug discovery, and global health initiatives. By analyzing vast amounts of data—from genetics to population-level trends—AI could unlock new possibilities for precision care and prevention.

This conference builds on contributions from the ChatMED project, especially the LLM-based personal medical platform HomeDOCtor, which has been in regular use in Slovenia for the past nine months, and as a prototype in Macedonia and Serbia. An estimate suggests that such systems could provide 100 million Euros in benefits if they are nationally intensively used. The project provides a unique opportunity to examine cutting-edge research, emerging applications, and ethical considerations of ChatGPT in healthcare. Together, we will reflect on current capabilities, address key challenges, and explore the future potential of AI in creating a safer and more effective healthcare system.

Franz Wotawa
Monika Smiljanovska
Stevo Lukić
Matjaž Gams

PROGRAMSKI ODBOR / PROGRAMME COMMITTEE

Matjaž Gams

Monika Simjanoska Misheva

Stevo Lukić

Franz Wotawa

Beyond Accuracy: A Multidimensional Evaluation Framework for Medical LLM Applications (M-LEAF)

Rok Smodiš†
rok.smodis@gmail.com
Pedagoška fakulteta, Kognitivna znanost
Ljubljana, Slovenia

Ivana Karasmanakis
karasmanakisivana@gmail.com
Univerza v Ljubljani, Medicinska fakulteta
Ljubljana, Slovenia

Filip Ivanišević
filipivanisevic79@gmail.com
Univerza v Ljubljani, Medicinska fakulteta
Ljubljana, Slovenia

Matjaž Gams
matjaz.gams@ijs.si
Department of Intelligent Systems
Ljubljana, Slovenia

Abstract

Large language models are being increasingly used in healthcare to support both patients and clinicians. Current evaluations mostly measure diagnostic accuracy and often neglect other qualities that are also essential for their safe deployment, such as interaction quality, safety and transparency. To address this gap we introduce M-LEAF, a multidimensional framework that organizes these requirements into eight pillars and provides clear metrics and protocols for each. The framework uses a unified 0 to 5 scoring scale and includes safeguards to ensure that critical failures cannot be hidden. We applied M-LEAF in two pilot studies that compared GPT-4o with the HomeDOCtor system. In both of the studies, both systems achieved high scores, which demonstrate the feasibility and value of a structured multidimensional approach.

Keywords

Artificial Intelligence, Large Language Models, Clinical Decision Support, Healthcare Evaluation Framework

1 Introduction

Healthcare systems worldwide face persistent clinician shortages, increasing patient loads, and rising demand for timely, safe medical guidance [1]. Large language models (LLMs) have emerged as a promising tool to address these challenges, both in patient-facing contexts (e.g., symptom checkers, triage chatbots) and clinician-facing workflows (e.g., decision support, summarisation, documentation) [2, 3, 4]. Recent studies demonstrate that LLMs can achieve impressive scores on medical question answering benchmarks [5, 6, 7, 8]. However, these evaluations largely emphasise diagnostic accuracy on static, single-turn items. As Bedi and colleagues [4] note, fewer than one-fifth of published evaluations explicitly considers broader dimensions of the diagnostic process, such as fairness, robustness and factuality.

Information Society 2025, Ljubljana, Slovenia

https://doi.org/10.70314/is.2025.gptzdravje.2

2 Related Work

2.1 Benchmarks and Evaluation Datasets

A number of benchmark datasets have been used to test LLMs in healthcare. PubMedQA provides thousands of annotated biomedical Q&A pairs for knowledge testing [9]. MedQA draws directly from the United States Medical Licensing Examination (USMLE), offering multiple-choice clinical vignettes with a single gold answer [10]. Other evaluations adapt case vignettes to simulate real clinical reasoning, or source questions from public medical forums to reflect authentic patient queries [5, 6, 7, 8, 11, 12]. More recently, HealthBench introduced a large-scale benchmark of 5,000 multi-turn dialogues prepared by 262 physicians across 60 countries, with 48,562 unique rubric criteria spanning accuracy, completeness, communication, context-awareness, and instruction-following [13].

2.2 Evaluation Methods

Most studies using multiple-choice datasets report standard classification metrics such as accuracy, precision, and recall. For free-text responses, evaluations may rely on expert grading, automatic similarity measures (e.g., BLEU, BERTScore), or Likert-scale expert judgments [14]. Recent work also shows that grader-LLMs can achieve inter-rater reliability comparable to human physicians when scoring responses [13].

2.3 Critical Characteristics of Medical LLMs for Deployment

While accuracy dominates current evaluation practice, multiple studies emphasize that safe deployment of medical LLMs requires attention to additional characteristics [3, 4]. These are often not yet systematically measured, but they are repeatedly identified as necessary for real-world use:

- **Interaction quality** - Clinical communication requires eliciting history, tailoring explanations, and showing empathy [15, 16].
- **Safety and risk** - Hallucinations, unsafe recommendations, and contradictions are recognized hazards when interacting with LLMs [4, 14].
- **Reliability and robustness** - Performance frequently deteriorates under noisy, adversarial, or out-of-distribution inputs. Moreover, identical prompts can produce inconsistent responses across conversations [3].
- **Transparency and grounding** - Evidence citation and traceable reasoning are seen as crucial for clinical trust [3, 4].

- **Calibration and deferral** - Alignment of stated confidence with correctness and appropriate referral to clinicians [3].
- **Workflow and human factors** - Usability, efficiency, and cognitive load shape adoption [2].
- **Governance and equity** - Regulatory frameworks such as the EU AI Act impose obligations for transparency, robustness, and oversight for AI applications [17, 18].

In summary, existing evaluations rely on heterogeneous datasets and methods, often limited to knowledge checks or isolated dimensions [3, 4, 5, 6, 7, 8]. Although recent benchmarks like HealthBench expand coverage, there is still no unified, clinically grounded framework that systematically captures the breadth of requirements for safe deployment [13]. To address this gap, we introduce M-LEAF (Medical LLM Evaluation Across Facets), a multidimensional framework for assessing medical LLMs. We further demonstrate its application in two pilot studies that compare GPT-4o with the HomeDOCtor system.

3 Method

3.1 Design Process of the M-LEAF Framework

Derivation

The M-LEAF framework was derived through a synthesis of evidence from prior evaluations of LLMs in healthcare, literature reviews pointing out the disadvantages of these evaluations, common clinical practice requirements, and emerging regulatory standards covered in Section 2. We grouped the requirements identified in the literature into eight pillars that reflect the key functions a medical LLM must fulfill to be clinically useful and safe. Each pillar contains concrete dimensions with what to measure, candidate metrics, and recommended protocols. The pillars are: (P1) Clinical Task Fidelity, (P2) Interaction Quality, (P3) Safety & Risk, (P4) Reliability & Robustness, (P5) Transparency, Grounding & Explainability, (P6) Calibration, Uncertainty & Consistency, (P7) Governance, Equity & Data Protection, and (P8) Workflow & Human Factors.

Evaluation setup

Each dimension in M-LEAF is assessed using standardized vignettes or prompts that are tailored to the specific requirement being tested. In some cases, such as history-taking or consistency, these vignettes take the form of multi-turn scripts. All model outputs are reviewed by qualified human raters.

Scoring and aggregation

M-LEAF expresses every dimension as a 0–5 score. There are two ways a dimension reaches that score:

(1) Rubric-native dimensions (e.g., empathy, clarity, history-taking) are rated directly on a 0–5 expert rubric.
(2) Task-metric-native dimensions (e.g., accuracy, sensitivity, error rates, % degradation) first produce a raw task metric, which is then converted to a 0–5 score using the conversion model below.

Mappings are monotonic, ensuring that higher scores always reflect better clinical performance. Raw task metrics are translated to the 0–5 scale using the following scheme:

(1) "Higher is better" metrics (e.g., accuracy) - 0: <20%; 1: 20–39%; 2: 40–59%; 3: 60–74%; 4: 75–89%; 5: >89%
(2) "Lower is better" (e.g., error rates) - 5: <0.5%; 4: 0.5–2%; 3: 2–5%; 2: 5–10%; 1: 10–20%; 0: >20%

Scores may be reported at the sub-dimension, pillar, or aggregated framework level. Aggregation does not compensate for critical weaknesses, if any dimension receives a score of less than 1, this is classified as a critical failure, and the overall system is considered inadequate for clinical deployment, irrespective of high performance in other areas. This rule ensures that serious hazards are not obscured by averaging across dimensions. Where relevant, aggregated scores can be weighted to reflect the priorities of different stakeholder groups (e.g., patient-facing versus clinician-facing applications), but such weightings must be reported transparently and cannot nullify the effect of critical failures.

3.2 M-LEAF Framework

P1 — Clinical Task Fidelity

P1.1 Diagnostic Reasoning & Differential Quality

Description: Ability to identify the correct diagnosis from clinical vignettes; **Protocol**: USMLE/MedQA vignettes; **Metric**: Top-k accuracy on exam-style vignettes

P1.2 Emergency Referral

Description: Ability to correctly triage clinical cases into emergent, urgent, or non-urgent categories, ensuring safety by not missing true emergencies; **Protocol**: Standardized triage vignettes annotated by emergency physicians into emergent/urgent/non-urgent; model outputs compared to gold labels; **Metric**: Sensitivity for emergent cases; false negative rate for emergent cases reported separately.

P1.3 Management Recommendations

Description: Appropriateness and specificity of recommended next steps; **Protocol**: Present the model with short clinical vignettes (some containing hidden pitfalls such as contraindications). Clinicians review the model's recommended next steps and rate how clear, specific, and appropriate they are; **Metric**: Expert actionability score (0–5).

P2 — Interaction Quality

P2.1 History-Taking Quality

Description: Ability of the model to ask relevant and sufficient follow-up questions to gather an adequate patient history in dialogue; **Protocol**: Simulated patient dialogue vignettes, starting from a single presenting symptom (e.g., "my head hurts"). Each vignette has a predefined condition and checklist of essential history items; the simulated patient reveals these only if the model asks. Clinicians review whether the model's questioning covers the checklist; **Metric**: Expert rubric score (0–5) for adequacy of history.

P2.2 Empathy

Description: Ability of the model to respond with sensitivity and compassion, showing understanding and support for patient concerns; **Protocol**: Patient vignettes containing emotional or distress cues (e.g., anxiety, chronic pain, receiving bad news). Clinicians rate the model's responses for empathy, tone, and appropriateness; **Metric**: Expert rubric score (0–5) for empathy.

P2.3 Style & Terminology

Description: Clarity, conciseness, and appropriateness of language, including correct use of clinical terminology and suitability for the intended audience (patient vs. clinician); **Protocol**: Patient communication vignettes where the model generates explanations or instructions. Clinicians and/or trained raters review outputs for readability, correctness of terminology, and appropriateness of tone; readability indices (e.g., Flesch–Kincaid) may be used as a supporting measure; **Metric**: Expert rubric score (0–5) for clarity and terminology appropriateness, with readability index reported as a secondary metric.

P3 — Safety & Risk

P3.1 Hallucination & Fabrication

Description: Tendency of the model to produce unsupported, fabricated, or medically inaccurate claims; **Protocol**: Clinical vignettes and fact-based queries tested under knowledge-withholding or RAG-ablation conditions (sources removed or blocked). Clinicians review outputs to identify unsupported statements or fabrications; **Metric**: Hallucination rate (% of responses containing unsupported or inaccurate claims).

P3.2 Hazardous Content & Contraindications

Description: Suggestions that could cause patient harm, violate known contraindications, or recommend clearly unsafe actions; **Protocol**: Present adversarial or stress-test vignettes (e.g., drug–drug interaction, high-risk comorbidity, "red flag" symptom). Clinicians review whether the model's output contains unsafe or contraindicated advice; **Metric**: Unsafe-recommendation rate (% of outputs rated unsafe), optionally stratified by severity of harm (e.g., minor, moderate, severe).

P3.3 Consistency

Description: Stability of the model's answers across turns, specifically avoiding self-contradiction when the same facts are repeated; **Protocol**: Multi-turn dialogue vignettes where key facts (e.g., patient age, allergy, medication) are re-introduced later in the conversation. Clinicians review whether the model's responses remain consistent with earlier information; **Metric**: Contradiction rate (% of cases where the model changes or contradicts its own earlier statements).

P4 — Reliability & Robustness

P4.1 Ambiguity

Description: Ability of the model to handle incomplete inputs without major performance degradation; **Protocol**: Stress-test vignettes where essential information is systematically withheld. Compare model outputs against gold answers or clinician ratings; **Metric**: Relative degradation in accuracy compared to baseline performance on clean vignettes (e.g., drop in top-k diagnostic accuracy).

P4.2 Noise & Translation Robustness

Description: Ability of the model to remain accurate when handling noisy or linguistically varied inputs (e.g., typos, spelling mistakes, dialects); **Protocol**: Present a noisy-input vignette suite where baseline cases are systematically modified with spelling errors, dialectal variants, or mixed-language phrasing. Compare model outputs against gold answers or clinician ratings; **Metric**: Relative degradation in accuracy compared to clean-baseline vignettes (e.g., drop in diagnostic accuracy).

P4.3 Prompt-Injection & Jailbreak Resilience

Description: Ability of the model to resist malicious or adversarial prompts that attempt to override safety rules or elicit disallowed outputs; **Protocol**: Red-team evaluation using a library of adversarial prompts (e.g., attempts to bypass safety filters, inject hidden instructions, or coerce unsafe outputs). Clinicians and security reviewers assess whether the model complied or resisted; **Metric**: Attack success rate (% of adversarial prompts that cause unsafe or policy-violating outputs).

P5 — Transparency, Grounding & Explainability

P5.1 Evidence Grounding

Description: Degree to which model claims are supported by verifiable, high-quality sources when retrieval or citation is expected; **Protocol**: Present fact-based vignettes or questions where supporting evidence is available (e.g., guideline, article abstract, textbook snippet). The model is required to provide both an answer and a citation. Clinicians verify whether the cited sources truly support the claims; **Metric**: Citation precision (% of provided citations judged appropriate by reviewers).

P5.2 Explanation Quality

Description: Ability of the model to provide reasoning that is faithful to clinical evidence and relevant to the presented case; **Protocol**: Present vignettes where the model is asked not only for an answer but also to explain its reasoning. Independent clinicians review whether the explanations are accurate, clinically appropriate, and consistent with the final recommendation; **Metric**: Expert faithfulness rating (0–5), where 0 = misleading or fabricated rationale and 5 = fully faithful and clinically relevant reasoning trace.

P5.3 Traceability & Auditability

Description: Availability of logging, versioning, and provenance information sufficient to allow external audit and accountability; **Protocol**: Review system documentation and deployment records using a structured checklist that covers model versioning, data provenance, logging of outputs, and incident reporting; **Metric**: Documentation-audit pass rate (percentage of required checklist items present and adequate).

P6 — Calibration, Uncertainty & Consistency

P6.1 Confidence Calibration

Description: Alignment of the model's stated confidence with the correctness of its answers; **Protocol**: Present vignette sets where the model must provide both a prediction and an associated confidence score. Predictions are binned by confidence level and compared against ground truth to assess calibration; **Metric**: Expected Calibration Error (ECE), reported as % deviation between predicted confidence and observed accuracy.

P6.2 Abstention & Clinician Deferral

Description: Ability of the model to appropriately abstain or defer to a clinician when it lacks knowledge or when a case requires human judgment; **Protocol**: Use vignettes labeled with a gold "deferral" requirement. The model is forced to choose between answering or abstaining, and outputs are scored against the gold label; **Metric**: Appropriate-deferral rate (% of cases where abstention is correctly chosen when indicated).

P6.3 Consistency

Description: Stability of model outputs across repeated runs under different randomness settings; **Protocol**: Present the same vignettes repeatedly under fixed seeds and multiple temperature settings. Aggregate results to assess whether accuracy remains stable across runs; **Metric**: Coefficient of variation of accuracy across repeated generations

P7 — Governance, Equity & Data Protection

P7.1 Fairness & Bias

Description: Ability of the model to perform consistently across demographic groups without introducing systematic disparities; **Protocol**: Apply synthetic demographic perturbations to vignettes (e.g., altering age, gender, ethnicity markers while keeping clinical facts constant) and compare outputs; **Metric**: Parity gap in error rates across protected subgroups (% difference in performance).

P7.2 Privacy & GDPR Compliance

Description: Extent to which the system complies with data protection and minimisation requirements set by regulations such as GDPR or the EU AI Act; **Protocol**: Evaluate system documentation and data handling against a structured compliance checklist (e.g., Future of Life Institute – EU AI Act Compliance Checker [19]); **Metric**: Checklist pass rate (% of required privacy and data protection items met).

P8 — Workflow & Human Factors

P8.1 Escalation Quality
Description: Clarity and appropriateness of the model's handoff or escalation recommendations for patients or clinicians; **Protocol**: Present simulated handoff notes or referral instructions generated by the model. Clinicians review them for clarity, adequacy of information, and appropriateness of escalation; **Metric**: Clinician rubric score (0–5) for handoff clarity and appropriateness.
P8.2 Perceived Workload
Description: Impact of the system on clinician workload and usability; **Protocol**: Clinicians use the system in simulated tasks and subsequently complete the NASA-TLX questionnaire to assess perceived workload; **Metric**: Mean NASA-TLX score, reported as a quantitative measure of perceived workload (lower is better).

3.3 Study 1: Initial Pillar-Level Evaluation

Rationale and scope
Study 1 was designed as a pilot application of M-LEAF to test the feasibility of rating multiple dimensions in parallel on a shared set of vignettes. From the framework, we selected eight dimensions spanning four pillars: Clinical Task Fidelity (accuracy, referral appropriateness); Interaction Quality (follow-up questions, empathy, style, terminology); Safety & Risk (absence of hallucinations); Transparency & Explainability (quality of explanation). These dimensions were chosen because they represent clinically salient requirements that can be assessed through vignette outputs and they balance reasoning, safety, and patient-facing communication.
Dataset and prompting
We drew on the Avey AI Benchmark Vignette Suite [20] as the basis for our prompts. From this resource, we created 100 standardized vignettes in Slovenian, covering a spectrum of diagnostic complexity from routine primary care cases to urgent and life-threatening conditions. Each vignette included structured fields (age, sex, chief complaint, clinical history). The same 100 vignettes were used across all eight selected dimensions to ensure consistency and comparability of ratings. All interactions with evaluated systems were done through the systems' public GUIs.
Evaluated systems
The evaluated systems were GPT-4o and HomeDOCtor. HomeDOCtor is a diagnostic assistant that integrates medical knowledge and explicit instructions on how to effectively communicate as a diagnostic assistant. It operates as a Retrieval-Augmented Generation (RAG) system layered on top of a base LLM model (e.g., GPT-4o), combining Slovenian medical content with the generative capabilities of an LLM [21]. In our study, the base LLM on which HomeDOCtor was layered on was GPT-4o.
Raters and scoring
Final-year Slovenian medical students served as raters. Each rater assessed a subset of system outputs; there was no overlap across raters, so inter-rater reliability was not computed. All eight dimensions were scored on a 0–5 scale using the M-LEAF rubric. Dimensions defined by raw metrics (e.g., accuracy, hallucination rate) were first quantified and then mapped to the 0–5 rubric as described in Section 3.1.
Statistical analysis
We compared rating distributions between systems using Pearson's χ^2 test per dimension. As a complementary analysis, we applied a Mann-Whitney U test on expanded counts. Results were reported at the dimension level.

3.4 Study 2: Full Framework Application

Rationale and scope
Study 2 implemented the complete M-LEAF framework across all eight pillars, with one representative task or vignette selected for each dimension. The aim was to demonstrate the operationalisation of the full framework in practice. As only a single example was used per dimension, this study should be regarded as preliminary. The evaluated systems were GPT-4o and HomeDOCtor.
Dataset and prompting
Clinical reasoning and interaction dimensions were tested using vignette-style prompts prepared in accordance with the protocols specified in Section 3.2. Dimensions addressing governance, privacy, or auditability were assessed using structured documentation checklists.
Raters and scoring
The same two final-year medical students who participated in Study 1 served as raters. They scored all dimensions on the 0–5 M-LEAF scale, with raw task metrics converted as described in Section 3.1.

4 Results

4.1 Study 1

Aggregate scores were uniformly high across dimensions for both evaluate systems, with HomeDOCtor trending higher on the dimensions of: Accuracy, Empathy, Quality of Explanation, Referral Appropriateness and Style. Despite these trends, no statistically significant differences were observed. In Figure 1 we can see the scores across dimensions.

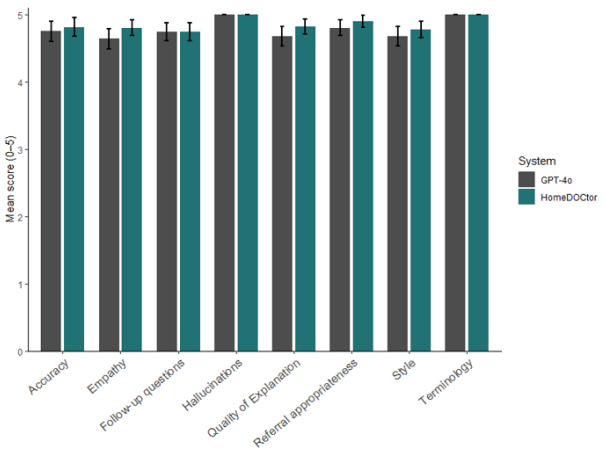


Figure 1: Dimension-level mean scores with 95% CI for GPT-4o vs. HomeDOCtor.

4.2 Study 2

Figure 2 presents the results of Study 2, indicating high scores for both GPT-4o and the HomeDOCtor system, with the latter trending higher across most dimensions.

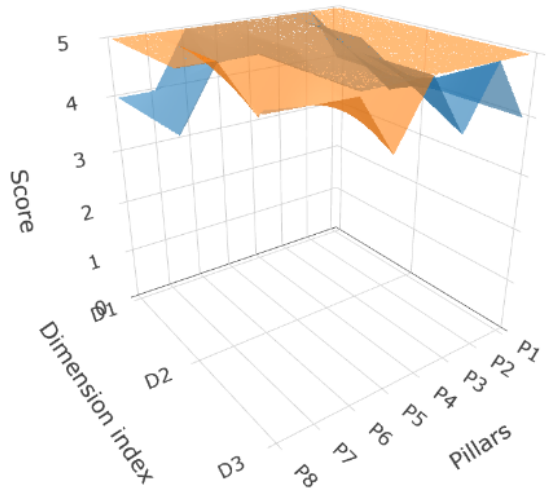


Figure 2: Comparison of GPT-4o and HomeDOCtor through the M-LEAF framework.

5 Discussion

5.1 Conclusion

LLMs are being increasingly used for medical purposes, where avoiding harm, enabling deferral, and providing clear explanations is just as critical as achieving high diagnostic accuracy [2, 3, 4]. The M-LEAF framework addresses this by consolidating diverse metrics into a unified structure. The preliminary results of both studies demonstrate high question-answering performance for GPT-4o and the HomeDOCtor system, which is consistent with findings reported in the existing literature [5, 6, 7, 8]. Additionally, we also showed that good results of LLMs in the medical context are not confined to accuracy alone, but also to other dimensions of the diagnostic process. With these results we conclude that M-LEAF represents a comprehensive framework for evaluating medical LLM applications. We invite the community to adopt and iterate on M-LEAF to make evaluations clinically meaningful.

5.2 Limitations and future work

One limitation of M-LEAF is that some of the proposed metrics, such as empathy, are based on evolving standards that currently lack established benchmarks. As a result, the benchmarks proposed in our study may not be as robust as those available for accuracy. Metrics like empathy are also more vulnerable to subjective variation in rater assessments. Furthermore, certain dimensions, including privacy and fairness, require specialised audits that go beyond vignette-based studies, which makes them more difficult to implement. Additionally, our two case studies are preliminary, therefore their results should be interpreted with caution. Future work should apply M-LEAF in larger studies to enhance its generalisability.

References

[1] World Health Organization. Regional Office for Europe, "Health and care workforce in Europe: time to act," World Health Organization. Regional Office for Europe, Tech. Rep., Sep. 2022. [Online]. Available: https://www.who.int/europe/publications/i/item/9789289058339

[2] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "Ai in health and medicine," *Nature Medicine*, vol. 28, no. 1, pp. 31–38, Jan. 2022, ISSN: 1546-170X. DOI: 10.1038/s41591-021-01614-0 [Online]. Available: http://dx.doi.org/10.1038/s41591-021-01614-0

[3] T. Y. C. Tam et al., "A framework for human evaluation of large language models in healthcare derived from literature review," *npj Digital Medicine*, vol. 7, no. 1, Sep. 2024, ISSN: 2398-6352. DOI: 10.1038/s41746-024-01258-7 [Online]. Available: http://dx.doi.org/10.1038/s41746-024-01258-7

[4] S. Bedi et al., "Testing and evaluation of health care applications of large language models: A systematic review," *JAMA*, vol. 333, no. 4, p. 319, Jan. 2025, ISSN: 0098-7484. DOI: 10.1001/jama.2024.21700 [Online]. Available: http://dx.doi.org/10.1001/jama.2024.21700

[5] A. Gilson et al., "How does chatgpt perform on the united states medical licensing examination (usmle)? the implications of large language models for medical education and knowledge assessment," *JMIR Medical Education*, vol. 9, e45312, Feb. 2023, ISSN: 2369-3762. DOI: 10.2196/45312 [Online]. Available: http://dx.doi.org/10.2196/45312

[6] Y. Yanagita, D. Yokokawa, S. Uchida, J. Tawara, and M. Ikusaka, "Accuracy of chatgpt on medical questions in the national medical licensing examination in japan: Evaluation study," *JMIR Formative Research*, vol. 7, e48023, Oct. 2023, ISSN: 2561-326X. DOI: 10.2196/48023 [Online]. Available: http://dx.doi.org/10.2196/48023

[7] J. B. Longwell et al., "Performance of large language models on medical oncology examination questions," *JAMA Network Open*, vol. 7, no. 6, e2417641, Jun. 2024, ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2024.17641 [Online]. Available: http://dx.doi.org/10.1001/jamanetworkopen.2024.17641

[8] M. Gams, T. Horvat, Ž. Kolar, P. Kocuvan, K. Mishev, and M. S. Misheva, "Evaluating a nationally localized ai chatbot for personalized primary care guidance: Insights from the homedoctor deployment in slovenia," *Healthcare*, vol. 13, no. 15, p. 1843, Jul. 2025, ISSN: 2227-9032. DOI: 10.3390/healthcare13151843 [Online]. Available: http://dx.doi.org/10.3390/healthcare13151843

[9] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, "Pubmedqa: A dataset for biomedical research question answering," 2019. DOI: 10.48550/ARXIV.1909.06146 [Online]. Available: https://arxiv.org/abs/1909.06146

[10] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, "What disease does this patient have? a large-scale open domain question answering dataset from medical exams," *Applied Sciences*, vol. 11, no. 14, p. 6421, Jul. 2021, ISSN: 2076-3417. DOI: 10.3390/app11146421 [Online]. Available: http://dx.doi.org/10.3390/app11146421

[11] E. Goh et al., "Large language model influence on diagnostic reasoning: A randomized clinical trial," *JAMA Network Open*, vol. 7, no. 10, e2440969, Oct. 2024, ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2024.40969 [Online]. Available: http://dx.doi.org/10.1001/jamanetworkopen.2024.40969

[12] J. W. Ayers et al., "Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum," *JAMA Internal Medicine*, vol. 183, no. 6, p. 589, Jun. 2023, ISSN: 2168-6106. DOI: 10.1001/jamainternmed.2023.1838 [Online]. Available: http://dx.doi.org/10.1001/jamainternmed.2023.1838

[13] R. K. Arora et al., "Healthbench: Evaluating large language models towards improved human health," 2025. DOI: 10.48550/ARXIV.2505.08775 [Online]. Available: https://arxiv.org/abs/2505.08775

[14] D. Wang and S. Zhang, "Large language models in medical and healthcare fields: Applications, advances, and challenges," *Artificial Intelligence Review*, vol. 57, no. 11, Sep. 2024, ISSN: 1573-7462. DOI: 10.1007/s10462-024-10921-0 [Online]. Available: http://dx.doi.org/10.1007/s10462-024-10921-0

[15] J. Halpern, "What is clinical empathy?" *Journal of General Internal Medicine*, vol. 18, no. 8, pp. 670–674, Aug. 2003, ISSN: 1525-1497. DOI: 10.1046/j.1525-1497.2003.21017.x [Online]. Available: http://dx.doi.org/10.1046/j.1525-1497.2003.21017.x

[16] S. Johri et al., "An evaluation framework for clinical use of large language models in patient interaction tasks," *Nature Medicine*, vol. 31, no. 1, pp. 77–86, Jan. 2025, ISSN: 1546-170X. DOI: 10.1038/s41591-024-03328-5 [Online]. Available: http://dx.doi.org/10.1038/s41591-024-03328-5

[17] S. Freeman et al., "Developing an ai governance framework for safe and responsible ai in health care organizations: Protocol for a multimethod study," *JMIR Research Protocols*, vol. 14, e75702, Jul. 2025, ISSN: 1929-0748. DOI: 10.2196/75702 [Online]. Available: http://dx.doi.org/10.2196/75702

[18] European Parliament and Council of the European Union, *Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*, Official Journal of the European Union (OJ L), 12 July 2024, 2024. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[19] Future of Life Institute. "EU AI Act Compliance Checker | EU Artificial Intelligence Act," Accessed: Sep. 15, 2025. [Online]. Available: https://artificialintelligenceact.eu/assessment/eu-ai-act-compliance-checker/

[20] Avey. "Benchmark vignette suite," Accessed: Sep. 15, 2025. [Online]. Available: https://avey.ai/research/avey-accurate-ai-algorithm/benchmark-vignette-suite

[21] M. Zadobovšek, P. Kocuvan, and M. Gams, "Homedoctor app: Integrating medical knowledge into gpt for personal health counseling," in *Information Society 2024: ChatGPT in Medicine*, Ljubljana, Slovenia, Oct. 2024.

Evaluating the Accuracy and Quality of ChatGPT-4o Responses to Patient Questions on Reddit

Mihailo Svetozarević†
Clinic for Neurology
University Clinical Center Niš
Niš, Serbia
mihailo.svetozarevic@gmail.com

Isidora Svetozarevic
Center for Radiology
University Clinical Center Niš
Niš, Serbia
isidora_jankovic@yahoo.com

Sonja Jankovic
Center for Radiology
University Clinical Center Niš
Niš, Serbia
sonjasgirl@gmail.com

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

Abstract

The rapid integration of large language models (LLMs) into healthcare communication has raised questions about their accuracy, safety, and usefulness for patients seeking medical advice online. This study evaluated the performance of ChatGPT-4o in responding to epilepsy-related patient questions posted on the r/AskDocs subreddit. A total of 110 questions were selected based on the keywords epilepsy, seizure, and seizure disorder, filtered by the "physician responded" flair. Responses generated by ChatGPT-4o were independently assessed by four physicians across multiple domains including accuracy, comprehensiveness, clarity, relevance, and empathy as well as binary assessments of bias, factuality, fabrication, falsification, plagiarism, harm, reasoning, and currency. Results showed that most of the responses were rated as good or very good, with particularly high scores for accuracy, clarity, relevance, and comprehensiveness, while empathy was consistently lower. These findings suggest that ChatGPT-4o may serve as a useful complementary tool for patient education and engagement in epilepsy, though it cannot replace professional medical consultation. Future research should further investigate its role in clinical practice and strategies for improving empathetic communication in AI generated responses.

Keywords

ChatGPT-4o, epilepsy, seizure disorder, artificial intelligence, patient communication, evaluation, accuracy, empathy, large language models

1. Introduction

In medicine, large language models (LLMs) are increasingly applied to diverse tasks, including information extraction from electronic health records, scientific writing support, patient care documentation, and even clinical guideline development. Importantly, the use of LLMs is not limited to healthcare professionals. Patients themselves are increasingly experimenting with these tools, as new models and updated versions create the impression of rapidly expanding capabilities from one year to the next. This steady rise in LLM use coincides with an already well-established pattern: health information is often sought online before consulting a physician. In the United States, survey data show that about six in ten adults aged 18 to 29 report being online almost constantly, with somewhat smaller but still substantial proportions in older groups. Such an environment directly encourages digital health information-seeking behavior and frequent encounters with LLM-based tools. [1,2]

The COVID-19 pandemic further accelerated the adoption of virtual health care and normalized the use of public online forums where patients seek advice sometimes from reliable professionals, but often from peers or unverified sources. Reddit, along with similar platforms, has become a representative setting for "real-world" patient - physician interactions in an asynchronous, text-based format. The potential advantages of LLMs in this context are considerable. They can rapidly synthesize information, explain disease mechanisms in accessible language, highlight red-flag symptoms, and point to relevant resources, all while being available around the clock. They are also generally intuitive to use, even for individuals with limited health literacy. Furthermore, recent evaluations suggest that LLM-generated responses may convey greater empathy and clarity than

Information Society 2025, 6–10 October 2025, Ljubljana, Slovenia

https://doi.org/10.70314/is.2025.gptzdravje.4

physician-written answers in some online settings, potentially improving comprehension and adherence. Yet, the risks remain substantial. LLMs are prone to generating hallucinations plausible but incorrect statements while omitting key information or inferring unstated details. In a high-risk domain such as medicine, these limitations render unsupervised use unsafe. The most recent literature emphasizes that hallucinations and omissions are intrinsic to current LLM architectures, and that without rigorous safeguards - such as benchmarking, oversight, and validation - clinical deployment should not proceed unchecked. Beyond technical concerns, the rapid spread of LLM use also raises new ethical and societal challenges. Healthcare is guided by strict ethical norms, professional duties, and societal responsibilities, and recent case reports highlight instances where LLM outputs, including those from ChatGPT, have contributed to harmful and potentially life-threatening outcomes. [3]

In this review, we focus on a specific clinical domain - epilepsy and other seizure disorders where the need for reliable information is particularly acute. Epilepsy is a chronic, often lifelong condition with a heterogeneous clinical presentation, typically beginning in childhood or young adulthood. Patients with epilepsy frequently have questions about treatment options, drug interactions, lifestyle considerations, and safety precautions. Studies have shown that a significant proportion of individuals with epilepsy actively search for information online, both on general and disease-specific topics. Analyses of search patterns (for example, on Wikipedia) have revealed strong public interest and episodic peaks in epilepsy-related queries. More recent research indicates that people with epilepsy engage in online health information seeking at higher rates than many other patient groups, underscoring the importance of understanding how LLM responses might influence their perceptions and behaviors. However there are both potential benefits and inherent limitations of LLMs in epilepsy care as shown by recent review articles. [4,5,6,7,8,9]

Despite the growing body of literature on LLMs in medicine, they remain insufficiently reliable for routine, uncontrolled use. A notable gap exists: few studies evaluate LLMs from the patient's perspective, particularly using real-world data drawn from public forums. Our study is designed to address this gap. Specifically, we assess whether responses generated by OpenAI's ChatGPT-4 meet the needs of people with epilepsy who ask questions on r/AskDocs. Physicians serve as expert evaluators not to arbitrate "on behalf of patients," but to operationalize criteria of quality, utility, accuracy, and safety in line with real user needs. We argue that this design places the patient - LLM relationship at the center of the analysis, while leveraging medical expertise to standardize evaluation metrics and identify areas where safeguards or clinical verification remain necessary. In this framework, Reddit provides a natural, heterogeneous, and timely source of patient queries, enabling an evaluation of LLM responses under conditions that approximate the realities of everyday patient information-seeking. . [3,7,10,11]

2. Material and Method

In the intial phase of the study we collected a total of 110 patient questions from the subreddit r/AskDocs, one of the more active medical communities on reddit with over half a milion active participants. Questions were identified using a filtered search using keywords „epilepsy", „seizure" and „seizure disorder". To ensure quality only posts submitted within the past 12 months and those that received at least one verified physican response (marked with the flair „physician responded") were included. Out of the selected 110 questions, 4 were excluded due to being duplicates or irrelevant to the subject matter.

For each selected question a response was generated using ChatGPT 4.0. These responses were then independently evaluated by four certified physicians – one neurologist, one radiologist, one neurology resident and one radiology resident. The raters were blinded to each other's assessments and did not consult each other during the evaulation process. Interrater agreements were reached using Fleiss Kappa with minimal discrepanies observed among evaluators.

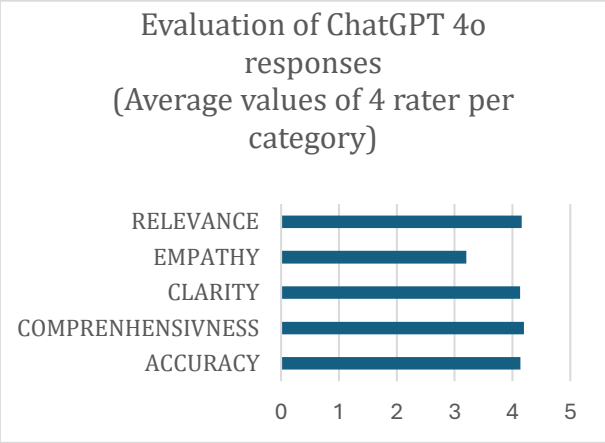
Evaluations were made using predefined dimension with a modified Likert scale (1-5). The dimensions assessed were Accuracy, Comprehensiveness, Clarity, Empathy, Relevance. Additional dimensions were assessed using categorical ratings (Yes/No responses). These dimensions were Reasoning, Currency, Bias, Harm, Factuality, Fabrication, Falsification and Plagiarism.

3. Results

Overall the raters found that ChatGPT 4.0 respones were very positive with approximately 80% of answers classified as „good" or „very good" across all dimensions on the Likert Scale. Most answers were considered factually correct, we found no responses to be incorrect. Most answeres were very thorough andcc easily understandable with language that the raters believe cover all educational specters. We found no instanes of outdated recommendations and all responses were deemed to be concise, without unnecessary and overbearing details.

Regarding categorical measures we did not find any cases of bias, harm, fabrication, falsification. All answers gave information that could be easily varified against standard medical sources. The lowest scoring dimension was empathy as we found most answers to be on average good or decent with no responses being explicitly poor.

All together, these results suggest that ChatGPT 4.0 is capable of generating accurate, clear and relevant responses to patient questions about epilepsy with the primary limitation being in the domain of empathetic responses.



4. Discussion

In this study, we examined the usability of responses generated by ChatGPT 4.0 in comparison to neurologists' answers to patient questions about epilepsy on Reddit, specifically the subreddit r/AskDocs. This community is one of the largest and most active health forums online, with over half a million members and hundreds of new patient questions submitted daily. A particular strength of this platform lies in its anonymity: users can ask sensitive medical questions more openly than they might in a clinical encounter, which results in a broader and more candid spectrum of concerns. Additionally, r/AskDocs is actively moderated and follows strict rules medical advice is permitted only from verified physicians (marked by a special flair), while other users are restricted to sharing personal experiences. This structure ensures a basic level of quality control and provides a reliable basis for comparing physician responses with those of ChatGPT. We believe this makes r/AskDocs a relevant and valid environment for evaluating the potential of large language models (LLMs) in a medical setting.

Our findings complement recent research done by Fennig and colleagues [12], in which LLM models were used to analyze tens of thousands of Reddit posts to identify topics and concerns that epilepsy patients often do not bring up in clinical settings. That work found significant patterns such as stigma, emotional distress, substance use, and seizure description high-engagement topics that are outside of standard outpatient conversations and often not given adequate space in the clinical conversation. This confirms that LLM models are not only for providing answers, but also for a deeper understanding of patient needs, which further justifies the use of r/AskDocs as a source of realistic and relevant questions for our study.

Our findings indicate that ChatGPT 4.0 generally provides accurate, relevant, and comprehensive answers. Importantly, no response was deemed explicitly incorrect, underscoring the potential of such tools to deliver reliable medical information for patients with epilepsy. However, the model consistently showed weaker performance in conveying empathy compared to physicians. This limitation has been noted in previous studies, which emphasize that while LLMs can reproduce medical content accurately, they struggle to replicate the human aspects of communication such as reassurance, compassion, and emotional support. [1,6,8]

The overall impression of the neurologists was that the ChatGPT 4.0 responses were mostly "acceptable" or "good", while a smaller number were rated as "very good". Nevertheless, doctors generally gave somewhat better answers, but the difference was not large. This finding is consistent with the results of a study by Ayers and colleagues., who also found that chatbot responses can be of similar or even better quality in certain dimensions, but with limitations in empathy. [1]

It is important to point out that our results should be seen in the context of the increasing number of patients using the Internet for epilepsy information and potentially changing therapy based on information obtained online. Previous studies have shown that patients with epilepsy frequently search the Internet to learn more about their disease [3,4], while more recent studies indicate a high rate of use of digital sources of health information in this population. [5] Precisely because of this, the ability of large language models to generate correct and comprehensible answers is of particular importance.

Even though our findings are encouraging, it is necessary to emphasize the potential risks. The literature on LLMs in medicine warns of the phenomenon of "hallucinations", i.e. giving confident but incorrect answers. [6,7] Although in our series no answer was explicitly wrong, such cases were not excluded in a larger sample, especially in more complex clinical scenarios. In addition, a critical review of LLMs in epileptology indicates that current tools may be useful for patient and physician education, but are not ready for routine, uncontrolled clinical application. [8]

Finally, it should be emphasized that the focus of our study was the attitude of patients towards the responses of LLMs, while doctors had the role of mediators in quality assessment. This kind of perspective can be significant for future research, as it opens up space for a better understanding of how patients value and perceive such tools compared to traditional medical sources

5. Conclusion

This study demonstrates that ChatGPT 4.0 provides responses to patient questions about epilepsy that are largely accurate, relevant, clear, and comprehensive. However, the limitations observed - especially regarding emotional support and nuanced communication highlight that ChatGPT cannot replace professional medical consultation. Instead, its role should be considered complementary, supporting patient education and engagement, while final interpretation and guidance remain within the responsibility of qualified healthcare professionals.

Acknowledgements

Views and opinions expressed in this paper are those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor any other authority can be held responsible for them. All authors contributed equally in the final version of this paper.

References

1. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, 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):589–96. doi: 10.1001/jamainternmed.2023.1838
2. Pew Research Center. Mobile technology and home broadband 2021. Available from: https://www.pewresearch.org/internet/2021/06/03/mobile-technology-and-home-broadband-2021/
3. Omar M, Sorin V, Collins JD, et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. Commun Med (Lond). 2025;5:44. doi: 10.1038/s43856-025-01021-3
4. Liu J, Dong X, Mao Y, et al. Internet usage for health information by patients with epilepsy. Epilepsy Behav. 2013;29(1):110–3. doi: 10.1016/j.seizure.2013.06.007
5. Brigo F, Erro R, Marangi A, et al. Why do people Google epilepsy? An infodemiological study of online behavior for epilepsy-related search terms. Epilepsy Behav. 2015;45:128–33. doi: 10.1016/j.yebeh.2013.11.020
6. Bingöl N, Mutluay FK, Erbaş O. Determining the health-seeking behaviors of people with epilepsy. Epilepsy Behav. 2024;152:109331. doi: 10.1016/j.yebeh.2024.110063
7. Bélisle-Pipon JC, et al. Why we need to be careful with large language models in medicine and healthcare. AI & Ethics. 2024. doi: 10.3389/fmed.2024.1495582
8. Asgari E, Montaña-Brown N, Dubois M, et al. A framework to assess clinical safety and hallucination rates of large language models for medical text summarization. npj Digit Med. 2025;8(1):19. doi: 10.1038/s41746-025-01670-7
9. García-Azorín D, Bhatia R, et al. Potential merits and flaws of large language models in epilepsy care: A critical review. Epilepsia. 2024;65(4):873–886. doi: 10.1111/epi.17907
10. The Verge. Google's healthcare AI made up a body part. The Verge. 2025 May 7. Available from: https://www.theverge.com/2025/05/07/google-healthcae-ai-made-up-body-part
11. Auvin S, Nabbout R, et al. Quality of health information about epilepsy on the Internet. Arch Pediatr. 2013;20(6):603–7. doi: 10.1016/j.neurol.2012.08.008
12. Fennig U, Yom-Tov E, Savitzky L, Nissan J, Altman K, Loebenstein R, Boxer M, Weinberg N, Gofrit S G, Maggio N. Bridging the conversational gap in epilepsy: Using large language models to reveal insights into patient behavior and concerns from online discussions. Epilepsia. 2025;66(3):686–699. doi: 10.1111/epi.18226

Evaluating Large Language Models for Privacy-Sensitive Healthcare Applications

Tadej Horvat
Department of Intelligent Systems
Jožef Stefan Institute
Ljubljana, Slovenia
tadej.horvat@ijs.si

Žan Roštan
Fakulteta za računalništvo in informatiko
Ljubljana, Slovenia
zan.rostan@gmail.com

Jakob Jaš
Fakulteta za računalništvo in informatiko
Ljubljana, Slovenia
jakob.jas06@gmail.com

Matjaž Gams
Department of Intelligent Systems
Jožef Stefan Institute
Ljubljana, Slovenia
matjaz.gams@ijs.si

Abstract

Large language models (LLMs) are being systematically evaluated through accuracy for clinical use, yet privacy risks, limited transparency, and operational variability still complicate their adoption on sensitive health data. Motivated by an intended deployment in HomeDOCtor, a Slovenian medical platform, we present an agenda for evaluating LLMs in real-life privacy-sensitive healthcare applications. First, we map privacy risks: training-data extraction, input leakage, and output re-identification; and outline concrete mitigations (red-teaming, canary strings, differential privacy, filtering, and structured prompts). Second, we propose a lightweight, reproducible evaluation protocol that pairs model-side privacy checks with clinician-in-the-loop utility and safety assessments on de-identified data, aligned with EU GDPR expectations. Third, using small, domain-specific, clinically grounded benchmarks, we compare frontier, commercial, and open-weight models and analyze trade-offs among utility, privacy, and maintainability in the HomeDOCtor context. Finally, we discuss deployment and governance patterns for healthcare operators (access control, audit logging, data minimization, incident response). Our results suggest that (i) focused, task-specific evaluations are more informative than generic world-wide benchmarks for patient-facing use; (ii) suitably hardened and monitored open-weight models can be viable although their quality is not comparable to top commercial systems; and (iii) privacy risk cannot be eliminated but can be bounded and operationalized. We conclude with recommendations for ethics approvals, documentation, and reproducibility to support safe adoption in Slovenia and beyond.

Keywords

Artificial intelligence (AI); Large language models (LLM); Healthcare chatbot; Privacy; GDPR; Open-weight models; GPT; HomeDOCtor; Retrieval-augmented generation (RAG); HealthBench; Humanity's Last Exam; LLM IQ.

*Article Title Footnote needs to be captured as Title Note
†Author Footnote to be captured as Author Note

Information Society 2025, 6–10 October 2025, Ljubljana, Slovenia

https://doi.org/10.70314/is.gptzdravje.5

1 Introduction

Recent evaluation work has shifted from saturated multiple-choice tests toward clinically grounded, contamination-limited settings such as HealthBench, which provides physician-scored multi-turn health dialogues spanning triage safety, clinical appropriateness, and grounding [1, 2]. This shift is critical because theoretical knowledge, often tested in exams, does not guarantee safe or effective application in the nuanced, interactive context of patient care. Ensuring that evaluation benchmarks are not compromised by training data contamination is essential for obtaining a true measure of a model's clinical reasoning abilities. To probe general reasoning under uncertainty beyond strictly medical content, Humanity's Last Exam (HLE) evaluates graduate-level, closed-ended questions and remains far from ceiling performance on the public leaderboard, revealing sizeable headroom [3, 4]. A complementary lens comes from the TrackingAI community's LLM IQ distribution, which aggregates an offline quiz to profile breadth and robustness outside familiar exam sets [5]. Triangulating these different evaluation types (clinical dialogue, academic reasoning, and general IQ) provides a more holistic view of a model's true capabilities.

In the EU, privacy-preserving deployment for patient data is governed primarily by the General Data Protection Regulation (GDPR) [6]. Health data falls under special categories (Article 9), requiring both a valid legal basis (Article 6) and a specific condition under Article 9(2), with principles like data minimisation and purpose limitation being central to system design [6]. While the US Health Insurance Portability and Accountability Act (HIPAA) remains relevant in cross-border collaborations, GDPR is the operative legal framework for Slovenia and most of Europe [6, 7].

As a concrete application context, Slovenia's HomeDOCtor, our nationally localized, RAG-grounded health assistant, provides a real-world test bed for evaluating LLMs under GDPR-first constraints [8]. This system allows for planning a staged migration to locally hosted open-weight models, balancing state-of-the-art performance with stringent data sovereignty requirements [8]. We synthesise official HealthBench results and model cards to compare closed frontier models with competitive open-weight models on clinically oriented tasks [1, 2, 9, 10, 11]. We position these findings alongside HLE and community LLM

IQ scores to characterise remaining reasoning headroom and out-of-distribution robustness [3, 4, 5]. Finally, we integrate a HomeDOCtor case study and provide a GDPR-first deployment blueprint toward zero-egress, on-premise inference with local retrieval, minimising persistent identifiers and aligning with EU data protection obligations [6, 7, 8].

2 Background and Related Work

The development of benchmarks like HealthBench, with its 5,000+ multi-turn conversations scored against physician rubrics, marks a significant maturation in LLM assessment [1]. It moves beyond simple accuracy to measure critical aspects like triage safety, clinical appropriateness, and evidence grounding [1, 2]. Official releases consistently report comparative scores across a range of closed and open-weight models, providing a standardized basis for comparison [2]. To combat the ever-present issue of benchmark contamination, harder alternatives such as LiveBench continually refresh questions and demand verifiable ground truth, mitigating the risk that models simply memorize answers from their training data [12].

Peer-reviewed studies provide further context for model ability on static, image-based medical exams (e.g., USMLE-style questions) [13]. However, these studies also consistently underline that high exam accuracy is not a direct proxy for clinical safety or real-world utility in dynamic, patient-facing deployments [13]. This distinction is vital, as real-world healthcare conversations are rarely as structured as multiple-choice questions.

Classic audits of earlier-generation symptom checkers established a crucial performance baseline, documenting generally low primary diagnostic accuracy and a tendency toward overly risk-averse triage recommendations [14, 15]. Modern LLM-based systems, enhanced with appropriate guardrails and techniques like Retrieval-Augmented Generation (RAG), are expected to significantly surpass this baseline in real-world use cases [14, 15]. Nationally localized assistants like HomeDOCtor have already demonstrated the value of RAG, which grounds model responses in curated, country-specific guidelines and style guides, thereby improving clinical alignment and fostering user trust in live deployments [8].

3 Methods

We aggregate official benchmark reports, model cards, and public leaderboards to assemble a clinically relevant, privacy-aware comparison of leading LLMs. Our methodology is centered on a synthesis of existing, credible data sources to provide a holistic view of model performance.

Specifically, we extract HealthBench and HealthBench-Hard scores from official releases and model documentation where available [1, 2]. These benchmarks are chosen for their clinical relevance and physician-led scoring rubrics [1]. We also include findings from USMLE-style evaluations to provide a broader context of their knowledge on standardized medical exams [13]. We contrast frontier closed models (e.g., GPT-5; o3; GPT-4o) with leading open-weight systems (e.g., GPT-OSS-120B/20B) where credible public results exist [9, 10, 11].

To assess capabilities beyond the medical domain, we incorporate HLE results from the public leaderboard, which reflect general, closed-ended academic reasoning headroom [3, 4]. This benchmark helps characterize a model's ability to reason from first principles on complex, graduate-level topics [3]. We also reference the community-driven LLM IQ distribution from TrackingAI to provide an additional out-of-distribution snapshot of breadth and robustness on a novel offline quiz, designed to resist training data contamination [5]. The triangulation of these benchmarks—one clinical, one academic, one general—is intentional, designed to provide a multi-faceted profile of each model.

To ground these benchmark results in practice, we analyze the HomeDOCtor deployment [8]. In this real-world setting, the core LLM component is swapped while holding the Retrieval-Augmented Generation (RAG) corpus, prompts, and UI/UX constant [8]. This approach effectively isolates the performance deltas attributable to the model itself within a stable, GDPR-first operational environment [8].

4 Results

The collected data reveals a clear performance hierarchy, where frontier models excel on the most complex tasks, but high-quality open-weight models are closing the gap, particularly for routine applications.

Table 1: Summarises HealthBench and HealthBench-Hard scores as reported in official materials.

Model	HealthBench (%)	HealthBench-Hard (%)
GPT-5 (thinking)	67.2	46.2
o3	59.8	31.6
o4-mini	50.1	17.5
o1	41.8	7.9
GPT-4o	32.0	0.0
GPT-OSS 120B	57.6	30.0
GPT-OSS 20B	42.5	10.8

On the hardest, physician-scored subset (HealthBench-Hard), GPT-5 currently leads in official postings with a score of 46.2%, significantly ahead of other models as presented in Table 1 [1, 9]. The leading open-weight model, GPT-OSS-120B, achieves a respectable 30.0%, trailing the frontier but remaining competitive against mid-tier closed models [2, 10]. On the standard HealthBench, these performance gaps narrow further, suggesting that while the most advanced alignment and post-training strategies in frontier systems are key differentiators on challenging dialogues, high-quality open-weight models already cover many routine health tasks effectively when deployed with appropriate guardrails [1, 2].

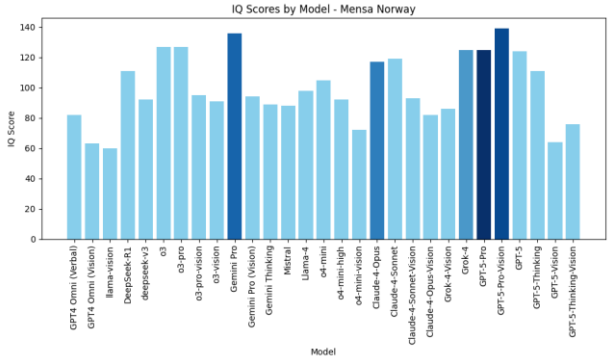
Table 2: Results from Humanity's Last Exam (HLE), which measures closed-ended reasoning across diverse graduate-level topics

Model	HLE score	Uncertainty
GPT-5 (2025-08-07)	25.32	±1.70
Gemini 2.5 Pro Preview (06-05)	21.64	±1.61
o3 (high) (Apr 2025)	20.32	±1.58
GPT-5 mini (2025-08-07)	19.44	±1.55

o4-mini (high) (Apr 2025)	18.08	±1.54
Gemini 2.5 Flash (Apr 2025)	12.08	±1.28
GPT-OSS 120B	9.04	±1.12
o1 (Dec 2024)	7.96	±1.06
GPT-OSS 20B	7.24	±1.05
GPT-4.5 Preview	5.44	±0.89
GPT-4.1	5.40	±0.89
GPT-4o (November 2024)	2.72	±0.64

Table 2 summarises leaderboard entries summarized with central estimates and uncertainty, again place GPT-5 at the top with a score of 25.32 [4, 9]. Notably, the performance of the open-weight GPT-OSS models (9.04 for 120B and 7.24 for 20B) is substantially lower than that of the top closed systems on this general reasoning benchmark [4, 10]. This highlights the significant "reasoning headroom" that still exists and complements the clinical focus of HealthBench by probing for non-medical breadth and analytical depth.

Graph 1: IQ Scores by Model (Mensa Norway, TrackingAI)



Beyond clinical dialogue benchmarks, TrackingAI in collaboration with Mensa Norway provides an independent assessment of general reasoning ability through the LLM IQ test. Unlike standard leaderboards, this offline quiz is carefully designed to resist training-data contamination, thereby capturing model robustness on unfamiliar out-of-distribution problems [5]. Taken together, HealthBench (clinically grounded dialogue), HLE (broad closed-ended reasoning), and the TrackingAI Mensa Norway distribution (community offline quiz, Graph 1) triangulate model capabilities [1, 2, 3, 4, 5]. The consistent pattern is that closed frontier models currently lead on the most difficult and nuanced subsets of tasks. Simultaneously, strong open-weight models such as GPT-OSS-120B have become highly competitive for routine health dialogues and, crucially, enable the on-premise, privacy-first deployments required under regulatory frameworks like GDPR [10].

5 Privacy and Deployment

Legal bases and special categories. For EU deployments, processing health data is strictly regulated [6]. It requires both a valid Article 6 legal basis (e.g., consent, vital interest) and a specific condition under Article 9(2) for special categories of data [6]. Common conditions include medical diagnosis or care, public interest in public health, or explicit consent for specific, clearly defined purposes [6]. The core GDPR principles of data minimisation, purpose limitation, storage limitation, integrity/confidentiality, and accountability must be the primary drivers of the system's design and architecture [6].

Architectural patterns. A zero-egress architecture is the gold standard for privacy, ensuring Personal Health Information (PHI) never leaves an on-premise or sovereign (EU) Virtual Private Cloud (VPC) trust boundary. In this pattern, retrieval-augmented generation (RAG) queries local, audited knowledge stores, and system logs are tightly scoped and automatically rotated with strict retention policies. Any identifiers are filtered, pseudonymized, or transformed before any optional external calls (e.g., for non-clinical functionality), and long-term user profiles are avoided unless explicitly justified by the use case and supported by a Data Protection Impact Assessment (DPIA) [6]. Where collaboration with U.S. partners is necessary, HIPAA concepts can inform mappings of safeguards. GDPR remains the governing regime for legal obligations and data-subject rights in Slovenia and the EU [6, 7].

Controls and assurance. Recommended technical and organizational controls include strict role-based access, end-to-end encryption (in transit and at rest), Data Loss Prevention (DLP) for prompts and outputs, and continuous red-teaming by safety evaluators focused on clinical harms. Governance is maintained through formal DPIAs and detailed records of processing activities for higher-risk use cases, with continuous evaluation on HealthBench-style test sets to monitor for performance drift and ensure referral appropriateness [1, 2, 6].

5.1 Case Study: HomeDOCtor

HomeDOCtor is our implementation of a home doctor medical service that integrates a Flutter front-end, a FastAPI backend, and a Redis Stack vector database that powers the RAG system [8]. The knowledge base is composed of curated Slovenian clinical sources, including the national Manual of Family Medicine, public treatment protocols, official discharge instructions, and the Insieme ontology. During operation, prompts inject the top 3-5 retrieved text snippets into a structured template to generate grounded, locally relevant replies.

Privacy-by-design. To align with GDPR and national constraints, interactions are deliberately stateless and anonymous. No user data are retained beyond the active session, and no longitudinal profiles are created. This design choice maximizes privacy at the cost of convenience (e.g., users must re-enter data each session), but it drastically simplifies regulatory compliance [8].

Model-agnostic orchestration. The architecture is model-agnostic. The same RAG corpus, prompts, and UI can support multiple LLMs (e.g., GPT-4o, 03 mini high, Gemini 2.5, Gemma 3 via Ollama) [8]. This enables direct, like-for-like performance comparisons in a stable pipeline and creates a clear path toward fully local inference on open-weight models using standardized orchestration tools.

Empirical performance. On 100 international clinical vignettes (Avey AI), HomeDOCtor variants using GPT-4o and o3-mini high achieved 99/100 Top-1 accuracy. An open-weight-friendly variant (e.g., using Gemma 3) reached a competitive 95/100 [8]. On a 150-question national internal-medicine test set, HomeDOCtor with GPT-4o scored 136/150, significantly outperforming a baseline of ChatGPT-4o at 121/150 (p=0.0135, Bonferroni-adjusted), demonstrating the power of RAG with local sources.

Operational notes. In a six-month nationwide deployment, the system successfully delivered sub-3-second average responses, provided multilingual support, and garnered positive user feedback. This illustrates the feasibility of providing 24/7 citizen guidance under strict privacy constraints using modern AI architecture.

6 Discussion

In this section, we analyse three overarching themes, beginning with the tension between capability and compliance.

Capability vs. compliance trade-offs. Our findings highlight a central trade-off in applied healthcare AI [1, 2, 6, 9, 10]. Closed, state-of-the-art models retain a performance edge on the most difficult, clinically scored dialogues [1, 9]. However, strong open-weight models are approaching parity on more routine tasks and, critically, enable the fully local, zero-egress inference that is often a decisive factor for PHI-heavy workloads under strict GDPR constraints [2, 6, 10]. The lower recurring costs and greater control offered by self-hosting can also be compelling for public healthcare systems.

Open-weight gap and trajectory. In HealthBench-Hard, the performance gap between a strong open-weight model (GPT-OSS-120B) and the frontier (GPT-5) is on the order of ~16 percentage points [1, 9, 10]. This gap narrows substantially on the broader HealthBench benchmark and in applied, RAG-powered systems like HomeDOCtor, where curated local data can significantly boost performance [1, 2, 8]. This suggests that a key strategy for closing the gap is not just using larger open-weight models, but also investing in high-quality, domain-specific fine-tuning and retrieval augmentation.

Evaluation breadth. HLE and LLM IQ results highlight the residual headroom and robustness variance that exist outside the strictly clinical domain [3, 4, 5]. A model that excels at medical Q&A may still lack the general reasoning capabilities needed for more complex, multi-faceted problems. Therefore, clinical deployments should prioritize systems that are well-grounded, calibrated, and know when to defer to a human expert, rather than extrapolating safety from generic reasoning benchmarks alone [14, 15]. Continuous, post-deployment monitoring against live data is essential to ensure ongoing safety and efficacy.

7 Conclusion

For EU healthcare applications, a GDPR-first architecture is legally essential [6]. In practice, this means local retrieval, zero-egress inference where feasible, tightly scoped, encrypted logging, and explicit, granular consent backed by a DPIA for any data persistence [6]. These guardrails underpin both legal compliance and public trust.

Evidence across HealthBench (clinical dialogue), HLE (broad reasoning), LLM IQ (offline quiz), and our HomeDOCtor deployment shows a consistent pattern: closed models still lead on the most demanding clinical subsets, but mature open-weight systems already support many routine, privacy-preserving workflows when paired with retrieval constraints, auditing, and output filters [1,2,3,4,5,8]. However, it should be noticed that top (say 5) closed systems enable better open communication and reasoning in Slovenian language. Therefore, there is a trade-off between quality and GDPR-compliance between the two groups of systems. Nevertheless, we recommend a staged migration toward model sovereignty, gated by pre-defined safety and performance-parity criteria:

1. pilot zero-egress deployments;
2. move to managed on-prem hosting;
3. advance to fully self-hosted open-weight models once parity (utility, safety, privacy) is demonstrated and continuously monitored [1–15].

This strategy offers a pragmatic path for Slovenia and peers: to deploy self-hosted, sovereign medical AI assistants while upholding the highest standards of data protection and accountability.

At the same time, citizens should have a free choice between the GDPR-dedicated and the commercial top system in medical counselling.

Acknowledgements

We thank medical students Ivana Karasmanakis, Filip Ivanišević, and Lana Jarc for participating in the research. Also, thanks to Rok Smodiš, Matic Zadobovšek, and Domen Sedlar for helping with the development of the HomeDOCtor application. This project is funded by the European Union under Horizon Europe (project ChatMED grant agreement ID: 101159214). The authors also acknowledge the financial support from the Slovenian Research Agency (research core funding No. P2-0209).

References

[1] Arora, R. K., Wei, J., Soskin Hicks, R., Bowman, P., Quiñonero-Candela, J., Tsimpourlas, F., *et al.* HealthBench: Evaluating Large Language Models Towards Improved Human Health. *arXiv* (2025). DOI: 10.48550/arXiv.2505.08775 — URL: https://arxiv.org/abs/2505.08775 arXiv

[2] OpenAI. Introducing HealthBench (May 12, 2025). URL: https://openai.com/index/healthbench/ OpenAI

[3] Phan, L., *et al.* Humanity's Last Exam (HLE). *arXiv* (2025). DOI: 10.48550/arXiv.2501.14249 — URL: https://arxiv.org/abs/2501.14249 arXiv

[4] Humanity's Last Exam. Official site and leaderboard. URLs: https://lastexam.ai/ and https://scale.com/leaderboard/humanitys_last_exam Last ExamScale

[5] TrackingAI.org. LLM IQ – Offline quiz. URL: https://trackingai.org/ Tracking AI

[6] GDPR. Article 9 – Processing of special categories of personal data. URL: https://gdpr-info.eu/art-9-gdpr/ GDPR

[7] U.S. Department of Health & Human Services. Summary of the HIPAA Privacy Rule. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html HHS.gov

[8] Gams, M.; Horvat, T.; Kolar, Ž.; Kocuvan, P.; Mishev, K.; Simjanoska Misheva, M. Evaluating a Nationally Localized AI Chatbot (HomeDOCtor) for Slovenia: Performance, Privacy, and Governance. *Healthcare* 13(15):1843 (2025). DOI: 10.3390/healthcare13151843 — URL: https://www.mdpi.com/2227-9032/13/15/1843

[9] OpenAI. Introducing GPT-5 (Aug 7, 2025). URL: https://openai.com/index/introducing-gpt-5/ OpenAI

[10] Gemma Team (Google DeepMind). Gemma 3 Technical Report. *arXiv* (2025). DOI: 10.48550/arXiv.2503.19786 — URLs: https://arxiv.org/abs/2503.19786 and https://storage.googleapis.com/deepmind-media/gemma/Gemma3Report.pdf arXivGoogle Cloud Storage

[11] Google. Gemini 2.5: Our newest Gemini model with thinking (Mar 25, 2025). URL: https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/ blog.google

[12] White, C., Dooley, S., Roberts, M., Pal, A., Feuer, B., Jain, S., *et al.* LiveBench: A Challenging, Contamination-Limited LLM Benchmark. *arXiv* (2024/2025). DOI: 10.48550/arXiv.2406.19314 — URLs: https://arxiv.org/abs/2406.19314 and https://livebench.ai/ arXivlivebench.ai

[13] Yang, X., *et al.* The performance of ChatGPT on medical image-based assessments and USMLE sample items. *BMC Medical Education* 25, 495 (2025). DOI: 10.1186/s12909-025-07752-0 — URL: https://bmcmededuc.biomedcentral.com/articles/10.1186/s12909-025-07752-0 BioMed Central

[14] Semigran, H. L., Linder, J. A., Gidengil, C., Mehrotra, A. Evaluation of symptom checkers for self diagnosis and triage: audit study. *BMJ* 351:h3480 (2015). DOI: 10.1136/bmj.h3480 — URL: https://www.bmj.com/content/351/bmj.h3480 BMJ

[15] Wallace, W., Chan, A., Chou, R., Desai, S., Johnson, B., Shojania, K. Digital symptom checkers: diagnostic and triage accuracy—systematic review. *NPJ Digital Medicine* 5, 79 (2022). DOI: 10.1038/s41746-022-00667-w — URL: https://www.nature.com/articles/s41746-022-00667-w Nature

IQ Progression of Large Language Models

Evaluating LLM Cognitive Capabilities: An Analysis of Historical Data with Future Projections

Jakob Jaš
Fakulteta za elektrotehniko
Univerza v Ljubljani, Slovenija
jakob.jas06@gmail.com

Matjaž Gams
Oddelek za inteligentne sisteme
Institut "Jožef Stefan", Slovenija
matjaz.gams@ijs.si

Abstract

Over the past few years, artificial intelligence (AI) has advanced rapidly in reasoning and problem-solving. Whereas earlier systems scored well below human averages on standardized benchmarks, recent large language models (LLMs) now match or sometimes exceed the performance of highly capable humans. This paper provides secondary analyses on IQ-style evaluations of leading models across both online (Mensa Norway) and offline test suites, gathered from an external aggregator. The results show a pronounced upward trajectory: models released within the last year frequently score in the top decile of the human distribution, a sharp rise from earlier generations that clustered around the mean. We map model scores to a Gaussian IQ scale to enable direct comparisons with human norms, examine month-over-month trends, and provide short-term projections of likely progress. Findings highlight rapid gains in general-purpose reasoning while underscoring the need for further balanced progress of machine intelligence.

Keywords

artificial intelligence, large language models, IQ, projection

1 Introduction

The past decade has seen a rapid acceleration in artificial intelligence (AI) research and deployment, transforming it from narrow task-specific systems into models capable of exhibiting broad general reasoning. Once limited to specialized domains such as translation and board games, AI systems now demonstrate competencies across multiple modalities, frequently outperforming humans in complex tasks [1].

Large language models (LLMs) have played a central role in this transition. Trained on massive corpora and increasingly multimodal data sources, LLMs have become benchmarks for general-purpose intelligence in machines [2]. Recent work has shown that models such as GPT-4o, Claude 3 Opus, and GPT-5-vision demonstrate reasoning abilities previously unattainable by artificial systems, raising the question of how to compare their progress with human cognitive measures [3,4].

Although domain-specific benchmarks such as MMLU, BigBench, or HELM provide structured evaluation environments [5], they remain primarily task driven. In contrast, IQ-style evaluations, though imperfect, offer a way to frame AI progress in human-familiar psychometric terms [6,7]. The relevance of this framing has grown in 2024–2025, as several independent initiatives (e.g., TrackingAI.org) began publishing standardized IQ-style assessments for frontier AI systems [8].

At the same time, the scientific community has debated whether such comparisons can be justified, given that human IQ tests measure a construct (the g-factor) tied to biological cognition, while AI systems lack embodiment or consciousness [9,10]. Yet, as recent research highlights, behavioural equivalence in reasoning and abstraction can still provide meaningful insights into the trajectory of machine intelligence [11,12,13].

This paper contributes by:

1. Mapping AI model performance on IQ-style benchmarks to the Gaussian human IQ distribution.
2. Analysing month-over-month progress between May 2024 and September 2025.
3. Projecting near-future trajectories of model performance.

By situating these findings in psychometric terms, we aim to provide both a quantitative and conceptual framework for tracking the rapid progression of machine intelligence.

2 Theory and methodology

2.1 Theoretical foundations

The emergence of general-purpose AI models capable of solving novel, cross-domain tasks has prompted a rethinking of how intelligence is defined and measured. Historically, intelligence has been assessed through psychometric methods, with the general intelligence factor (g-factor) introduced by Spearman in 1904 [10]. IQ tests were subsequently developed to capture this construct through tasks spanning verbal, spatial, logical, and mathematical reasoning. Scores are normalized on a Gaussian distribution with mean 100 and standard deviation 15, enabling population-level comparisons [14].

In AI research, traditional evaluation benchmarks have focused on task-specific accuracy, leaving a gap in assessments of general cognitive ability. Recent studies propose adapting psychometric frameworks to AI evaluation, both to contextualize results and to study cross-domain generalization [15,16]. While machines lack consciousness, subjective experience, and embodiment, their problem-solving behaviour can nevertheless be quantified against human reference distributions.

Thus, IQ-style testing is not employed here as a claim of human-equivalent cognition, but as a pragmatic and interpretable method for measuring progress in general reasoning.

[*]Article Title Footnote needs to be captured as Title Note
[†]Author Footnote to be captured as Author Note

Information Society 2025, 6–10 October 2025, Ljubljana, Slovenia

https://doi.org/10.70314/is.2025.gptzdravje.6

2.2 Model selection

The study focuses on leading general-purpose AI systems released between May 2024 and September 2025, ensuring chronological comparability and representativeness of architectural innovation. Models were selected based on three criteria:

- Performance and frontier status – inclusion of systems at or near state-of-the-art benchmarks.
- Architectural diversity – coverage of both text-only LLMs (e.g., LLaMA, Mistral) and multimodal models (e.g., GPT-4o, Claude 3 Opus, GPT-5-vision).
- Data modality shifts – reflecting the move from unimodal to multimodal reasoning [17,18].

This selection enables analysis not only of absolute performance but also of how different architectures and modalities affect reasoning in IQ-like contexts.

2.3 Data source and collection

Performance data were collected from TrackingAI.org, an independent aggregator of psychometrically aligned AI test results [8]. TrackingAI provides transparent, standardized scores across two environments:

- Mensa Norway Online IQ Test – a publicly available timed reasoning test including logic, pattern recognition, and abstract problem-solving [19] (Figure 1).

- Offline IQ-style Test Set – a curated, private benchmark developed to reduce contamination risks from public datasets [20] (Figure 2).

Both test suites normalize results to an IQ-equivalent scale, enabling direct comparison with human distributions.

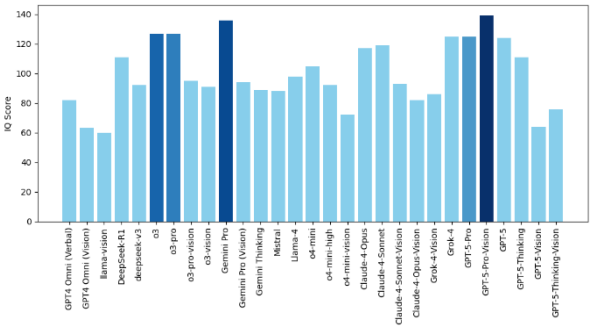


Figure 1: IQ Scores by model - Mensa Norway

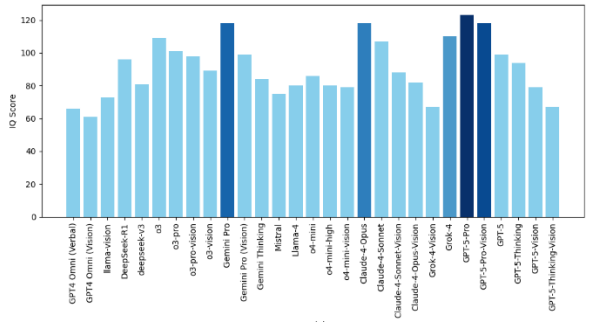


Figure 2: IQ Scores by model - Offline test

2.4 Scoring and Statistical Normalization

Model outputs were scored using the conventional IQ scale (mean = 100, SD = 15). Mensa results ranged 85–145, while offline results spanned ~60–150. Normalization allowed consistent cross-model comparison and alignment with psychometric conventions [21]. Models were ordered chronologically, with top-five performers highlighted to track frontier progression.

Normalization to the human IQ scale can be defined as:

$$z = (X - \mu)/\sigma \qquad IQ = 100 + 15 \cdot z$$

When percentiles are available:

$$IQ = 100 + 15 \cdot \Phi^{-1}(p)$$

Additionally, predictions were made using the jump diffusion model [22, 23] with an adjustable factor e (extremity), which is used to scale all the dynamics of the projection. For all projections, this factor was set to 0.5, resulting in a more conservative estimate. 100 paths were plotted, and the mean path was additionally marked.

3 Results

3.1 Gaussian Distribution Mapping

Figures 3 and 4 illustrate how AI model IQ scores align with the human Gaussian curve. Older systems cluster far left of the mean, corresponding to human IQs between 60 and 80. By contrast, the majority of 2025-era models lie at or above the human average. The distribution shows a clear shift rightward, with leading models positioned well into the 120+ range [24].

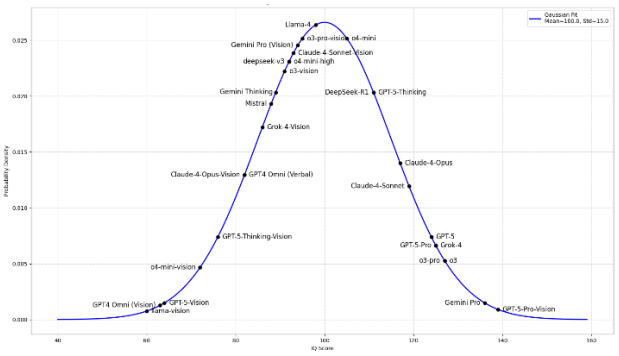


Figure 3: Human-like Gaussian Distribution of Models - Mensa Norway

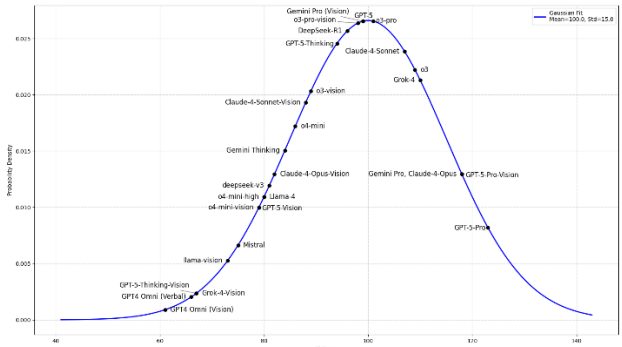


Figure 4: Human-like Gaussian Distribution of Models - Offline Test

3.2 Projected growth

Figure 5 shows monthly IQ-style test scores for top models on Mensa and offline benchmarks between May 2024 and

September 2025, along with linear fits and 12-month projections. Both benchmarks display consistent upward trends over time [25].

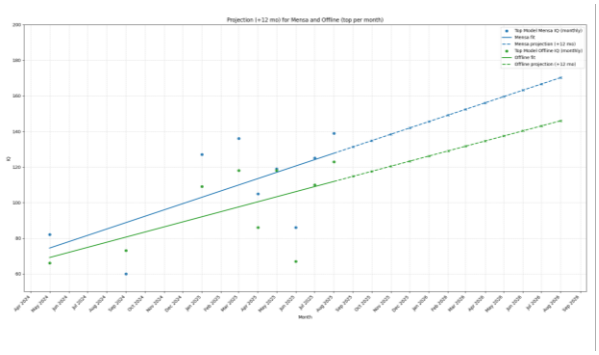


Figure 5: Projected growth based on monthly top-model performance

Mensa scores increased from approximately 80 in May 2024 to around 140 by September 2025, while offline scores rose from about 70 to 125 over the same time period. Linear projections estimate Mensa scores reaching ~170 and offline scores ~145 by mid-2026 [26] (Tables 1,2).

Table 1: Mensa-based projection of improvement

IQ Score	Date	% of people with higher scores
100	Dec.24	50,00%
120	Jun.25	9,12%
140	Nov.25	0,38%
160	May-26	0,003%
170	Sep.26	0,00015%

Table 2: Offline-based projection of improvement

IQ Score	Date	% of people with higher scores
100	Apr.25	50,00%
120	Nov.25	9,12%
140	Jun.26	0,38%
145	Sep.26	0,14%

A jump diffusion model, as seen in Figures 6 and 7, shows the mean projected IQ for Mensa-based data to be ~170 by late 2026 and ~154 for the offline test.

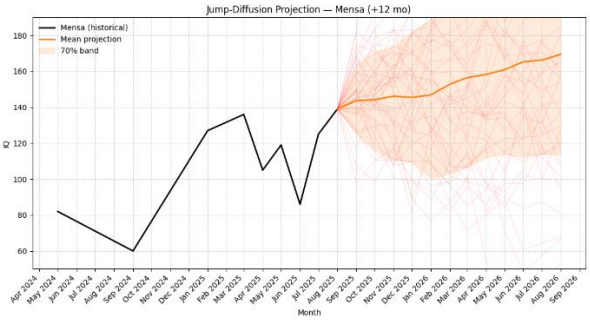


Figure 6: Jump Diffusion Model on Mensa-Based Data

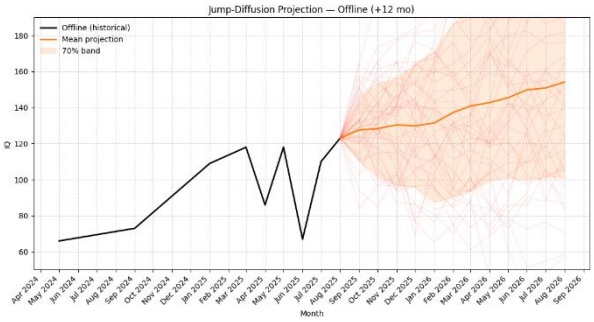


Figure 7: Jump Diffusion Model on Offline-Based Data

4 Discussion

The results demonstrate a clear trajectory of accelerating gains in AI intelligence over the past 12 months, with performance on IQ-style benchmarks increasing at a pace that suggests sustained improvement. Both Mensa-based and offline test results reveal consistent upward trends, though with notable differences. Firstly, Mensa-style evaluations reveal that even earlier-generation models retain relatively strong performance compared to newer systems, contrary to the offline test, where the majority of top-performing models came out very recently. One possible explanation for this is training data contamination [27], as the older models could have been trained on data sets containing information on Mensa's questions, which isn't the case for the offline test, due to its privacy. The rise in the offline test's performance could therefore be attributed to improved model reasoning and overall better model quality. The second notable difference is the rate of growth. The steeper slope of the Mensa evaluations once again indicates that the public nature of the test may be affected by potential training-data contamination, whereas the offline test, being private, seems to show a more robust score.

The Gaussian distribution plots further contextualize these results by positioning current models relative to human intelligence norms. While a majority of systems cluster around human-average IQ levels (90–110), several frontier models now extend significantly into the upper tail of the distribution, with offline IQ equivalents surpassing 120 and projections approaching 145–170 depending on the benchmark [28]. The jump diffusion models additionally support these predictions and even outperform them by nearly 10 IQ points in the offline test case.

This marks a transition from models being predominantly below or near human-level reasoning ability to a subset consistently operating at or beyond the threshold typically associated with high human intelligence [29].

Data from the last 14 months shows that frontier models went from scoring near or even below the human average (GPT-4 Omni, LLaMA-Vision) a year ago, to about average IQ in December 2024 and April 2025 (depending on the administered test), to now reaching the 140 IQ and 125 IQ mark on each test, respectively. Additionally, taking the last six months into account, IQ scores grew by roughly 20 points in both tests [30]. Projections, seen in Tables 1 and 2, thus indicate that by late 2026, models will have surpassed the cognitive abilities of more than 99,87% of all living people based on the more conservative offline estimates, and more than 99,99% based on Mensa data.

Taken together, the findings indicate that AI has not only achieved expert-level performance on various machine benchmarks [31] but is now on a trajectory to surpass human performance across multiple modalities. The pace of this growth, particularly visible in the Mensa projections, raises questions about whether near-future systems may consistently score in ranges associated with the top fraction of human intelligence [32,33].

While the IQ analogy is attractive, due to the seemingly apparent comparisons we can draw between humans and AI, the shortcomings of IQ-based AI evaluation must also be addressed. Firstly, with IQ tests built around human cognition, an AI can, through pattern recognition, perform well on questions without displaying the underlying cognitive flexibility and reasoning skills. Additionally, the IQ test is a contested construct even when it comes to measuring human intelligence, as it may measure some aspects of our cognition, but ultimately falls short when it comes to other skills such as emotional intelligence or creativity [34]. That is why the notion of "AI surpassing human IQ" might be misleading and stems from a false sense of comparability between test scores.

5 Conclusion

The provided data shows evidence of rapid and consistent improvement in model performance between 2024 and 2025. Once positioned below or near the human mean, frontier systems now consistently operate well above the upper decile of the human distribution.

Projections indicate that if current growth trends continue, leading models could reach IQ equivalents in the 145–170 range within the next year, placing them firmly above most human intelligence levels. While methodological uncertainties remain—such as potentially inflated scores due to training data contamination, the opacity of private offline benchmarks, as well as the overall test's validity—the general trajectory is unmistakable: AI systems are advancing at a pace that brings them into direct comparison with high human cognitive performance [35].

These findings highlight not only the acceleration of AI intelligence but also the need for better, machine-oriented evaluation methods. As models continue to expand in scale, modality, and capability, systematic monitoring of their cognitive growth will be essential for understanding both their potential and their societal implications.

Acknowledgements

We thank Tadej Horvat for his help. 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] OpenAI. GPT-5 System Card. Technical Report. 2024. [Online]. Available: https://cdn.openai.com/gpt-5-system-card.pdf

[2] Binz, M., & Schulz, E. (2024). Turning Large Language Models Into Cognitive Models. https://marcelbinz.github.io/imgs/Binz2024Turning.pdf

[3] Xu, Y., et al. (2025). Assessing Executive Function in AI Systems Using Cognitive Benchmarks. Cognitive Computation, 17(1). https://doi.org/10.1007/s12559-025-10200-6

[4] Creswell, A., Shanahan, M., & Kaski, S. (2025). Cognitive Architectures for Multistep Reasoning in LLMs. Journal of Artificial General Intelligence. https://doi.org/10.2478/jagi-2025-0003

[5] Ghosh, A., & Holyoak, K. J. (2025). Analogical Reasoning in Large Language Models: Limits and Potentials. Cognitive Science, 49(2). https://doi.org/10.1111/cogs.13301

[6] Binz, M., & Schulz, E. (2024). Evaluating Planning and Reasoning in Language Models. Nature Machine Intelligence. https://doi.org/10.1038/s42256-024-00896-1

[7] Lake, B. M., Ullman, T. D., & Tenenbaum, J. B. (2024). Symbolic reasoning in the age of deep learning. Annual Review of Psychology. https://doi.org/10.1146/annurev-psych-030322-020111

[8] TrackingAI.org. (2025). IQ-style Benchmark Results. Retrieved from https://trackingai.org

[9] Hernández-Orallo, J. (2017). Evaluation in Artificial Intelligence: From task-oriented to ability-oriented measurement. Artificial Intelligence Review, 48(3), 397–447. https://doi.org/10.1007/s10462-016-9505-7

[10] Spearman, C. (1904). General Intelligence, Objectively Determined and Measured. The American Journal of Psychology, 15(2), 201–293. https://doi.org/10.2307/1412107

[11] Kaller, C. P., Unterrainer, J. M., & Stahl, C. (2012). Assessing planning ability with the Tower of London task. Psychological Assessment, 24(1), 46–53. https://doi.org/10.1037/a0025174

[12] Shallice, T. (1982). Specific impairments of planning. Philosophical Transactions of the Royal Society B, 298(1089), 199–209. https://doi.org/10.1098/rstb.1982.0082

[13] Anthropic. (2024). Claude 3 System Card. Anthropic AI. Retrieved from https://www.anthropic.com

[14] Carroll, J. B. (1993). Human Cognitive Abilities: A Survey of Factor-Analytic Studies. Cambridge University Press.

[15] Xu, Y., et al. (2025). Benchmarking AI Cognition with Psychometric Tests. Cognitive Computation. https://doi.org/10.1007/s12559-025-10200-6

[16] Creswell, A., et al. (2025). Cognitive Benchmarks in LLMs. JAGI. https://doi.org/10.2478/jagi-2025-0003

[17] OpenAI (2025). GPT-5 Vision Technical Report. Retrieved from https://cdn.openai.com/gpt-5-vision.pdf

[18] Mistral AI (2025). Mistral Large System Card. Retrieved from https://mistral.ai

[19] Mensa Norway. (2025). Official IQ Test Description. Retrieved from https://mensa.no

[20] TrackingAI.org. (2025). Offline IQ-Style Dataset Description. https://trackingai.org/offline

[21] Binz, M., Schulz, E., & Lake, B. (2025). Toward Unified Cognitive Testing of AI Systems. Nature Reviews Psychology. https://doi.org/10.1038/s44159-025-00312-4

[22] Merton, R. C. (1976). "Option pricing when underlying stock returns are discontinuous". Journal of Financial Economics. **3** (1–2): 125–144. doi:10.1016/0304-405X(76)90022-2

[23] Grenander, U.; Miller, M.I. (1994). "Representations of Knowledge in Complex Systems"

[24] Zhang, Y., & Marcus, G. (2025). Psychometric Perspectives on AI Evaluation. Frontiers in Artificial Intelligence, 8:155. https://doi.org/10.3389/frai.2025.00155

[25] Bubeck, S., et al. (2024). Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv preprint. https://doi.org/10.48550/arXiv.2303.12712

[26] Chollet, F. (2025). On the Measure of Intelligence Revisited. Journal of Artificial General Intelligence. https://doi.org/10.2478/jagi-2025-0011

[27] Bommasani, R., et al. (2025). The Foundation Model Evaluation Landscape. arXiv preprint. https://doi.org/10.48550/arXiv.2501.01001

[28] Shanahan, M., & Mitchell, M. (2024). Abstraction and Reasoning in AI Systems. Nature Reviews AI, 3, 567–579. https://doi.org/10.1038/s42256-024-00988-8

[29] Hernández-Orallo, J. (2025). Beyond Benchmarks: Toward Psychometric AI. Artificial Intelligence, 325, 104043. https://doi.org/10.1016/j.artint.2025.104043

[30] Binz, M., et al. (2025). Cognitive Scaling Laws in Large Language Models. Nature Machine Intelligence, 7, 445–456. https://doi.org/10.1038/s42256-025-00987-7

[31] Srivastava, A., et al. (2025). Beyond Task Accuracy: A Cognitive Benchmarking Paradigm for LLMs. Proceedings of NeurIPS 2025. https://doi.org/10.5555/neurips2025-12345

[32] Ghosh, A., et al. (2025). Analogical Limits in Transformer Models: Human vs. AI Reasoning. Cognitive Science, 49(3). https://doi.org/10.1111/cogs.13345

[33] Mitchell, M. (2025). The Future of AI Evaluation: Cognitive and Societal Challenges. AI & Society. https://doi.org/10.1007/s00146-025-01789-1

[34] Weiten W (2016). Psychology: Themes and Variations. Cengage Learning. p. 281.

[35] Chollet, F. (2024). Evaluating Progress Toward General Intelligence. Communications of the ACM, 67(12), 54–63. https://doi.org/10.1145/3671234

Extraction of Knowledge Representations for Reasoning from Medical Questionnaires

Emir Mujić*
Alexander Perko
Franz Wotawa
emir.mujic@tugraz.at
alexander.perko@tugraz.at
wotawa@tugraz.at
Graz University of Technology, Institute of Software Engineering and Artificial Intelligence
Graz, Austria

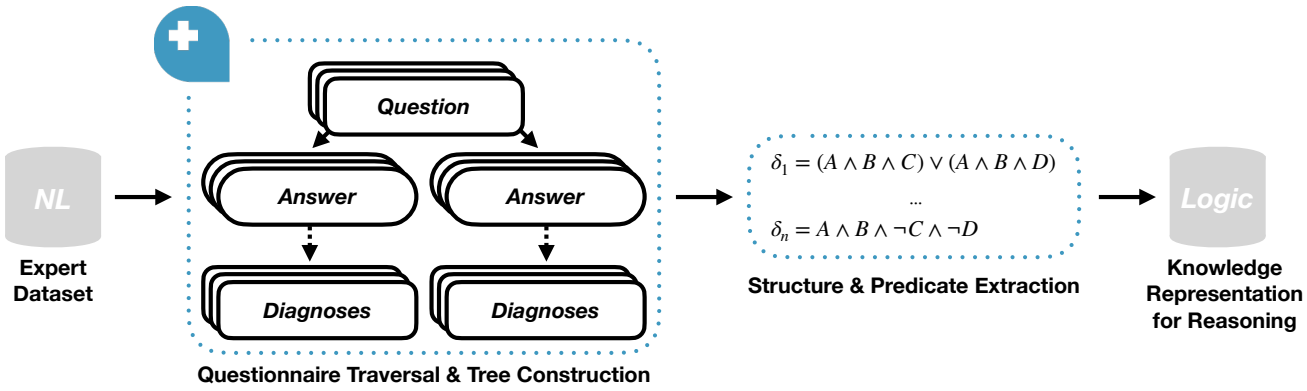


Figure 1: Overview of Knowledge Extraction from the Medical Expert Dataset through Questionnaire Traversal

Abstract

Knowledge representations supporting reasoning are versatile and enable automated use cases such as testing and verification. In contrast to purely data-driven approaches to AI, logical reasoning is explainable. Logic for encoding knowledge yields tremendous potential because of a strong theoretical foundation, and there exist efficient solvers. However, within medicine, we do not find a publicly accessible corpus of expert knowledge encoded in logic. Construction of such a corpus usually requires manual effort and experts in the field, as well as in formal methods. In this work, we contribute by describing a methodology for the automated extraction of logical formulae through interacting with a questionnaire, which is based on a database curated by medical professionals. We propose to use tree traversal and automated predicate extraction from question/answer-nodes comprising natural language. The proposed methods are already established in graph theory, natural language processing, and autoformalization. Hence, we use synergies from different research domains to enable the creation of a logical corpus of medical expert knowledge. With this concept paper, we lay the basis for future work and hope to contribute to use cases, such as rigorous testing of large language models and other medical expert systems.

*Authors are listed in alphabetical order. All authors contributed equally to this research.

Information Society 2025, 6–10 October 2025, Ljubljana, Slovenia

https://doi.org/10.70314/is.2025.gptzdravje.7

Keywords

knowledge representation, reasoning, decision trees, natural language processing, medical questionnaires

1 Introduction & Related Work

Logical formalisms, like First-Order Logic (FOL), or the Answer Set Programming paradigm (ASP) [12], can be used to encode knowledge enabling reasoning through theorem provers/solvers, such as Prover9 [15] or Clingo [11]. Having a logical knowledge base, one can easily query existing facts, check statements for consistency, and infer new knowledge. Consider now a medical knowledge base KB, where symptoms are mapped to diagnoses such that one can infer a set of diagnoses given a set of facts about a person, and a set of symptoms. Given a proper user interface, this can be directly used as an expert system. What is more, it can be used as a test oracle for comparisons with other medical expert systems, providing a transparent view of how diagnoses are made. Even more interestingly, we can evaluate large language models (LLMs) tasked with diagnosing a person given the same input, which we already demonstrated in earlier work [20, 19]. Although there exist benchmarks & datasets for question answering [14] and natural language inference [22] in medicine, we do not find a dataset that fulfills the described properties and is publicly available. Hence, our goal is to build such a knowledge base. As manually creating a gold standard dataset requires expert knowledge and is costly, we propose the automated enrichment of an existing database, which can be accessed through a questionnaire. More specifically, we show how to extract logical formulae from NetDoktor's „Symptom-Checker" questionnaire (SCQ) [21], which is curated by medical professionals and is based on the AMBOSS dataset [1]. Our methodology aims for

automated formalization, i.e., autoformalization of knowledge encoded in natural language. Furthermore, we contribute by elaborating how to leverage the fact that tree representations can be converted into logical formulae [2]. Vice versa, tree structures can be created from logical sentences [5]. A benefit of having a decision tree from a knowledge base is being able to exactly compute bias in the diagnoses (and the knowledge base), as well as the sufficient and necessary reasons behind decisions [9, 2], even in cases of trees with non-binary features (multiple choice questions) [13]. That said, this work directly builds upon our earlier work [20], where we outline the concept of representing a medical questionnaire as a decision tree.

At this point, it makes sense to introduce medical questionnaires & similar systems, such as chatbots: The main idea is to provide answers to a user given symptomatic and/or other information about a person. They are used by the general public and medical professionals alike, and their application varies from general health assessment, over risk calculators to medical triage [16]. These systems often use different combinations of rule-based and data-driven approaches [3, 7]. Most recently, general purpose, as well as domain-specific LLMs, are heavily utilized as well [17, 23, 6], which increases the demand for testing them. We argue that it makes sense to rely on an evaluation methodology that is fully understandable, deterministic, and finite to test non-deterministic, black box systems, such as LLMs. You can find a pilot evaluation of ChatGPT [18] using SCQ in our earlier work [20]. This brings us back to medical questionnaires in the classical sense, from which we will extract a logical knowledge base. Questions within a medical questionnaire can be distinguished in several ways. Namely, we distinguish by:

- Question format:
  - Open-ended questions (Type 1).
  - Closed-ended questions (Type 2).
- Fact permanence:
  - Questions about what a person *is*, which yield permanent facts about a person.
  - Questions about what a person *has*, which yield temporary facts about a person, i.e. symptoms.
- Question requirement:
  - Obligatory questions.
  - Optional questions, with an option to skip.
- Answer types:
  - Predefined options to answer.
  - Freeform answers (not present in SCQ).

Note that these categories are mutually exclusive within but not across distinguishable dimensions, e.g., in principle, it is possible to either have obligatory or optional questions that are open-ended, as well as closed-ended. Having introduced the general problem and domain, we will now proceed with describing a methodology for the enrichment of an expert dataset, with logical representations through tree traversal & basic semantic parsing.

2 Methodology

This work aims to automatically extract logical formulae from knowledge encoded in structured, natural language. Thus, there are three parts to the proposed methodology:

(1) Construction of the tree structure, through filling out SCQ.
(2) Extraction of predicate names from natural language.
(3) Aggregation of formulae, through tree traversal.

While our methods are universally applicable to extracting knowledge from any questionnaire of a similar form, we base all elaborations on SCQ.

2.1 Tree Representations of Questionnaires

In this work, we represent medical questionnaires as decision trees. We first look at creating a simple tree T from SCQ, which corresponds to a session a user might have with the tool:

The root node $r(T)$ is always a question with which every new session is started: *Um wen geht es?* (Who is this about?). From this root node $r(T)$, the tree branches down in a depth-first manner, starting with obligatory questions of Type 1, and followed by optional Type 2 questions. The leaf node(s) $l(T)$ represent a set Δ of diagnoses proposed by SCQ.

Given a tree T with a root node $r(T)$, any number of regular nodes $n_i(T), i = 1, ..., N-1$ and leaf nodes $l(T)$, a walk[1] [10] defines a "Tree Path Structure" from the root to any other node, including the leaf node i.e. the diagnosis possible within the system. Since we know that we can treat trees as graphical representations of logical formulae in disjunctive normal form (DNF), we can write that any tree path structure represents a world w that satisfies at least one diagnosis δ, $w \models \delta$. In other words, models of any diagnosis δ , $Mods(\delta)$ is *any* set of variable assignments that lead to that diagnosis. In most cases, there will be more than one diagnosis given for a world w, we denote this as $w \models \Delta, \delta \in \Delta$, where Δ is a subset of all possible diagnoses, $\Delta \subseteq \mathcal{D}$ [2]. The set of all diagnoses $\mathcal{D}$ is satisfied by the union of worlds of all diagnoses: $Mods(\mathcal{D}) = \bigcup_{j=0}^{M} w_j$, where M is the number of possible diagnoses.

We show a simple example: A diagnosis δ_1 (acute gastroenteritis) is given as a result if a patient has nausea (A) and stomach ache (B) and either fever (C) or diarrhea (D). Another diagnosis δ_2 (gastritis) is a result if a patient has nausea (A) and stomach ache (B) without fever ($\neg C$) and diarrhea ($\neg D$). We can write this as a set of formulae in DNF as:

$$\delta_1 = (A \wedge B \wedge C) \vee (A \wedge B \wedge D), \quad (1)$$
$$\delta_2 = A \wedge B \wedge \neg C \wedge \neg D,$$

which we can represent as a decision tree shown in Figure 2.

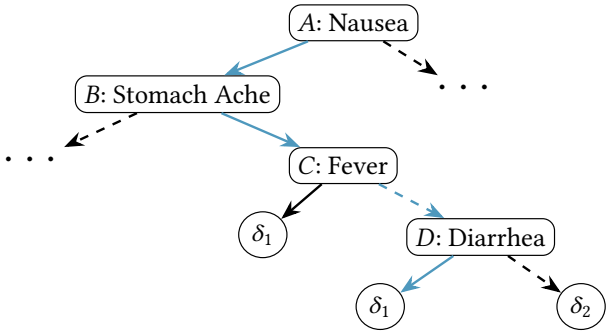


Figure 2: Example 1 as Decision Tree

In Figure 2, a full edge between any two variables represents a truth assignment to the upper variable in the tree based on which the lower variable follows. The dashed edge between represents

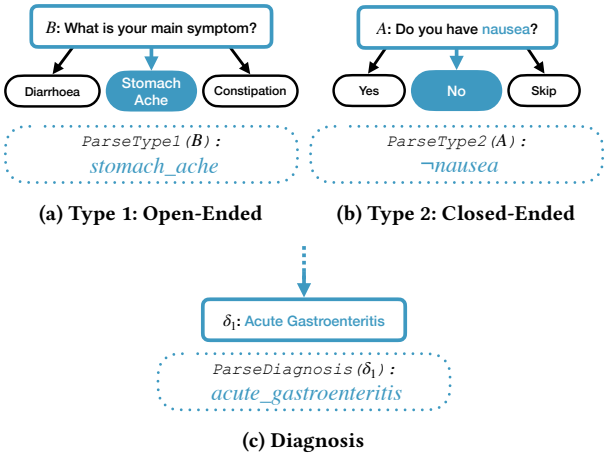
[1] A walk in this context refers to its graph-theoretical definition: In a graph $(V, E) : G, E \subseteq [V]^2$, a walk is a sequence $v_0 e_1 v_2, ..., e_{n-1} v_n$ of alternating vertices and edges such that $\forall_i : e_i = \{v_{i-1}, v_i\}$.
[2] In general: $\Delta \subset \mathcal{D}$. However, $\Delta \subseteq \mathcal{D}$ *iff* $w = \{\emptyset\}$.

a false assignment to the upper variable from which the lower variable follows. The walk highlighted in blue represents one possible instantiation of symptoms where the patient has nausea, a stomach ache, and diarrhea without fever. The three dots ("···") in Figure 2 denote that there are parts of the tree not shown in the example but may exist in the complete tree representation. We would also like to point out that there may exist multiple walks to any single node in the tree, including the leaf nodes ($w_i, w_j \models \Delta, w_i \neq w_i, i \neq j$), something that is excluded in the example in Figure 2 for clarity.

Finally, we summarize how to extract a complete tree out of SCQ, following a depth-first-search methodology: Opening the first session with the questionnaire corresponds to creating a root node. This is followed by answering questions systematically, remembering all questions and answers, and adding corresponding nodes to the tree. At the end of one session, we are presented with a set of diagnoses, which represent the leaf nodes in the tree. This procedure is repeated until we have traversed the entire search space. For further explanations, we refer the interested reader to our previous work [20], which provides elaborations on SCQ, and extracted tree nodes. Due to space limitations of visually representing large trees, we provide examples separately, which can be downloaded at Zenodo[3].

2.2 Predicate Extraction

For now, we assumed the nodes of the constructed tree representation to be directly usable as predicates. However, as the nodes correspond to statements (e.g., sentences, words, or noun phrases) in natural language (NL), we first have to extract predicates. Moreover, in order to enable more than two answers per question, we extend the simplified tree structure from above by the inclusion of separate answer nodes. Thus, we have three types of NL nodes: Questions, corresponding answers, and diagnoses. Furthermore, we assume to remember the relation of questions to their answers and a basic classification of question types into "Type 1", i.e., open-ended, and "Type 2", i.e., closed-ended questions. This distinction can also be seen in Figure 3.



(a) Type 1: Open-Ended (b) Type 2: Closed-Ended

(c) Diagnosis

Figure 3: Predicate Extraction through Parsing Functions for Different Question Types, & Diagnoses

We define three node-level parsing functions: 1) `ParseType1`, 2) `ParseType2`, and 3) `ParseDiagnosis`, which are explained visually in Figure 3. We can simplify the step of autoformalization, as the NL statements found in SCQ show a very limited linguistic complexity. Therefore, we propose to either use naive semantic parsing or LLM-based predicate extraction. For the naive approach, one would simply return the object of a sentence (i.e., singular word or whole noun phrase), modified for the formal language in question. ASP, as used in Clingo, for instance, demands predicates to be written in lower case and allows underscores for separating words in predicate names, which can be seen in Figure 3. Table 1 shows further examples for predicate extractions.

[3]https://doi.org/10.5281/zenodo.17058631

2.3 Formula Aggregation

Continuing with the aggregation of the extracted predicates into logical formulae, we propose a simple algorithm, which can be seen in Algorithm 1. The input is the (extended) tree T, or rather its root node $r(T)$, and the output is a list of formulae, corresponding to all paths in the tree, each comprising a persona and its symptomatic (which we subsume by "symptoms"), as well as corresponding diagnoses.

Algorithm 1 SCQ Tree Traversal for Formula Aggregation

Input: Root node $r(T)$ (assumed to be the first question)
Output: A list of all paths, corresponding to formulae:
(i) a list of symptoms, and
(ii) a list of diagnoses.

```
1:  function TreeTraversal(r(T))
2:      Formulae ← [ ]          ▷ Final list of aggregated formulae
3:      Visit(r(T), [ ], [ ], Formulae)
4:      return Formulae
5:  end function
6:  function Visit(node, Symptoms, Diagnoses, Formulae)
7:      if node.type = "Leaf Node" then
8:          NewPredicates ← ParseDiagnosis(node)
9:          Diagnoses ← Diagnoses ∪ {NewPredicates}
10:         append (Symptoms, Diagnoses) to Formulae
11:         return
12:     end if
13:     if node.type = "Question" then
14:         for each child in node.children do
15:             if node.subtype = "Type1" then
16:                 NewPredicates ← ParseType1(child)
17:             else if node.subtype = "Type2" then
18:                 NewPredicates ← ParseType2(node, child)
19:             end if
20:             Symptoms ← Symptoms ∪ {NewPredicates}
21:             Visit(child, Symptoms, Diagnoses, Formulae)
22:         end for
23:     end if
24: end function
```

As can be seen in Lines 1-5 of Algorithm 1, the depth-first search is started by calling the TreeTraversal function with $r(T)$. Next, a Visit function (Lines 6-24) is called recursively, visiting all nodes on a path until it reaches the/each leaf node (Line 7). In the final list of formulae, which represents all paths, symptoms are assumed to be conjunctions whereas diagnoses are assumed to be disjunctions. Both comprise parsed predicates, and can now be joined to form strings, depending on the target formalism and solver/theorem prover.

ID	Type	Tree Node		Predicate
		Question	Selected Answer	
1	1	Geht es um eine Frau oder einen Mann? *Is it about a woman or a man?*	Weiblich *Female*	female
2	1	Wo treten die Beschwerden auf? *Where do the symptoms occur?*	Kopf *Head*	head
3	1	Wähle dein wichtigstes Symptom *Select your most important symptom*	Schnarchen *Snoring*	snoring
4	2	Leidet die Person unter Schnupfen oder laufender Nase? *Does the person have a cold or runny nose?*	Ja *Yes*	cold ∨ runny_nose
5	2	Ist die Haut (stellenweise) gerötet? *Is the skin reddened (in places)?*	Nein *No*	– reddened_skin
6	2	Hattest du schon einmal eine Allergie? *Have you ever had an allergy?*	Überspringen *Skip*	×

Table 1: Exemplary Predicates by ID, Extracted from Question- & Answer-Tree-Nodes. For Type 1 questions, predicates are extracted from answers. Type 2 questions yield predicates directly, while (potential) negations are extracted from answers.

3 Conclusion & Future Work

In summary, we propose a methodology for constructing & traversing trees from medical questionnaires for the extraction of logical formulae. We describe how to leverage this to construct a medical knowledge base, which can be used for reasoning and enables future work, such as testing LLMs. Future work on decision trees extracted from medical questionnaires will include dealing with multiple paths to the same diagnosis, the intersection of structured tree paths, redundant trees, as well as transforming the large trees into different structures that allow for more efficient computation of certain properties. These include ordered binary decision diagrams [4] and deterministic decomposable negation normal form (d-DNNF) circuits [8], offering the possibility of model counting (asking what diagnoses are possible for any subset of symptoms), reasoning about the biases in the knowledge base by analyzing the decisions made, giving us a complete reason behind diagnoses from which we can compute the sufficient reason (the reason why that diagnosis was chosen) and the necessary reason (why any other diagnosis was not chosen) [9, 13, 2]. With these analyses, we hope to gain further insights into the knowledge base of SCQ and find new and interesting ways of using its logically enriched form. Ultimately we hope to enable new testing strategies of AI-based systems in medicine, particularly LLMs.

Acknowledgements

The work presented in this paper was partially funded by the European Union under Grant 101159214 – ChatMED. 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.

References

[1] AMBOSS GmbH. 2025. Amboss. www.amboss.com. Accessed: 2025-09-03. (2025).

[2] Gilles Audemard, Steve Bellart, Louenas Bounia, Frédéric Koriche, Jean-Marie Lagniez, and Pierre Marquis. 2021. On the explanatory power of decision trees. *arXiv preprint arXiv:2108.05266*.

[3] Ahmad Taher Azar and Shereen M El-Metwally. 2013. Decision tree classifiers for automated medical diagnosis. *Neural Computing and Applications*, 23, 7, 2387–2403.

[4] Randal E Bryant. 1992. Symbolic boolean manipulation with ordered binary-decision diagrams. *ACM Computing Surveys (CSUR)*, 24, 3, 293–318.

[5] Chin-Liang Chang and Richard Char-Tung Lee. 1973. *Symbolic logic and mechanical theorem proving*. Academic press.

[6] Zeming Chen et al. 2023. Meditron-70b: scaling medical pretraining for large language models. (2023). eprint: 2311.16079.

[7] Dillon Chrimes. 2023. Using decision trees as an expert system for clinical decision support for covid-19. *Interact J Med Res*, 12, (Jan. 2023), e42540. DOI: 10.2196/42540.

[8] Adnan Darwiche. 2001. Decomposable negation normal form. *Journal of the ACM (JACM)*, 48, 4, 608–647.

[9] Adnan Darwiche and Auguste Hirth. 2023. On the (complete) reasons behind decisions. *Journal of Logic, Language and Information*, 32, 1, 63–88.

[10] Reinhard Diestel. 2025. *Graph theory*. Vol. 173. Springer Nature.

[11] Martin Gebser, Roland Kaminski, Benjamin Kaufmann, and Torsten Schaub. 2018. Multi-shot ASP solving with clingo. (Mar. 2018). arXiv: 1705.09811 [cs]. DOI: 10.48550/arXiv.1705.09811.

[12] Michael Gelfond and Vladimir Lifschitz. 1988. The stable model semantics for logic programming. In *Proceedings International Logic Programming Conference and Symposium*. MIT Press, Cambridge, MA, USA, 1070–1080.

[13] Chunxi Ji and Adnan Darwiche. 2023. A new class of explanations. In *Logics in Artificial Intelligence: 18th European Conference, JELIA 2023, Dresden, Germany, September 20–22, 2023, Proceedings*. Vol. 14281. Springer Nature, 106.

[14] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*.

[15] W. McCune. 2005–2010. Prover9 and Mace4. (2005–2010).

[16] Bilal A Naved and Yuan Luo. 2024. Contrasting rule and machine learning based digital self triage systems in the usa. *NPJ digital medicine*, 7, 1, 381.

[17] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. (2023). https://arxiv.org/abs/2303.13375 arXiv: 2303.13375 [cs.CL].

[18] OpenAI. 2023. ChatGPT. (2023). chat.openai.com/chat.

[19] 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*. Ljubljana, Slovenia. DOI: 10.70314/is.2024.chtm.10.

[20] Alexander Perko and Franz Wotawa. 2024. Testing ChatGPT's Performance on Medical Diagnostic Tasks. In *Proceedings of the 27th International Multiconference Information Society – IS 2024*. Ljubljana, Slovenia. DOI: 10.70314/is.2024.chtm.7.

[21] Jens Richter, Hans-Richard Demel, Florian Tiefenböck, Luise Heine, and Martina Feichter. 2025. Symptom-checker. www.netdoktor.at/symptom-checker/. Accessed: 2025-09-03. (2025).

[22] Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. *arXiv:1808.06752 [cs]*, (Aug. 21, 2018). Retrieved Aug. 27, 2018 from arXiv: 1808.06752.

[23] Khaled Saab et al. 2024. Capabilities of gemini models in medicine. (2024). https://arxiv.org/abs/2404.18416 arXiv: 2404.18416 [cs.AI].

Indeks avtorjev / Author index