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.

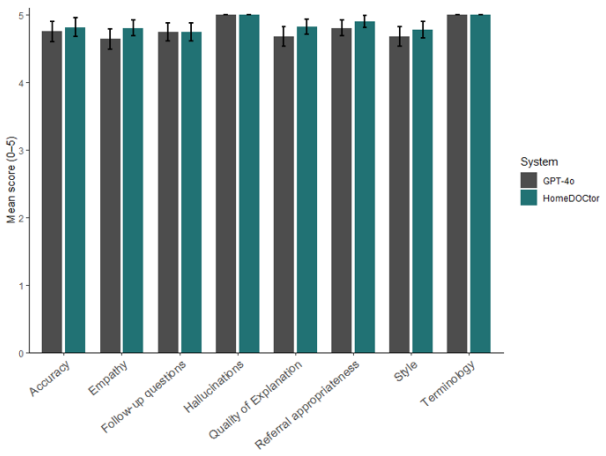

Figure 1: Dimension-level mean scores with 95% CI for GPT-4o vs. HomeDOCtor.

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.

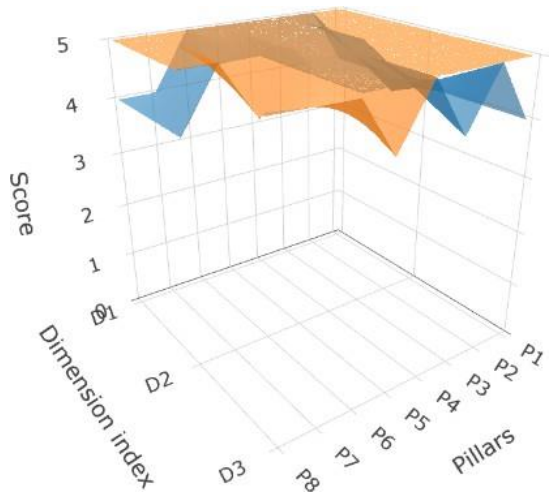

Figure 2: Comparison of GPT-4o and HomeDOCtor through the M-LEAF framework.

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.

Acknowledgements

This project is funded by the European Union under Horizon Europe (project ChatMED grant agreement ID: 101159214).

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.