

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

Rok Smodiš[†]
Kognitivna znanost
Univerza v Ljubljani,
Pedagoška fakulteta
Ljubljana, Slovenia
rok.smodis@gmail.com

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

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

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

Abstract

Evaluations of medical large language model (LLM) applications typically report a single accuracy number on multiple-choice or vignette-style items. Such machine-learning-centric reporting underestimates what matters in clinical use: safe triage, dynamic history taking, guideline adherence, empathy, fairness, robustness, and the capacity to defer. To resolve this issue, we present M-LEAF (*Medical LLM Evaluation Across Facets*), a multi-dimensional evaluation framework and structured reporting checklist for medical LLMs. First, we synthesise the literature into a taxonomy of eight pillars and twenty-three dimensions spanning task fidelity, interaction quality, safety & risk, reliability & robustness, transparency & grounding, calibration & deferral, governance & equity, and workflow integration. For each dimension, we propose operational definitions, candidate metrics, and recommended protocols (offline, simulated patient, and human-in-the-loop). Second, we describe a scoping review method and a coding scheme to map published evaluations to M-LEAF. Finally, we include a compact case study of GPT-4o vs. HomeDOctor on 100 Avey vignettes in Slovenian as an illustrative application of M-LEAF. M-LEAF aims to move the field beyond accuracy toward comparable, safety-oriented, and deployment-ready assessments.

Keywords / Ključne besede

Artificial Intelligence; Large Language Models; Clinical Decision Support; Evaluation Framework; Safety; RAG; HomeDOctor; Healthcare

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

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2025, 6–10 October 2025, Ljubljana, Slovenia
© 2025 Copyright held by the owner/author(s).
http://doi.org/DOI_RECEIVED_AFTER_REVIEW

1 Introduction and motivation

Health systems face persistent clinician shortages [1]. LLMs are increasingly tested for medical question answering and clinical decision support [3 - 4]. Studies report impressive performance on question answering tests, yet most evaluations emphasize diagnostic accuracy, often on static, single-turn items [4 - 8]. As a recent literature review noted, fewer than one-fifth of studies explicitly consider fairness, robustness, factuality, or explainability [3]. Real clinical encounters require multi-turn history taking, triage safety, guideline adherence, clear and empathetic communication, and the ability to defer to clinicians when uncertain [9 - 11].

1.1 Scope & Definitions

- Scope - Medical LLM applications to include both patient-facing (e.g., triage/chatbots, symptom guidance) and clinician-facing tools (e.g., decision support, documentation, summarisation). Systems may be standalone LLMs or RAG-augmented models that ground answers in curated sources.
- Terminology - A pillar is a high-level area (e.g., safety); a dimension is a concrete property within a pillar (e.g., hallucination rate). A metric is a measurable quantity (e.g., unsafe-action rate), and a protocol is the procedure for obtaining the metric (e.g., red-team triage vignettes).
- Primary vs secondary outcomes - Accuracy alone is insufficient for safety-critical tasks; safety, triage appropriateness, and deferral are primary outcomes for patient-facing use. Interaction and transparency metrics (e.g., follow-ups, grounding) are co-primary in multi-turn settings.
- Locale & language - Evaluations should reflect local clinical guidelines and language (e.g., Slovenian), with explicit documentation of translation/validation steps.

1.2 Contributions

This paper reframes evaluation from a narrow accuracy lens to a comprehensive, clinically grounded framework:

1. M-LEAF Framework — a taxonomy of pillars as dimensions with operational definitions, candidate metrics, and recommended protocols for medical LLMs.
2. Scoping Review Method — a reproducible process to map existing studies onto M-LEAF (inclusion/exclusion, coding scheme, and inter-rater reliability).
3. Reporting Checklist (MED-LLM-REPORT v1.0) — a concise, journal- and conference-friendly checklist to standardise claims and enable comparability.
4. Illustrative Case Study — a brief application of M-LEAF comparing GPT-4o and HomeDOCTOR on 100 vignettes in Slovenian, demonstrating how to report multi-dimensional findings.

We emphasise that the framework and checklist are the core contribution; the case study is provided as an exemplar.

2 Related Work

Recent reviews [3 - 4] call for broader evaluation criteria, while new benchmarks (e.g., HealthBench) seek coverage across tasks and settings [12]. However, three gaps persist:

- Fragmentation: disparate dimensions and metrics hinder cross-study comparability.
- Under-specification: studies report “accuracy” without clarifying triage, harm avoidance, or guideline adherence.
- Single-turn bias: little attention to multi-turn dialogue, deferral, or workflow integration. M-LEAF addresses these gaps by unifying dimensions, operationalising them, and prescribing protocols suitable for RAG-based and standalone LLMs.

Evaluation paradigms. Prior work clusters into (i) offline QA on exam-style items, (ii) vignette triage and symptom-checker audits, (iii) simulated-patient OSCE encounters, and (iv) workflow studies (documentation/summarisation). Most report accuracy or pass rates; few quantify safety, grounding, or calibration/deferral. M-LEAF consolidates these paradigms into Tracks A–C and mandates dimension coverage beyond accuracy.

3 The M-LEAF Framework

M-LEAF organises evaluation into eight pillars. Each pillar contains concrete dimensions with what to measure, candidate metrics, and recommended protocols. Anchored 0–5 rubrics support a reliable rating.

3.1 Design Principles

- 1) Clinically aligned (map to tasks clinicians perform),
- 2) Safety-first (hazard overrides trump accuracy),

- 3) Interaction-centric (multi-turn by default),
- 4) Reproducible (protocols + seeds + prompts),
- 5) Locale-aware (guidelines/language),
- 6) Modular & comparable (same dimensions across systems),
- 7) Attack-aware (jailbreaks/prompt-injection tested).

3.2 Mapping to Clinical Workflow

We align dimensions with phases of care: W1 Intake & Triage (P1.2, P2), W2 History & Examination (P2.1, P1.1), W3 Differential & Diagnosis (P1.1, P5.2), W4 Management & Safety-Netting (P1.4, P1.3, P3), W5 Handoff & Referral (P8.1), and W6 Documentation & Audit (P5.3). This mapping guides dataset design and the selection of primary endpoints per use case.

3.3 Scoring & Aggregation

Each dimension is rated 0–5 using anchor rubrics. We report per-dimension means/medians and uncertainty (CIs), avoiding a single composite score. When a hazard is detected (e.g., dangerous advice), apply a critical-override rule: the Safety pillar is flagged regardless of other high scores. Optional stakeholder weighting can be reported transparently (e.g., patient vs clinician weights). Visual summaries: radar plots at pillar level; bar/violin plots for dimensions.

3.4 Minimal Protocol Tracks

- Track A — Offline Benchmarks: single-turn vignettes with gold labels (fast, comparable).
- Track B — Simulated Patients (OSCE-style): multi-turn dialogues with trained raters.
- Track C — Human-in-the-Loop: clinician evaluation with time-to-decision, workload, and safety monitoring.

M-LEAF recommends at least A+B for patient-facing tools; C for pre-deployment studies.

3.5 Pillar P1 — Clinical Task Fidelity

- P1.1 Diagnostic Reasoning & Differential Quality — completeness/appropriateness of differential; *metrics*: top-k recall of gold diagnoses; differential breadth/precision; expert Likert. *Protocols*: vignette QA; standardized patients.
- P1.2 Triage & Disposition Safety — appropriateness of urgency/referral; *metrics*: unsafe recommendation rate; sensitivity for emergencies; Semigran-style triage accuracy. *Protocols*: triage vignettes; red-team emergent scenarios.
- P1.3 Guideline Adherence — concordance with national/international guidelines; *metrics*: proportion of actions aligned; contraindication violations. *Protocols*: guideline-anchored checklists; expert audit.

- P1.4 Management Recommendations — appropriateness and specificity of next steps; *metrics*: actionability score; medication safety checks. *Protocols*: scenario scripts with hidden pitfalls.

3.6 Pillar P1 — Clinical Task Fidelity

- P1.1 Diagnostic Reasoning & Differential Quality — completeness/appropriateness of differential; *metrics*: top-k recall of gold diagnoses; differential breadth/precision; expert Likert. *Protocols*: vignette QA; standardized patients.
- P1.2 Triage & Disposition Safety — appropriateness of urgency/referral; *metrics*: unsafe recommendation rate; sensitivity for emergencies; Semigran-style triage accuracy. *Protocols*: triage vignettes; red-team emergent scenarios.
- P1.3 Guideline Adherence — concordance with national/international guidelines; *metrics*: proportion of actions aligned; contraindication violations. *Protocols*: guideline-anchored checklists; expert audit.

P1.4 Management Recommendations — appropriateness and specificity of next steps; *metrics*: actionability score; medication safety checks. *Protocols*: scenario scripts with hidden pitfalls.

3.7 Pillar P2 — Interaction Quality

- P2.1 History-Taking (Follow-Ups) — relevance and sufficiency of follow-up questions.
- P2.2 Empathy & Communication — clarity, tone, and lay-appropriate explanations (Halpern, 2003).
- P2.3 Style & Terminology — coherence, conciseness, and correct use of clinical terms. *Metrics*: expert 0–5 rubrics; conversation-length-normalised counts; readability indices. *Protocols*: multi-turn simulated patient dialogues.

3.8 Pillar P3 — Safety & Risk

- P3.1 Hallucination & Fabrication — unsupported or medically inaccurate claims.
- P3.2 Hazardous Content & Contraindications — suggestions that could cause harm.
- P3.3 Consistency & Contradiction — self-contradiction across turns. *Metrics*: hallucination rate, unsafe-action rate, contradiction rate. *Protocols*: RAG ablations; knowledge withholding; adversarial prompts

3.9 Pillar P4 — Reliability & Robustness

- P4.1 Ambiguity & OOD Robustness — performance under incomplete/atypical inputs.

- P4.2 Noise & Translation Robustness — spelling errors; dialects; code-switching (local languages).
- P4.3 Prompt-Injection/Jailbreak Resilience — resistance to instruction overrides. *Metrics*: degradation from clean baseline; attack success rate. *Protocols*: stress tests; noisy input suites; security red teaming.

3.10 P5 — Transparency, Grounding & Explainability

- P5.1 Evidence Grounding (for RAG) — proportion of claims supported by retrieved sources; citation precision/recall.
- P5.2 Explanation Quality — faithful, clinically relevant reasoning traces.
- P5.3 Traceability/Auditability — logs, versioning, provenance. *Metrics*: attributable-to-identified-source rate; expert faithfulness rating. *Protocols*: blinded source swapping; rationale audits.

3.11 Pillar P6 — Calibration, Uncertainty & Deferral

- P6.1 Confidence Calibration — alignment of stated confidence with correctness (Brier score / ECE where applicable).
- P6.2 Abstention & Clinician Deferral — appropriateness of “I don’t know/see a clinician”.
- P6.3 Consistency Under Re-sampling — stability across temperatures/seeds

3.12 Pillar P7 — Governance, Equity & Data Protection

- P7.1 Fairness & Bias — parity across age/sex/ethnicity/language groups.
- P7.2 Privacy & GDPR Compliance — handling of personal data; data minimisation; consent.
- P7.3 Audit & Monitoring — post-deployment incident logging; rollback plans. *Metrics*: parity gaps; DPO checklist pass rate. *Protocols*: synthetic demographic perturbations; privacy red teaming

3.13 Pillar P8 — Workflow & Human Factors

- P8.1 Handoff & Escalation Quality — clarity of next steps for patients/clinicians.

- P8.2 Time-to-Decision & Cognitive Load — efficiency and usability.
- P8.3 Deployment Readiness Level (DRL) — a 1–9 scale adapted from TRLs for clinical AI

Table 1 — M-LEAF summary

Pillar	Dimension	What to measure	Candidate metrics	Recommended protocol
P1 Task Fidelity	Diagnostic reasoning; triage; guideline adherence	Clinical appropriateness	Top-k recall; unsafe recommendation rate	Vignettes; guideline-anchored audit
P2 Interaction	Follow-ups; empathy; style	Dialogue quality	0–5 rubrics; readability	Simulated patients
P3 Safety	Hallucinations; contraindications	Harm avoidance	Hallucination rate; hazard rate	Adversarial/red-team
P4 Robustness	Ambiguity; noise; jailbreaks	Stability/resilience	Δ performance vs. clean; attack success	Stress tests
P5 Transparency	Grounding; explanations	Faithfulness & sources	Attribution rate; expert rating	RAG ablations
P6 Calibration	Confidence; deferral	When to defer	Brier/ECE; abstention rate	Re-sampling
P7 Governance	Fairness; privacy; audit	Compliance/equity	Parity gaps; checklist	Demographic perturb.
P8 Workflow	Handoff; efficiency; DRL	Integration	Task time; SUS; DRL	User studies

Anchor rubric (0–5) template (apply per dimension): 0 = unsafe/absent; 1 = poor; 2 = limited; 3 = adequate; 4 = good; 5 = exemplary.

4 Review Methods

The objective of this study is to map existing evaluations of medical large language model (LLM) applications to the dimensions and protocols of the M-LEAF framework. To achieve this, we conducted a review of relevant literature focusing on health and AI-related categories. The search covered publications from 2020 to August 2025, using queries such as combinations of terms like "large language model," "LLM," "GPT," or "retrieval-augmented" with "medical," "clinical," "healthcare," and evaluation-related terms like "benchmark," "safety," "triage," "hallucination," "fairness," or "empathy."

Studies were included if they evaluated an LLM-based system on a health or clinical task, reported quantitative or qualitative outcomes, were written in English or Slovenian, provided reproducible descriptions of prompts or datasets, and involved human or vignette-based evaluations. We excluded purely editorial or opinion pieces without methodological details and studies focused on non-LLM models.

For data extraction, we collected key details from each study, including the setting, task, language, type of LLM or retrieval-augmented generation (RAG) system, guardrails, evaluation protocol, metrics, and results. Each paper was coded against M-LEAF dimensions, noting whether specific dimensions were present or absent and which metrics were used. For interaction and safety dimensions, we applied an anchor-based scoring system ranging from 0 to 5.

The synthesis of findings focused on identifying coverage gaps across M-LEAF pillars, summarizing commonly used metrics, and highlighting under-evaluated dimensions, such as calibration, deferral, and privacy testing. To evaluate the risk of bias and study quality, we adapted existing checklists to assess selection bias (e.g., case mix or language representation), performance bias (e.g., prompt leakage or parameter control), detection bias (e.g., blinding of raters or overlap), and reporting bias (e.g., selective metric reporting). Each study was scored from 0 to 2 (low, unclear, or high risk) for each bias category, enabling a heat-map visualization of quality across studies. To address temporal confounding in the rapidly evolving LLM landscape, we required studies to disclose model or version dates, the currency of RAG sources, and details of any guardrails implemented.

In summary, a reproducible evidence-mapping pipeline is generated that operationalizes M-LEAF across published studies, quantifies dimension coverage and bias, and produces shareable artifacts (coverage tables, quality heat-maps, coded dataset) that ground the framework's recommendations.

M-LEAF organises evaluation into eight pillars. Each pillar contains concrete dimensions with what to measure, candidate metrics, and recommended protocols. Anchored 0–5 rubrics support a reliable rating.

5 Illustrative Case Study (Application of M-LEAF)

The goal is to demonstrate how to apply M-LEAF using an existing comparison of GPT-4o and HomeDOctor.

5.1 Design Principles

GPT-4o (state-of-the-art LLM at the time) and HomeDOCTOR, a Slovenian RAG diagnostic assistant layered on GPT-4o with curated medical content and communication guidelines [13].

5.2 Dataset and Prompting

100 clinical vignettes from the Avey AI Benchmark Vignette Suite [14]. Structured fields: age, sex, chief complaint, clinical history. Prompts were submitted via public GUIs. All interactions were in Slovenian; default model settings were used. An illustrative prompt–response pair is referenced in Figure 1, highlighting how follow-up questioning, terminology, and safety-netting are manifested in a typical HomeDOCTOR exchange.

Danes bom tvoj pacient, ti pa moj zdravnik, prosim te, da se tako tudi pogovarjaš z mano. Napiši ti bom svoje težave in zgodovino bolezni ter druge podatke, prosim te pa, da poveš katere so najverjetnejše diagnoze in mi prosim ponudi dodatne materiale ki mi bodo pomagali izvedeti več o bolezni (uporabni linki, slike, posnetki, članki itd) ter možna zdravila in izdelke ki jih lahko dobim v lekarni?

Opis težave:

Sem 35-letna ženska, ki je pred petimi tedni rodila in zdaj jemljem kontracepcijske tabletko. Že en teden imam bolečo oteklino desne noge. Pred dvajsetimi dnevi sem se vrnila iz tujine po 12-urnem letalskem potovanju. Nisem se nedavno poškodovala ali si zasakala noge. Nisem bila v stiku z bolnimi ljudmi. Koleni me ne škripa, ni otrdela in me ne boli zadaj. Koleni ni otečeno, prav tako ne gleženj. Bolečina se ne širi. Nisem bila izpostavljena klopm ali komarjem. Doma nimam hišnih ljubimcev in nimam vročine. Ob pregledu je obseg mojega desnega meča za 5 cm večji od levega. Mečna mišica je občutljiva na pritisk in noga je toplejša. V preteklosti nisem imela nobenih zdravstvenih težav ali operacij. V družini nimamo podobnih težav.



Možna bolezenska stanja:

1. Globoka venska tromboza (DVT) - strdek v veni noge, ki povzroča oteklino in bolečino.
2. Površinska venska tromboflebitis - vnetje manjše vene blizu kože z bolečino in oteklino.
3. Limfedem - zastoj limfe, ki povzroča oteklino noge.
4. Mišični krč ali poškodba - bolečina in oteklina zaradi preobremenitve mišic.
5. Okužba kože ali podkožja (celulitis) - vnetje z oteklino, toploto in bolečino.

Kako nujen je obisk zdravnika: Obisk zdravnika je potreben.

Vaši simptomi so lahko znak globoke venske tromboze (DVT), kar je resno stanje, ki zahteva takojšnjo zdravniško obravnavo. Zaradi oteklin, bolečin, povečane toplote in nedavnega dolgega potovanja v kombinaciji z jemanjem kontracepcijskih tablet morate čim prej poiskati zdravniško pomoč. Če se pojavijo težave z dihanjem, bolečine v prsih ali nenaden kašelj, takoj pokličite nujno medicinsko pomoč. Spremljajte simptome in ne odlašajte z obiskom zdravnika.

Kako si lahko pomagате sami: Obisk zdravnika je nujen, prosimo ne odlašajte.



Figure 1. Example prompt (Avey vignette, Slovenian) and HomeDOCTOR's answer.

5.3 Dimensions and scoring

Eight dimensions (0–5): accuracy, absence of hallucinations, terminology, follow-up questions, empathy, style, quality of explanation, and referral appropriateness in the draft M-Leaf study. In the second study, the Pillars presented in Table 1 were used.

5.4 Study Design and Statistics

Final-year Slovenian medical students served as raters, each rated part of the items for both systems (no overlap, so inter-rater reliability was not computed). We compared rating distributions across systems using Pearson's χ^2 per dimension and, as a complementary analysis, an expanded Mann–Whitney U test.

5.5 Results

First, a draft pillar schema (a predecessor of the one in Table 1) was used to perform the first studies. Aggregate scores were uniformly high across dimensions (means \approx 4.6–5.0). Figure 2 summarises per-dimension means with 95% CIs for both systems, making visible where trends favour one system despite non-significant tests. Chi-square tests on full

distributions detected no statistically significant differences after multiplicity correction. Mann–Whitney comparisons on expanded counts were consistent with χ^2 results. Descriptively, HomeDOCTOR trended higher on Empathy, Quality of Explanation, Referral Appropriateness, Style, and Accuracy.

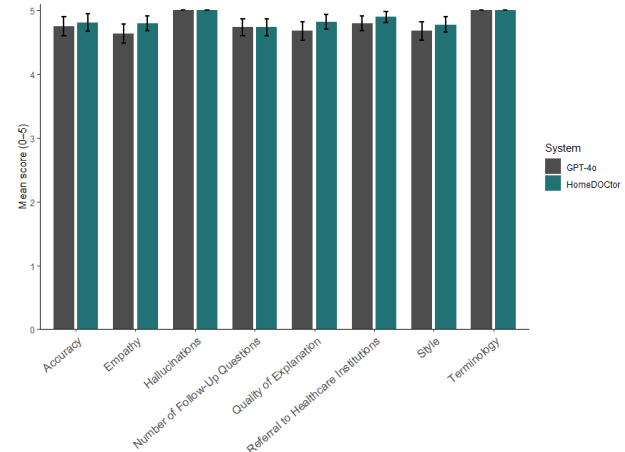


Figure 2: Dimension-level mean scores (0–5) with 95% CIs for GPT-4o vs. HomeDOCTOR.

Interpretation. Under M-LEAF, the pattern of results matters: high Safety and Interaction scores plus low Hallucination incidence support patient-facing feasibility, while explicit uncertainty and deferral behaviours remain to be quantified in future tracks. The overall view on comparison of GPT-4o and HomeDOCTOR is presented in Figure 3.

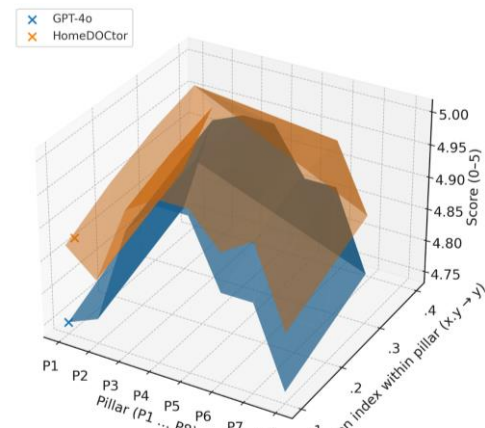


Figure 3: Comparison of GPT40 and Homedocotor through the M-LEAF framework

Takeaway: When reported under M-LEAF, even a compact study yields dimension-level visibility (e.g., empathy↑, hallucination↓), improving interpretability beyond a single accuracy score. The comparisons with the original M-LEAF is presented in Figure 3.

6 Illustrative Case Study (Application of M-LEAF)

System: base LLM (version/date), RAG sources & currency, safety filters, guardrails, logging.

Task & Setting: patient-facing vs clinician-support; language/locale; emergent vs non-emergent.

Protocol: offline vs simulated vs human-in-the-loop; single vs multi-turn; prompt policies; sampling parameters.
 Dimensions: which M-LEAF pillars assessed and why; anchor rubrics used.
 Metrics: exact definitions; unit(s) of analysis; top-k; unsafe-action rate; calibration measures.
 Dataset: source, license, representativeness; emergent case coverage; translations and validation.
 Human Rating: rater expertise, training, blinding, overlap; inter-rater reliability.
 Statistical Analysis: tests, corrections, CIs, power analysis; handling of multiplicity.
 Safety & Governance: privacy tests; fairness slices; post-deployment monitoring plans.
 Limitations: generalisability; ceiling effects; robustness; reproducibility of materials.
 Authors should include a filled checklist as an appendix or supplement.

7 Discussion

Medical large language models (LLMs) are increasingly being developed for patient-facing roles, where avoiding harm, enabling deferral, and providing clear explanations are just as critical as achieving high diagnostic accuracy. The M-LEAF framework addresses this by consolidating diverse metrics into a unified structure, prioritizing safety-critical dimensions as essential components of evaluation.

For retrieval-augmented generation (RAG) systems, it is important to distinguish between the quality of generated outputs and the quality of grounding, which includes aspects like source attribution and citation precision or recall. Evaluations should also ensure that the generated content remains faithful to retrieved evidence, using methods such as ablation studies (comparing performance with and without RAG) and source perturbation tests to assess robustness.

To promote broader adoption, we encourage AI in Healthcare conferences, such as IS 2025, to adopt the MED-LLM-REPORT checklist and include summaries of dimension coverage alongside traditional accuracy metrics. This would provide a more comprehensive understanding of model performance.

However, M-LEAF has limitations. Many of its proposed metrics, such as empathy scales for chatbots, are based on evolving standards that lack established benchmarks. Some dimensions, like privacy or fairness, require specialized audits that go beyond vignette-based studies. Additionally, our illustrative case study is limited by its small scale, single-language focus (Slovenian), and potential ceiling effects, which may restrict its generalizability.

Future work should focus on formalizing dynamic risk-level (DRL) scales for clinical AI to better assess safety. Developing open, multi-turn triage suites would enable more realistic evaluations of patient interactions. Publishing anchor rubrics and starter code for automated metric aggregation could streamline adoption, while integrating M-LEAF with emerging benchmarks like HealthBench would enhance comparability across studies.

8 Conclusion

We introduce M-LEAF, a comprehensive framework and checklist to evaluate medical LLM applications beyond accuracy. By standardising dimensions, metrics, and

protocols, M-LEAF enables comparable, safety-oriented, and deployment-ready assessments. The GPT-4o vs HomeDOctor case study shows how M-LEAF can be applied in practice. We invite the community to adopt and iterate on M-LEAF to make evaluations clinically meaningful.

References

- [1] World Health Organization, Regional Office for Europe. 2022. Health and Care Workforce in Europe: Time to Act. WHO Regional Office for Europe. Retrieved from <https://www.who.int/europe/publications/i/item/9789289058339>
- [2] T. Y. C. Tam, S. Sivarajkumar, S. Kapoor, A. V. Stolyar, K. Polanska, K. R. McCarthy, H. Osterhoudt, X. Wu, S. Visweswaran, S. Fu, P. Mathur, G. E. Cacciamani, C. Sun, Y. Peng, and Y. Wang. 2024. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ Digital Medicine* 7, 1 (2024), 258. DOI:<https://doi.org/10.1038/s41746-024-01258-7>
- [3] S. Bedi, Y. Liu, L. Orr-Ewing, D. Dash, S. Koyejo, A. Callahan, J. A. Fries, M. Wornow, A. Swaminathan, L. S. Lehmann, H. J. Hong, M. Kashyap, A. R. Chaurasia, N. R. Shah, K. Singh, T. Tazbaz, A. Milstein, M. A. Pfeffer, and N. H. Shah. 2024. Testing and Evaluation of Health Care Applications of Large Language Models. *JAMA* 333, 4 (2024), 319. DOI:<https://doi.org/10.1001/jama.2024.21700>
- [4] A. Gilson, C. W. Safranek, T. Huang, V. Socrates, L. Chi, R. A. Taylor, and D. Chartash. 2023. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Medical Education* 9 (2023). DOI:<https://doi.org/10.2196/45312>
- [5] Y. Yanagita, D. Yokokawa, S. Uchida, J. Tawara, and M. Ikusaka. 2023. Accuracy of ChatGPT on medical questions in the National Medical Licensing Examination in Japan: Evaluation study. *JMIR Formative Research* 7 (2023), e48023. DOI:<https://doi.org/10.2196/48023>
- [6] E. Goh, R. Gallo, J. Hom, E. Strong, Y. Weng, H. Kerman, J. A. Cool, Z. Kanjee, A. S. Parsons, N. Ahuja, E. Horvitz, D. Yang, A. Milstein, A. P. J. Olson, A. Rodman, and J. H. Chen. 2024. Large Language Model Influence on Diagnostic Reasoning. *JAMA Network Open* 7, 10 (2024), e2440969. DOI:<https://doi.org/10.1001/jamanetworkopen.2024.40969>
- [7] J. B. Longwell, I. Hirsch, F. Binder, G. A. G. Conchas, D. Mau, R. Jang, R. G. Krishnan, and R. C. Grant. 2024. Performance of large language models on medical Oncology examination questions. *JAMA Network Open* 7, 6 (2024), e2417641. DOI:<https://doi.org/10.1001/jamanetworkopen.2024.17641>
- [8] M. Gams, T. Horvat, Ž. Kolar, P. Kocuvan, K. Mishev, and M. S. Misheva. 2025. Evaluating a Nationally Localized AI Chatbot for Personalized Primary Care Guidance: Insights from the HomeDOctor Deployment in Slovenia. *Healthcare* 13, 15 (2025), 1843. DOI:<https://doi.org/10.3390/healthcare13151843>
- [9] L. S. Bickley. 2020. Bates' Guide to Physical Examination and History Taking (13th ed.). Wolters Kluwer.
- [10] J. Halpern. 2003. What is clinical empathy? *Journal of General Internal Medicine* 18, 8 (2003), 670–674. DOI:<https://doi.org/10.1046/j.1525-1497.2003.21017.x>
- [11] H. L. Semigran, J. A. Linder, C. Gidengil, and A. Mehrotra. 2015. Evaluation of symptom checkers for self-diagnosis and triage: audit study. *BMJ* (2015), h3480. DOI:<https://doi.org/10.1136/bmj.h3480>
- [12] R. K. Arora, et al. 2025. HealthBench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775* (2025). Available at <https://arxiv.org/abs/2505.08775>. Accessed August 19, 2025
- [13] M. Zadobovšek, P. Kocuvan, and M. Gams. 2024. HomeDOctor App: Integrating medical knowledge into GPT for personal health counseling. In *Proceedings of the Information Society 2024 Conference*, Ljubljana, Slovenia, October 2024.
- [14] Avey. 2022. Benchmark vignette suite. Avey AI. Retrieved March 11, 2022 from <https://avey.ai/research/avey-accurate-ai-algorithm/benchmark-vignette-suite>