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

http://doi.org/DOI_RECEIVED_AFTER_REVIEW

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].

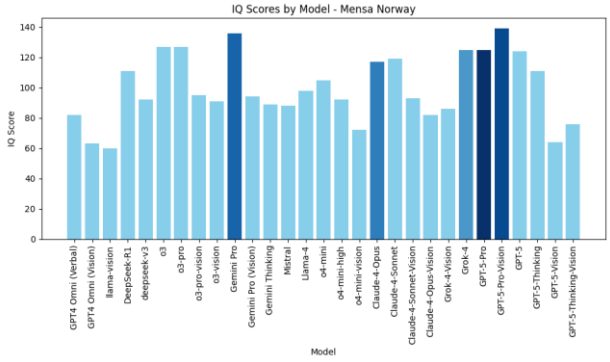
Table 2: Results from Humanity's Last Exam (HLE), which measures closed-ended reasoning across diverse graduate-level topics

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