FRONTIER MODELS

The 2026 frontier models and the health record: what the launch evidence measured, and what it did not

ABSTRACT

Between June and November 2026 four laboratories released a new generation of general-purpose language models, and each reported results on health care tasks at launch. This paper reviews what those reports and the independent measurements of the same models establish. On the benchmark all four now quote, current models score 57 to 68 against a specialist physician baseline of 43.7, and the ranking changes with who ran the test and how answer length is scored. No peer-reviewed study yet measures a patient outcome with any of the four models. The controlled evidence that exists describes products built on their predecessors.

Type Narrative review of vendor and independent evidence **References** 26 **Reading time** 15 min

Last reviewed November 2026

1 A generation that reports on health care

Physician use of AI has outrun the evidence for it. The American Medical Association's 2026 survey found 81% of physicians using AI in practice, against 38% in 2023, with an average of 2.3 use cases each and research summarization and documentation the most common.[1] Most of that use runs on general-purpose models that their makers revise several times a year.

Four such models define the current frontier: Anthropic's Claude Fable 5.2, Google DeepMind's Gemini 4 Argon, OpenAI's GPT-6 Astra, and Grok 4.7 from xAI. They differ from earlier generations in one respect that matters to a practice. Each maker published results on health care tasks in its launch material, next to the coding and reasoning scores that used to stand alone.[2,4,5,6] That gives a buyer something to read. It does not, by itself, give a buyer something to rely on.

43.7	57–68	81%	0
specialist physician baseline on HealthBench Professional	range of current frontier models on the same benchmark	of physicians reporting AI use in practice in 2026	peer-reviewed patient-outcome studies of any of the four models

This paper makes three claims. The launch figures are real measurements, but most are self-reported and the one shared benchmark was built by a competitor. Where a maker published its full evaluation, the picture is mixed in ways the launch table does not show. And the randomized evidence that clinical AI helps anyone was gathered on products built one or two model generations earlier, so it cannot be transferred to these models without argument. It is a review of published material and contains no new data.

2 The four models, as their makers document them

Table 1 sets out what each maker has published. Every entry in the last column is the maker's own claim or a customer statement the maker chose to publish.

TABLE 1

The four models named in this paper, with release date, price where published, and the health care evidence each maker offered at launch.

Model	Maker and release	Published price per million tokens	Health care evidence at launch
Claude Fable 5.2[2,3]	Anthropic, 9 November 2026	$8 input, $40 output	A health section in the system card: two benchmarks run by a third party at the maker's request, one run by the maker
Gemini 4 Argon[4]	Google DeepMind; release date not taken from a primary source	Not taken from a primary source	No health benchmark in the launch tables; one customer statement about synthetic charts
GPT-6 Astra[5,8]	OpenAI, September 2026	Not taken from a primary source	63.4% on HealthBench Professional, against 60.5% for the maker's previous model
Grok 4.7[6]	xAI, publishing as SpaceXSI, 21 September 2026	$2 input, $6 output	56.7% on HealthBench Professional, against 48.5% for the maker's previous model

The Google entry deserves a note. Its launch page carries no health benchmark. The health claim is a statement from a clinical documentation company that, on more than 200 synthetic longitudinal charts, the model had the highest recall and diagnosis-code specificity of the models that company tested, and that validation on real patient records had begun.[4] That is a customer's unpublished internal test on synthetic data. It is a reason to look, and nothing more.

OpenAI's release went first to a limited set of organizations, with paid ChatGPT plans, the API and two cloud platforms to follow.[5] Trade reporting of the launch quoted an outside consultant urging boards to recognize the risks before adopting it, which is an unusual thing to find in launch coverage and worth noting for that reason.[5]

3 The benchmark all four now quote

Three of the four makers report a score on HealthBench Professional. Google's launch tables do not. A reader who understands this one benchmark understands most of what the launch material says about health care.

What it is

HealthBench Professional was published by OpenAI in April 2026. It contains 525 tasks written by physicians, drawn from a pool of 15,079, across three kinds of work: care consultation, writing and documentation, and medical research. Responses are graded against rubrics that physicians wrote and adjudicated. About one third of the tasks are deliberate attempts by physicians to make a model fail, and tasks that OpenAI's own recent models found difficult are over-represented by a factor of roughly 3.5.[7]

The benchmark includes a human baseline. Physicians matched to the specialty of each task, given unlimited time and web access, wrote answers to all 525. Those answers scored 43.7. The best system in the original paper, a clinician-facing product built on GPT-5.4, scored 59.0, and the same model without that product scored 48.1.[7]

This is a better instrument than the examination questions that dominated earlier work, a problem a companion paper in this library sets out. The tasks resemble what clinicians ask. It is also a benchmark designed, sampled and first scored by one of the four competitors, selected partly on what that competitor's models found hard, and described in a preprint that has not been peer reviewed.[7]

What the scores are

TABLE 2

Published HealthBench Professional scores for current models, with who measured each. Scores are length-adjusted percentages of rubric points. Rows measured by different parties are not strictly comparable.

System	Score	Measured by
Claude Fable 5.2	67.7	Artificial Analysis, at Anthropic's request[3]
GPT-6 Astra	66.0	Artificial Analysis[9]
GPT-6 Astra	63.4	OpenAI[8]
Claude Sonnet 5.5	63.0	Artificial Analysis[9]
Claude Opus 5.5	59.9	Artificial Analysis[9]
ChatGPT for Clinicians on GPT-5.4	59.0	OpenAI, April 2026[7]
Claude Fable 5.1	57.8	Artificial Analysis[3]
Grok 4.7	56.7	xAI[6]
GPT-5.4 without the clinician product	48.1	OpenAI, April 2026[7]
Specialty-matched physicians	43.7	OpenAI, April 2026[7]

Two readings hold up. In about six months the best general-purpose score moved from 48.1 to 67.7. And every maker with a published 2026 score clears the physician baseline by at least 13 points. No Gemini 4 Argon score from a source opened for this review is included, so the table has no row for it.

Why the ranking is not stable

Three properties of the measurement move the order of the models.

The first is who ran it. xAI's launch table lists Claude Fable 5.1 at 62.1.[6] Artificial Analysis measured the same model at 57.8.[3] OpenAI reported its own model at 63.4 and the independent tester reported 66.0.[8,9] Differences of three to four points between runs of one model are as large as the differences between models.

The second is the grader. Rubric scoring is done by a language model. Anthropic's system card states that the scores it reports for its own models are about 4 to 6.5 points lower than in its earlier cards, because the earlier ones were graded by one of its own models and these were graded by an OpenAI model.[3] A change of grader moved every score by more than the gap between first and third place.

The third is length. The published score penalizes answers longer than 2,000 characters and rewards shorter ones. Claude Opus 5.5 had the highest unadjusted score of any model in the independent leaderboard, 71.4%, and fell to sixth after adjustment because its answers averaged about 5,900 characters. One OpenAI model was the only one to average under 2,000 characters and the only one

whose score rose.[9] Whether a longer and more complete answer serves a clinician worse is a judgment about clinical work. The benchmark makes that judgment in advance, and its author is also the maker of the only model that averaged under the threshold.

LIMITATIONS

No confidence intervals are published for the Artificial Analysis figures, so differences of a point or two may be noise.[3] The benchmark grades text against a rubric. It has no patient, no encounter and no outcome. A score above the physician baseline shows that a model writes answers a rubric prefers to answers physicians wrote unaided. It does not show that a model practices medicine better than a physician, and the preprint does not claim so.

4 What a full evaluation shows that a launch table does not

Anthropic published the complete health section of its evaluation for Claude Fable 5.2, and it is the most useful document in this review because it reports results the maker had no commercial reason to highlight.[3]

On HealthBench Professional the model leads only after length adjustment. Its unadjusted score, 71.1%, is 0.3 points below Claude Opus 5.5, another of the same maker's models.[3] On PhysicianBench, a 100-task evaluation the maker ran itself and graded with one of its own models, Fable 5.2 passed 64.4% of tasks against 68.4% for Opus 5.5, and the card states that the gaps among its models are within noise. On the patient-facing HealthBench, which holds 5,000 conversations with members of the public and health workers, the newer model's unadjusted score was 58.8%, below the 63.9% of the version it replaced.[3,10]

Effort settings matter as well. On PhysicianBench the same model passed 55.0% at its lowest reasoning setting and 61.8% at a high one, a difference the card reports as statistically reliable. On HealthBench the score fell slightly as effort rose.[3] A product that runs a model at a cheaper setting is not running the model whose score appears in the launch table.

The practical conclusion is that the newest model is not reliably the best at clinical text, and that another of the maker's own models can match it. One independent leaderboard makes the cost point sharply: a small OpenAI model scored four points below GPT-6 Astra at under 2% of the cost per task.[9]

Composite and long-record measures

Artificial Analysis also publishes a composite Healthcare and Medical Index that weights medical knowledge, agentic knowledge work, long-context reasoning, tool use, non-hallucination and general reasoning. Claude Fable 5.2 leads it at 65, ahead of Claude Opus 5.5 at 58 and Gemini 4 Carbon, a sibling of Argon, at 57.[11] Three of its six components use medical material and three are general tests, so the index says as much about general capability as about medicine.

Its long-context component is the closest published measure to reading a real chart. It asks a model to synthesize medical case files of 70 to 150 pages. When it was introduced in August 2026 the leading model scored 64.4%, and the report of the launch described most models as failing.[12] That account reaches this paper through a news summary, not the tester's own methodology, and should be read with that in mind.

5 Safeguards are a clinical variable

The most capable models carry controls against misuse in biology, and in 2026 those controls reached ordinary clinical questions.

The first Fable model shipped in June with a classifier that sent biology-related requests to an older, less capable model. On 7 August Anthropic reported that it had retrained the classifier and cut those reroutes by about 85%, so that everyday questions about laboratory results and symptoms, and more of the work health professionals bring, would be answered by the newer model. Requests the maker classes as dual-use, including virology, toxicology and molecular design, remained blocked, and the maker stated that the model was not yet usable for professional biology research or drug development.[13] With Fable 5.2 it reported about 55% fewer blocks overall than the previous version and about 40% fewer in biology.[2]

Research access moved to a separate track. A verification program opened in September gives vetted life-science organizations access under looser biology safeguards, replaces the blocking of single requests with monitoring of usage against a declared purpose, and retains that traffic for 30 days.[14]

Two consequences follow for anyone building on these models. A benchmark score measured with safeguards on is a score for a system that may answer some questions with a different model, which the system card says in terms.[3] And behaviour can change between a Tuesday and a Wednesday without a new model name. The governance questions this raises are the subject of the library's paper on monitoring after go-live.

6 From the model to the record

Capability matters to a practice only when it reaches the chart. Three developments in 2026 are documented well enough to describe.

Chart access inside the dominant EHR

In September OpenAI connected its health care product to Epic. Authorized users can bring appointment notes, laboratory results, medications and specialist documentation into the assistant, which is designed to point back to the underlying chart entries so that a summary can be checked against its source. Access is granted through an organization's administrators, not to individual accounts, and UCSF Health was piloting it. The reporting does not say which model powers the integration.[15] No evaluation of the integration has been published.

Administrative work

The day after the first Fable model launched, a health care software company announced that it could run the model across its products for call handling, patient intake and coding. It stated that the maker's own evaluation included prior authorization, denials and appeals, and equipment orders run in simulated EHR and payer-portal environments, and it published no accuracy or cost results for those tasks.[16] The health section of the later Fable 5.2 card lists no such evaluation.[3]

The payer side is automating as well. Medicare's WISeR model began on 1 January 2026 in six states and applies technology vendors' AI and machine learning, with clinician review, to prior authorization for 13 groups of services in traditional Medicare. In May the Government Accountability Office concluded that the notice creating the model was a rule that should have been submitted to Congress; a Senate resolution to overturn it was introduced and had not passed. CMS states that licensed clinicians make every recommendation not to pay.[17] Nothing public ties a WISeR contractor to any of the four models. What automation does to authorization and denial is the subject of the library's paper on the revenue cycle. No study yet measures the result when both sides of an authorization are automated.

What has not been shown

None of these is evidence of benefit. Each is a deployment or a claim of readiness. The distance between a model that scores well on written tasks and a product that improves a clinic day is where most of the earlier generation's promises were lost, and nothing published about these four models closes it.

7 The controlled evidence describes an earlier generation

Randomized evidence that language models help clinicians exists, and it is strongest for ambient documentation. It was gathered before any of the four models existed.

A three-arm randomized trial of 238 physicians ran from November 2024 to January 2025. One documentation product reduced time in the note by 9.5% against usual care. The second, randomized in the same trial, produced a 1.7% reduction that was not statistically significant. Clinicians reported occasional inaccuracies with both.[18] A stepped-wedge trial of 66 clinicians found a 20% reduction in burnout.[19] A crossover trial of 160 clinicians found that two products both lowered burnout scores and differed from each other by about three minutes of note time a day.[20] The largest observational study, of

more than 1,800 users and 6,770 controls at five hospitals, found about 16 fewer minutes of documentation a day, no significant change in time spent in the record outside working hours, and only 32% of users relying on the tool for more than half their visits.[21] The library's review of ambient documentation appraises these studies in detail.

The effect is real, modest and specific to the product. Two tools in one trial, built on models of the same period, produced different results. That is the strongest available evidence that the model is not the unit that matters. The product is.

Knowledge that did not transfer

The largest randomized study of the public using language models for medical decisions, published in February 2026 with 1,298 participants, points the same way. People assisted by a model identified conditions and chose a course of action no better than people using their usual sources. Participants often did not know what to tell the model, small changes in wording produced different answers, and replies mixed good and poor recommendations.[22] The models tested predate the 2026 frontier by more than a year. The finding is still the best available check on the inference from a benchmark to a person, because a benchmark supplies a complete, well-posed question and a patient does not.

LIMITATIONS

The gap in time cuts both ways. Trials of older models cannot show that the new ones fail, and they cannot show that the new ones succeed. A trial takes longer than a release cycle: the oldest of the four models was about ten weeks old when this paper was reviewed and the newest was four days old. Until trials report, the honest description of the evidence for these models in clinical use is that there is none of the randomized kind.

8 The rules these models are deployed under

Regulation moved in 2026, and mostly toward permitting more.

The Food and Drug Administration revised its guidance on clinical decision support software in January. The revision extends enforcement discretion to software that offers a single recommendation, provided a clinician can independently review the basis for it.[23] A language model whose reasoning a clinician cannot inspect does not obviously meet that condition, and the library's paper on what binds an EHR vendor treats the question at length.

In August the agency issued a discussion paper on medical devices that use generative AI. It floats assessment by competency, pairing benchmark testing with clinical confirmation scaled to risk, and a voluntary master file through which the maker of a foundation model could give the agency confidential information for device makers to reference. It also asks how a device maker should detect changes to a third-party model it does not control. The paper is not guidance and binds no one; comments closed in October.[24]

One clearance is often described as the first for a device built on a patient-facing language model. It was granted in December 2025 for software that delivers insulin dosing instructions to adults with type 2 diabetes within a plan set by a clinician, and it was cleared as substantially equivalent to an insulin titration product from 2019. The description comes from secondary commentary and should be checked against the agency's record before it is relied on.[25]

Retention is a contract term, not a model property

A business associate agreement does not settle how long a model vendor keeps what it is sent. Anthropic's launch post states that customers of its newest models have 30-day retention for safety monitoring, and that zero retention is available to eligible enterprise customers through a program, in private beta, that runs the monitoring inside the customer's own infrastructure.[2] A third-party compliance guide reports that the same maker's consumer and team plans carry no business associate agreement on any model; that is a guide's reading, not the maker's terms.[26] The general position, including what the HIPAA rules require when a model touches the chart, is set out in a companion paper. The point for this review is narrower: the terms differ by model, by plan and by product surface, and the most capable model may carry the least favourable terms.

9 What follows

The 2026 models have done three things that can be documented. They raised measured performance on written clinical tasks well past an unaided specialist baseline. They brought makers to publish health evaluations, in one case with enough detail to show where the newest model is not the best. And they reached the record through at least one major integration. Their effect on patients, on diagnostic error and on cost has not been measured.

For an independent practice, five questions separate a claim worth reading from one worth setting aside.

1. **Which model, at which setting.** A product may run a cheaper model or a lower effort level than the one in the launch table, and the full evaluations show that this changes the score.[3]
2. **Who measured it.** A maker's figure and an independent figure for the same model have differed by three to four points, and a change of grader by more.[3,6,9]
3. **What the comparison was.** A score above the physician baseline compares written answers under a rubric. It is not a comparison of care.[7]

4. **Whether the product was tested, or only the model.** In the one trial that randomized two products, one worked and one did not.[18]
5. **What the contract says about retention and safeguards.** Both vary by plan and can change without a new model name.[2,13]

Three studies would settle most of what is open: a randomized trial of a chart assistant built on a current model, with diagnostic and safety endpoints; an evaluation of automated prior authorization that measures time to decision and overturn rates on both sides; and an independent replication of the shared benchmark graded by physicians, with and without the length adjustment. Until they exist, a clinician's review of what the model wrote remains the control, for the reasons the library's paper on human oversight gives.

DECLARATION OF INTEREST

ChartVoyant builds its product on a language model from one of the four makers discussed here and competes with products built on the others. This paper recommends no model. Every score in it is attributed to the party that measured it, and where a maker's figure and an independent figure differ, both are given.

REFERENCES

Entries 2 to 6, 8, 13, 14 and 16 are makers' own publications or reports of them, and carry the weight of self-report. Entries 9 and 11 are independent measurements. Entries 7 and 10 are preprints written by the maker of one of the models, read in abstract. Entries 18 to 20 and 22 are randomized studies and entry 21 is a cohort study; they carry the most weight here, and none of them tested any of the four models. Entries 21 and 22 were read through their institutions' releases, entries 18 and 20 in abstract, and entry 19 through a summary of its authors' presentation. Entries 12, 15, 17, 24, 25 and 26 are secondary accounts, identified as such where the text relies on them. Every figure is quoted at the precision its source reports.

1. American Medical Association. *AMA: AI usage among doctors doubles as confidence in technology grows*. Press release on the 2026 Physician Survey on Augmented Intelligence. 12 March 2026. ama-assn.org SURVEY
2. Anthropic. *Claude Fable 5.2 and Mythos 5.2*. 9 November 2026, amended 11 November 2026. anthropic.com VENDOR REPORT

3. Anthropic. *System Card: Claude Fable 5.2 and Claude Mythos 5.2*. Section 8.16, Healthcare. 9 November 2026. anthropic.com VENDOR REPORT

4. Google DeepMind. *Gemini 4*. Model page with launch benchmark tables and customer statements. Accessed 13 November 2026. deepmind.google VENDOR REPORT

5. Fox A. *OpenAI launches GPT-6 Astra model*. Healthcare IT News. September 2026. healthcareitnews.com TRADE PRESS

6. SpaceXSI. *Grok 4.7*. Launch announcement with benchmark table. 21 September 2026. x.ai VENDOR REPORT

7. Hicks RS, Trofimov M, Lim D, et al. *HealthBench Professional: Evaluating Large Language Models on Real Clinician Chats*. arXiv preprint arXiv:2604.27470. 30 April 2026. arxiv.org/abs/2604.27470 PREPRINT

8. Makhene T. *OpenAI launches GPT-6 Astra with enhanced healthcare cybersecurity*. Paubox. 16 September 2026. Reports the maker's published benchmark figures. paubox.com TRADE PRESS

9. Artificial Analysis. *HealthBench and HealthBench Professional leaderboards*. 19 October 2026. artificialanalysis.ai INDEPENDENT BENCHMARK

10. Arora RK, Wei J, Hicks RS, et al. *HealthBench: Evaluating Large Language Models Towards Improved Human Health*. arXiv preprint arXiv:2505.08775. May 2025. arxiv.org/abs/2505.08775 PREPRINT

11. Artificial Analysis. *Healthcare & Medical Index*. Accessed 13 November 2026. artificialanalysis.ai INDEPENDENT BENCHMARK

12. AlphaSignal Newsroom. *Artificial Analysis' MLCR-AA Shows Most AI Models Fail Medical Reasoning*. Republished by Hermes AI. 21 August 2026. hermes-ai.net TRADE PRESS

13. Anthropic. *Improving Fable 5's biology safeguards*. 7 August 2026. anthropic.com VENDOR REPORT

14. Anthropic. *Introducing the Life Sciences Verification Program*. 17 September 2026. anthropic.com VENDOR REPORT

15. HLTH. *OpenAI Integrates ChatGPT With Epic EHRs*. 8 September 2026. hlth.com TRADE PRESS

16. Commure. *Anthropic's Mythos-Class AI, Secured for Healthcare, Now on Commure*. 10 June 2026. commure.com VENDOR REPORT

17. HCC Buddy. *Medicare's WISeR AI prior-auth model still binds your claims while Congress moves to repeal it*. 8 July 2026. Reports the CMS notice at 90 Fed. Reg. 28749, GAO decision B-337994 and S.J.Res.192. hccbuddy.com TRADE PRESS

18. Lukac PJ, Turner W, Vangala S, et al. *Ambient AI Scribes in Clinical Practice: A Randomized Trial*. NEJM AI. 2025;2(12). doi:10.1056/AIoa2501000 RCT

19. Afshar M, Baumann MR, Resnik F, et al. *A Pragmatic Randomized Controlled Trial of Ambient Artificial Intelligence to Improve Health Practitioner Well-Being*. NEJM AI. 2025;2(12). doi:10.1056/AIoa2500945 RCT

20. Chowdhury A, Casey M, Wilson J, et al. *Comparing ambient scribes: a randomized crossover clinical trial addressing ambient scribe technologies' impact on physician burnout*. Journal of the American Medical Informatics Association. 2026;33(5):990–999. doi:10.1093/jamia/ocag018 RCT

21. Rotenstein L, Mishuris RG, et al. *Changes in Clinician Time Expenditure and Visit Quantity With Adoption of Artificial Intelligence-Powered Scribes: A Multisite Study*. JAMA. 2026. doi:10.1001/jama.2026.2253 COHORT

22. Bean AM, Payne RE, Parsons G, et al. *Reliability of LLMs as medical assistants for the general public: a randomized preregistered study*. Nature Medicine. 2026;32(2):609–615. doi:10.1038/s41591-025-04074-y RCT

23. U.S. Food and Drug Administration. *Clinical Decision Support Software: Guidance for Industry and Food and Drug Administration Staff*. Final guidance issued 29 Jan. 2026, superseding the version issued 6 Jan. 2026 and the final guidance of 28 Sept. 2022; Docket No. FDA-2017-D-6569. fda.gov GUIDANCE

24. Arnold & Porter. *FDA Seeks Public Feedback on Regulatory Approach for Generative AI-Enabled Medical Devices*. Advisory, 1 September 2026, on the FDA discussion paper of 18 August 2026; Docket No. FDA-2026-N-7874. arnoldporter.com COMMENTARY

25. Topflight. *FDA-cleared LLM clinical AI*. 2026. Commentary on 510(k) K253281, cleared 23 December 2025 against predicate K181916. topflightapps.com COMMENTARY

26. Layer3 Labs. *Is Claude Fable 5 HIPAA compliant?* Updated 5 July 2026. layer3labs.io COMMENTARY

RELATED PAPERS

Evaluating clinical language models

What a benchmark score controls for, and what it does not.

Ambient documentation and the clinician’s day

What the randomized trials of ambient scribes actually measured.

PHI and the model

What HIPAA requires when a language model touches the chart.