Can AI Reliably Interpret Blood Test Results?

Hallucination, context and the limits of general-purpose AI.

AI can explain a blood result in seconds. The harder question is whether it knows what it should refuse to conclude.

USEFUL

Explain the terminology.

Define a marker, compare concepts, summarize a report or help formulate better questions.

RISKIER

Infer the cause.

Missing context, time, units or discordant data can make a plausible narrative feel more certain than the evidence allows.

GOVERNED

Know where the answer ends.

Structured inputs, evidence provenance, deterministic checks, uncertainty and handoff can narrow what the system is allowed to say.

The important question is not "Can AI answer?"
It is: **what kind of answer is this task safe to support, and what controls are needed before the answer should influence a health decision?**

IN 60 SECONDS

Can AI reliably interpret a blood test?

SHORT ANSWER

Sometimes, for bounded tasks. General-purpose AI can be useful for explaining terms, organizing information and identifying questions. Reliability becomes harder to assume when the task requires longitudinal interpretation, causal inference, diagnosis, treatment decisions or personalized action.

WHY

Generative AI can produce confident but incorrect content, omit relevant information, misread numerical context, flatten time, force discordant data into one story, or make a stronger causal claim than the evidence supports.

WHAT CHANGES THE INTERPRETATION

Task scope, source quality, units and metadata, longitudinal history, missing information, prompt wording, conversation context, evidence grounding, deterministic normalization, safety controls and the ability to lower confidence or hand off.

WHAT WE CANNOT CONCLUDE

There is no single universal “AI error rate” for health questions. Performance depends on the model, version, prompt, task, safeguards, evaluation method and time of testing. This guide also does not establish that Sentinel eliminates hallucination.

EVIDENCE FOUNDATION

NIST identifies confabulation as a generative-AI risk [1]; WHO calls for governance, oversight and evidence generation for large multimodal models in health [2]. Recent audits show meaningful safety and accuracy limitations in patient-facing medical advice, while other studies show that constrained tasks and retrieval/safety frameworks can materially improve performance. The practical conclusion is task- and system-specific, not “AI is safe” or “AI is unsafe.”

VALIDATION STATUS

This is an educational Exia Bio interpretation guide, not a comparative clinical validation study. Sentinel is not represented here as a validated diagnostic system or as a system proven to eliminate AI hallucination.

What general-purpose AI is genuinely good at

The useful question is not whether AI belongs in health at all. It is which tasks fit the system's capabilities and which tasks demand additional controls.

TASK	WHY AI CAN HELP	WHAT STILL NEEDS CHECKING
Explain a laboratory term	Natural-language explanation and comparison are core strengths.	Definition, units and source accuracy.
Summarize a long report	AI can compress text and organize repeated information.	Omissions, OCR/data-extraction errors and whether the summary preserves uncertainty.
Generate questions to ask next	AI can surface alternative questions and educational directions.	Whether the questions are relevant to the actual clinical context.
Compare repeated values	AI can describe numerical direction when the data are correctly supplied.	Comparability, timing, biological/analytical variation and interpretation of the change.
Recommend diagnosis or treatment	This is a high-consequence inference task, not merely language generation.	Clinical evaluation, validated decision pathways, contraindications, missing data and professional responsibility.

A system can be very useful at explanation and still require much stronger governance before it should influence diagnosis, treatment or medication decisions.

02 · RELIABILITY CHANGES WITH THE TASK

The same AI can be helpful at one level and unsafe at another

LEVEL 1

Explain

“What is ferritin?” Low interpretive burden if the definition is accurate.

LEVEL 2

Organize

“Summarize these results by date.” Requires correct extraction and provenance.

LEVEL 3

Interpret

“What changed?” Requires comparability, context, relationships and uncertainty.

LEVEL 4

Attribute cause

“What caused the change?” Competing explanations and causal evidence matter.

LEVEL 5

Act

“What should I take or stop?” Highest consequence; requires appropriate clinical safeguards.

This is why asking “Is AI accurate?” is usually too broad. A narrow calculation or constrained interpretation benchmark can perform well while open-ended patient advice still produces consequential failures. Different evaluations answer different questions.

03 · FAILURE MODES

Seven ways a plausible health answer can become unreliable

1

Invented or unsupported evidence

A source, quotation or factual statement may be wrong, nonexistent or unrelated to the claim it is used to support.

2

Numerical context errors

Units, reference intervals, collection conditions or assay-specific details can be lost or misread even when the raw number is copied correctly.

3

Missing information becomes invisible

A fluent answer can make an incomplete case look finished. The model may fill narrative gaps instead of keeping them explicit.

4

Time gets flattened

Three years of data can be treated like one snapshot. Intervention dates, lab changes and historical baselines may disappear from the reasoning.

5

Discordance gets smoothed away

Related markers that point in different directions may be averaged into one tidy story even though the disagreement itself is informative.

6

Association becomes cause

Mechanism, correlation or a before/after change may be upgraded into “this caused that” without evidence that supports the stronger sentence.

7

Confidence outruns the decision

The system may offer highly personalized next steps even when the available data do not justify diagnostic or treatment-level certainty.

▸ Go deeper: system-level manipulation can change a medical answer

Hallucination is only one part of the problem. A response can contain no fabricated fact and still be unsafe because it omits context, misuses a correct source, overstates causality or recommends action beyond what the evidence supports.

AI performance is real - and highly task dependent

Recent evidence supports a more nuanced position than either “AI is unreliable” or “AI is ready to replace clinical interpretation.”

Published AI error rates should not be compared without considering what was actually tested. A low error rate on a tightly constrained task and a higher problematic-response rate from open-ended advice are not measurements of the same construct.

EVIDENCE	WHAT IT SHOWS	WHAT IT DOES NOT ESTABLISH
NIST Generative AI Profile	Confabulation, information integrity and other generative-AI risks require explicit governance, measurement and monitoring.	A healthcare-specific error rate or a verdict on any one product.
WHO guidance on large multimodal models in health	Health uses require governance, oversight, transparency, stakeholder involvement and evidence generation proportionate to risk.	That general-purpose chatbots are validated clinical systems.
HealthAdvice red-team study (2026)	Across 222 realistic patient questions, four public chatbots differed materially in problematic and unsafe response rates; unsafe responses were observed in every evaluated system [3].	A universal failure rate for all medical questions, all model versions or blood-test interpretation specifically.
BMJ Open misinformation-prone audit (2026)	Five public chatbots showed limitations in scientific accuracy and reference quality under health prompts chosen for misinformation susceptibility [4].	That routine low-risk questions have the same failure rate or that the tested systems are unchanged today.
Med-RISE comparative study (2025)	Retrieval plus fact/safety filtering improved benchmark accuracy and reduced measured hallucinations across four medical QA datasets [5].	That retrieval alone solves medical reliability, or that multiple-choice benchmark gains equal safe personalized patient advice.
Constrained blood-gas studies	LLMs can show high concordance on narrowly defined arterial blood-gas interpretation tasks [8][9].	That the same performance transfers to open-ended longitudinal, causal or treatment questions.

The most defensible conclusion is therefore operational: **define the task, define the acceptable failure modes, measure them, and build controls around the specific use case.**

▸ Go deeper: why published AI error rates should not be compared casually

05 · GROUNDING

Why retrieval helps - but does not finish the job

Giving a language model better sources can reduce one major failure mode: answering from incomplete or outdated internal knowledge. But a retrieved source still has to be selected, interpreted and applied correctly.

Source quality

A retrieved document can be authoritative, weak, outdated or inappropriate for the question.

Population fit

A correct study may not match the user's population, assay, intervention or outcome.

Claim fit

A study showing a biomarker effect does not automatically support a health-outcome claim.

Conflict handling

Two credible sources can disagree. The correct output may be uncertainty rather than picking the convenient source.

Numerical normalization

Retrieval does not replace deterministic handling of units, dates, reference metadata or duplicate records.

Missing-data logic

Retrieval cannot recover patient context that was never supplied; the system still needs to represent what is unknown.

A wrong source can ground a wrong answer. A correct source can ground an overstated answer.

06 · WORKED EXAMPLE

Same data. Three different system behaviors.

Consider an illustrative user who asks: **“Everything is still within range. Does that mean nothing important has changed?”**

MARKER	TEST 1	TEST 2	TEST 3
HbA1c (%)	5.4	5.5	5.6
Fasting glucose (mmol/L)	5.0	5.1	5.1
Triglycerides (mmol/L)	0.9	1.2	1.8
ALT (U/L)	18	24	34

Fluent answer

“Most values are still within common ranges, so there may be no immediate cause for concern. Continue healthy habits and monitor routinely.”

Why it sounds reasonable

- Calm and readable.
- Does not invent an overt diagnosis.
- Uses familiar “within range” logic.

What it may miss

- The directional triglyceride/ALT change.
- Whether the tests are comparable.
- Context and intervention timing.
- The distinction between “not flagged” and “unchanged.”

Grounded answer

“Reference intervals do not answer every longitudinal question. Triglycerides and ALT have increased across the three tests, while HbA1c and fasting glucose changed less. Repeated within-range movement can still justify closer contextual review.”

What grounding improves

- More precise reference-range language.
- Recognition of longitudinal direction.
- Less likely to dismiss change solely because it is unflagged.

What still needs governance

- Source applicability.
- Measurement comparability.
- Competing explanations.
- What action, if any, is justified.

Governed interpretation

“The available data support a more specific follow-up question, not a diagnosis. Triglycerides and ALT show directional change across repeated tests, whereas the glycaemic markers are comparatively stable. Before narrowing the interpretation, confirm comparability and relevant context such as collection conditions, medication/supplement timing and other clinical information.”

What is preserved

- Relationships and trajectory.
- Discordance rather than forced reconciliation.
- Missing context as missing.
- A bounded conclusion.

What is explicitly withheld

- Diagnosis.
- Causal attribution.
- Treatment or supplement recommendation.
- False certainty from a tidy narrative.

The point of the third state is not that a governed system must always be more cautious or longer. It is that the system has an explicit reason for each sentence it is allowed to make - and an explicit place to stop.

07 · PROMPT AND CONVERSATION CONTEXT

The model also interprets the way you ask

A general-purpose chatbot does not receive only your laboratory values. It receives the wording of your question, the context you choose to include and,

in a continuing chat, the conversation that came before.

That matters because ordinary patient prompts are often incomplete, inaccurate or framed around a prior belief. In the HealthAdvice study, the authors specifically note that most patients are not trained in prompt engineering and that prompt quality can affect medical-chatbot responses [3].

Short is not the problem. Missing or misleading context is. A longer prompt can still be wrong, and a short prompt can be perfectly adequate for a bounded task. The reliability question is whether the system can recognize what information and safeguards the task actually requires.

Same blood data. Different instructions.

The values below are unchanged from the worked example. Only the user instruction changes. These are illustrative prompts, not benchmark outputs from a named product.

TYPICAL QUICK QUESTION

> **"Everything is still normal. Is there anything to worry about?"**

This framing foregrounds reassurance from reference ranges. It does not explicitly ask the model to compare trajectories, check whether tests are comparable, preserve missing context or resist causal inference.

CONTEXT-RICH QUESTION

> **"These are three blood tests over time. Compare the trajectory, not only whether each value is in range. Flag important directional changes, keep missing context explicit, and do not assume a cause. Tell me what would need to be checked before making a stronger interpretation."**

This wording explicitly supplies several safeguards that a knowledgeable user may want. It can improve task specification, but it does not turn a general-purpose chatbot into a validated longitudinal interpretation system.

The blood data did not change. The user's ability to specify the task did.

A consumer should not need to know how to prompt like a scientist before a health-AI system behaves responsibly. If safe interpretation depends on the user remembering to request longitudinal comparison, unit checks, uncertainty, competing explanations and causal restraint, part of the governance burden has effectively been outsourced to the user.

Conversation history can become hidden input

Conversation history can be useful when it adds accurate, relevant information. But it can also introduce an assumption, a prior model error or a strongly framed hypothesis that influences what comes next. Multi-turn medical studies have observed sycophantic shifts under user pushback [13], reliability degradation as alternatives are introduced sequentially [14], and propagation of earlier model errors when model-generated history is carried into later turns [15].

FRESH TURN

Laboratory facts only

“My ALT rose from 18 to 34 U/L. What does that change mean?”

The current question contains the measured change but no asserted explanation.

USER BELIEF IN HISTORY

An unverified cause is already suggested

“I've been drinking more recently, so I assume that's why my liver test went up.”

A later question may inherit that framing even though the causal link has not been established for this individual.

PRIOR AI HYPOTHESIS

The model's own earlier sentence remains in context

Earlier AI: “The rise may be related to alcohol.” Later user: “So what is causing my ALT increase?”

An earlier generated hypothesis can become conversational baggage unless its status remains explicit.

Not every history-dependent change is a hallucination. The broader reliability risk is context contamination: verified laboratory facts, user beliefs, generated hypotheses and external evidence can become blurred together unless the system preserves their provenance.

Lab fact	User-reported context	Model hypothesis	External evidence
Measured or reported data	Relevant but not automatically verified	Generated inference, not a patient fact	Source-bound scientific support

Conversation history is context. It is not automatically evidence. A previous AI-generated statement should not silently become a patient fact.

08 · SAFER SYSTEM DESIGN

What governed longitudinal interpretation needs beyond a fluent model

01

Structured input & context provenance

Dates, values, units and metadata - plus whether context is measured data, user-reported information, model inference or external evidence.

02

Deterministic normalization

Known transformations and data-quality checks before inference.

03

Authoritative evidence

Sources chosen for the actual claim and context.

04

Governed relationships

Longitudinal context, discordance and missing information stay visible.

05

Uncertainty & boundary

Confidence can fall and the system can refuse a narrower conclusion.

06

Auditable output

The conclusion can be traced to data, evidence and system rules.

This architecture does not guarantee correctness. It changes the engineering problem from “generate a plausible answer” to “constrain, test and monitor the class of answers the system is allowed to produce.”

09 · SENTINEL

How Sentinel is designed to differ from a general-purpose chatbot

Public design principles

- **Structured health-data inputs:** Sentinel is designed around governed ingestion/normalization rather than treating the report as unconstrained conversational text.
- **Context has provenance:** laboratory facts, user-reported context, generated hypotheses and external evidence should not silently collapse into the same kind of truth claim.
- **Missing information stays missing:** absent context is not silently completed into a story.
- **Time and discordance remain visible:** repeated values and disagreements can alter confidence rather than being flattened.
- **Evidence provenance matters:** the scientific source must fit the claim being expressed.
- **Generative language is not scientific authority:** where generative components are used, they should express governed conclusions rather than create new ones.
- **Handoff is a valid output:** the system can stop when the next useful question requires information, examination or decision-making outside the available data.

Disclosure boundary. This guide describes public product principles. It does not disclose proprietary Sentinel thresholds, internal rules, weighting, scoring, governed logic implementation or security architecture.

10 · VALIDATION AGENDA

What Exia still needs to prove empirically

A health-AI product should not claim reliability merely because its architecture sounds careful.

Citation accuracy

Do cited sources exist, and do they actually support the claim?

Data fidelity

Are units, dates, reference metadata and duplicate/ambiguous records handled correctly?

Missing-data behavior

Does the system keep absent information explicit rather than filling the gap?

Prompt robustness

Do materially equivalent questions produce materially consistent governed conclusions when phrased briefly, colloquially or differently?

Context-contamination resistance

Can an unsupported user belief, irrelevant earlier turn or prior generated hypothesis improperly change a data-grounded conclusion?

Longitudinal consistency

Does the same governed logic behave coherently across multi-time-point cases?

Causal restraint

Does the system avoid crediting interventions merely because a later value changed?

Discordance handling

Does disagreement remain visible instead of being forced into one score or narrative?

Confidence calibration

Does uncertainty language track the actual completeness and strength of the evidence?

Safety and handoff

How often are unsupported or unsafe recommendations produced, and does the system escalate appropriately?

Exia should not claim that Sentinel eliminates hallucination until a defined validation programme has measured the relevant failure modes under controlled conditions.

If you use general-purpose AI with blood results, use it deliberately

Use AI for explanation first

Terminology, organization and question generation are safer starting points than diagnosis or medication decisions.

Verify the numbers

Check units, dates, laboratory reference information and whether the values were extracted correctly.

Ask for sources - then inspect them

A citation is not evidence until the source exists, is relevant and supports the sentence being made.

Provide history when history matters

Longitudinal interpretation needs prior results and timing, not just today's report.

Watch causal language

“Associated with,” “changed after,” and “caused by” are different claims.

Be suspicious of false precision

Very specific personalized recommendations can sound authoritative even when key context is missing.

Know when to leave the chatbot

Urgent symptoms, diagnosis, treatment changes and medication decisions require appropriate professional evaluation.

Preserve your privacy

Before uploading health documents to any service, understand how the service handles data, retention and access.

Selected references

These references support propositions actually used in this guide. They are not intended to represent every study of medical AI.

1 National Institute of Standards and Technology (NIST). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. 2024. Source

2 World Health Organization. Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. 2025 publication page. Source

3 Draelos RL, et al. Large language models provide unsafe answers to patient-posed medical questions. npj Digital Medicine. 2026;9:241. Source

4 Tiller NB, et al. Generative artificial intelligence-driven chatbots and medical misinformation: an accuracy, referencing and readability audit. BMJ Open. 2026;16:e112695. Source

5 Wang D, et al. Enhancing Large Language Models for Improved Accuracy and Safety in Medical Question Answering: Comparative Study. JMIR Medical Education. 2025;11:e70190. Source

6 Lee RW, et al. Vulnerability of Large Language Models to Prompt Injection When Providing Medical Advice. JAMA Network Open. 2025;8(12):e2549963. Source

7 Asgari E, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. npj Digital Medicine. 2025;8:274. Source

8 Turan EI, et al. Assessing the accuracy of ChatGPT in interpreting blood gas analysis results ChatGPT-4 in blood gas analysis. Journal of Clinical Anesthesia. 2025;102:111787. Source

9 Gün M. AI-Assisted Blood Gas Interpretation: A Comparative Study With an Emergency Physician. American Journal of Emergency Medicine. 2025;94:1-2. Source

10 Basu S, Huynh B. Mitigating hallucinations in healthcare AI: a systematic review of evidence-based strategies. BMC Health Services Research. 2026;26:1115. Source

11 NIST. AI Risk Management Framework and Trustworthy and Responsible AI Resource Center. Source

12 World Health Organization. Harnessing artificial intelligence for health. Source

13 Kim TM, Luo L, Kim SE, Manrai AK, Topol E, Rajpurkar P. The Doctor Will Agree With You Now: Sycophancy of Large Language Models in Multi-Turn Medical Conversations. Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026). 2026:19-34. Source

14 Guo K, et al. Stop Listening to Me! How Multi-turn Conversations Can Degrade LLM Reliability. arXiv:2603.11394v3. 2026. Preprint. Source

15 Munnangi M, Savage S. Evaluating Large Language Models on Misconceptions in Multi-Turn Medical Conversations. arXiv:2607.12884. 2026; accepted to MLHC 2026. Source

SCIENTIFIC REVIEW STATUS

Review and validation status

Scientific review status: Exia Bio internal scientific and claims review completed 14 Aug 2026. Version 1.1 adds a bounded prompt/context-provenance section and corresponding validation domains. External clinical/laboratory review has not been performed or represented.

Educational content only; not medical advice. This page is not a clinical validation study, a diagnosis-tool evaluation or evidence that Sentinel eliminates hallucination.