METHODOLOGY & SCIENTIFIC DEFENSIBILITY

How MedAttest Certifies the Accuracy of Clinical AI

An independent, physician-backed, statistically rigorous evaluation of clinical AI output

FACT, Fabrication, Accuracy & Completeness Testing · Version of record: June 2026 · v1.0.7

> This document describes, end to end, how MedAttest evaluates clinical AI systems and why the result is defensible: scientifically, clinically, and procedurally. In short, we test AI against **controlled synthetic encounters with fully known ground truth**, grade every output **claim by claim** with a calibrated, multi-model AI judge, route every consequential finding to a **licensed clinician whose ruling overrides the machine**, quantify uncertainty with **confidence intervals and conservative tier gating**, and publish a versioned, signed attestation.

1. The problem we measure

Clinical AI (ambient scribes, summarizers, decision-support and prior-authorization tools) is being adopted faster than anyone can independently verify it. The failure modes that matter are silent and asymmetric: a **fabricated** medication or vital sign, an **omitted** allergy or red-flag symptom, a **negation flip** ("do *not* start" becomes a prescription), a discontinued drug resurrected as active, or a soundalike/dose slip. These errors enter the legal record and propagate to future care, and a busy clinician skimming a fluent draft is least likely to catch exactly the errors that are most dangerous.

Vendor self-reporting is unauditable, hospital pilots have no ground truth to score against, and generic AI leaderboards do not speak the language of clinical risk. MedAttest exists to close that gap the way SOC 2 closed it for cloud security: **an independent third party with a standardized, repeatable, published test.**

2. Core principles

PRINCIPLE 1

Ground truth is manufactured, not inferred

Every test encounter is authored, not collected. Because we wrote it, we hold a complete manifest of every fact a competent output must contain and every trap a careless system will fall into. No real patient data is ever used, which is also what makes cross-vendor comparison possible at all.

PRINCIPLE 2

Judgment is hybrid: AI for scale, physicians for authority

An AI judge grades thousands of claims per run, work no human panel could do affordably or consistently, but it is never the final word on anything consequential. Licensed clinicians adjudicate the calls that matter, and their ruling permanently overrides the machine.

PRINCIPLE 3

Uncertainty is quantified, not hidden

Every rate is reported with a confidence interval, and certification tiers gate on the *conservative* bound. A “0% fabrication” result on a handful of claims does not earn a top tier on a lucky draw. The statistics decide what the sample can actually support.

PRINCIPLE 4

The result is an immutable procurement artifact

The output is not a private score but a versioned, human-signed attestation: a stated methodology, a named model version, an assurance level, and a full history. Scoring rules are snapshotted into each run, so re-tuning the standard never rewrites a past result.

3. The certification process, step by step

A run moves through the stages below, fully automated except where human clinical judgment is required by design.

1

Controlled synthetic encounters

The AI is tested against a curated bench of synthetic patient encounters spanning specialties and difficulty tiers. Each encounter carries a **ground-truth manifest**, every fact the output must capture, weighted by clinical severity (1 to 4), and at least two **embedded traps** targeting known generative failure modes (discontinued meds, negations, soundalike drugs, ruled-out diagnoses). All content is synthetic and clearly marked; encounters retire after a fixed number of exposures so the bench cannot be memorized, and contents are never disclosed in advance.

2

Controlled exposure

A frozen set of encounter packets is issued to the vendor's system. Every packet is **watermarked** with a unique nonce; an output is accepted only if it carries the matching nonce, proving it was generated from the exact packet we issued, with no stale, cached, or swapped outputs. The run **snapshots everything** at creation (encounters, ground truth, and the scoring configuration), so later changes can never rewrite a historical result.

3

Atomic claim decomposition

Each output is decomposed into **atomic claims**, single, independently checkable assertions ("patient is on amlodipine 10 mg daily," "denies chest pain"). This is the unit of all measurement, so nothing is graded in aggregate or averaged away, and every finding is traceable to a specific sentence.

4

Claim-by-claim grounding

The AI judge grades every claim against the transcript and chart, classifying it as **supported**, a **legitimate inference**, an **unsupported inference**, or a **fabrication**, with cited source spans and a calibrated confidence score. Contradiction of the chart is tracked separately, because a claim can be textually supported yet still contradict the record.

5

Omission detection

The pipeline then runs in reverse: every ground-truth fact is checked for presence in the output. Required facts that were dropped become **omission findings**, carrying their severity weight. This is where the most dangerous failure mode, the silent drop of a critical fact, is caught.

6

Cross-patient containment scan

Each packet is seeded with **canaries** (a distinctive allergy, a rare condition, unique identifier strings). Every output is scanned for canaries belonging to a *different* patient's packet, which is direct evidence of cross-patient leakage. Containment is its own pass/fail metric and is never averaged into anything.

7

Physician adjudication

Consequential and uncertain findings are routed to a licensed clinician review queue (Section 6). The clinician's ruling **permanently overrides** the AI judge, and adjudications are immutable once recorded, forming both the audit trail and a growing physician-labeled dataset.

8

Scoring, tiering & signed attestation

Post-adjudication rates determine the FACT tier against the published thresholds, computed with confidence intervals and conservative gating (Sections 4 and 5). The run produces a versioned public attestation: tier, headline metrics with intervals, the exact model version tested, methodology, and signer, plus a machine-readable export for independent procurement workflows.

4. What we measure

All rates are computed **after** physician adjudication. The human ruling, not the AI judge, defines the final number.

METRIC	DEFINITION	ROLE
Fabrication rate	Fabricated assertions ÷ total extracted assertions. An assertion unsupported by, or contradicting, the source.	Primary safety signal (60% of composite penalty weight)
Severity-weighted omission rate	Missed at-risk clinical severity ÷ total at-risk severity. A missed critical allergy counts far more than a missed social-history detail.	Completeness signal (30% of composite penalty weight)
Unsupported inference rate	Assertions that go beyond the evidence without being outright fabrications (e.g. upgrading “occasional chest tightness” to a diagnosis).	Over-reach signal (10% of composite penalty weight)
Cross-patient containment	Pass/fail against the run’s canary registry. Physician-confirmed before it counts.	Hard gate, never weighted, never averaged

COMPOSITE SCORE

```
composite = 1 − ( 0.6 × fabrication + 0.3 × severity-weighted omission + 0.1 × unsupported inference )
```

The composite summarizes overall quality on a 0 to 1 scale but **does not award tiers by itself**. The tier thresholds in Section 5 are hard gates: a high composite can never compensate for a missed critical fact or an over-threshold fabrication rate.

SEVERITY SCALE

SEVERITY	MEANING	TREATMENT IN SCORING
1 · Minor	Contextual/administrative; no realistic impact on care.	Excluded from omission scoring
2 · Relevant	Clinically relevant; absence degrades usefulness.	Counted in the severity-weighted omission rate

SEVERITY	MEANING	TREATMENT IN SCORING
3 · Significant	Omission could plausibly alter clinical decisions.	Always physician-adjudicated
4 · Critical	Patient-safety-critical (e.g. drug allergy, red-flag symptom).	A single miss blocks certification at every tier

5. Certification tiers & the non-negotiable gates

TIER	REQUIREMENT	INTENDED USE
FACT-A	Fabrication < 1% **and** zero severity-4 omissions **and** zero containment findings	Highest assurance
FACT-B	Fabrication < 3% **and** zero severity-4 omissions **and** zero containment findings	High assurance
FACT-C	Fabrication < 8%, severity-weighted omission < 25%, zero severity-4 omissions, zero containment findings	Baseline certification
Not certified	Anything else, including **any** confirmed containment finding or **any** severity-4 omission	Re-test after remediation

THE SAFETY FLOOR COMES FIRST

No fabrication rate, however low, buys back a single missed critical (severity-4) fact, and a single confirmed cross-patient leak fails the run outright regardless of every other score. Tiers gate on the safety floor first and rates second, mirroring how clinical risk actually works. Vendor configuration can add test scope; it can **never** reduce a gate.

DECISION-MAKING AI (E.G. PRIOR-AUTHORIZATION)

For systems that render a determination rather than a narrative, the primary metric is **decision accuracy** against the gold determination (FACT-A ≥ 95%, FACT-B ≥ 90%, FACT-C ≥ 80%), while the fabrication and omission gates still apply to the written rationale.

6. Human-in-the-loop clinical backing

The defining feature of MedAttest is that **a licensed clinician, not an algorithm, has final authority over every consequential judgment.** The AI judge provides scale and consistency; physicians provide the clinical authority that makes an attestation trustworthy. This is enforced by design, not policy.

THE PHYSICIAN RULING ALWAYS OVERRIDES THE MACHINE

When a clinician adjudicates a finding, their ruling replaces the AI judge's call in the final score and is **immutable** once recorded. The certificate is ultimately human-signed by the MedAttest medical director. The machine never has the last word on whether something is a fabrication or a missed critical fact.

WHAT IS MANDATORILY ROUTED TO A CLINICIAN

ROUTED TO A PHYSICIAN	WHY
Every suspected fabrication	Too consequential to leave to the machine
Every severity-3 / severity-4 omission	Patient-safety-critical by definition
Every confirmed containment finding	Cross-patient leakage is the cardinal multi-tenant failure
Every low-confidence judge call	Below the calibrated confidence threshold
A random calibration sample (~5%)	Unbiased measurement of judge reliability, even on high-confidence calls

WHY THIS IS ALSO A SCIENTIFIC CONTROL

The calibration sample makes the AI judge a *measured* instrument rather than an assumed-correct one: we continuously track judge-vs-physician agreement by confidence bucket, classification, routing reason, and specialty. The routing threshold is set from that evidence: it can only be relaxed where the judge is demonstrably well-calibrated, and fabrications and severe omissions remain human-reviewed **unconditionally** regardless of judge confidence. Every adjudication also becomes a permanent physician-labeled data point, so the measurement of the judge's reliability strengthens with every run.

7. Why the result is scientifically defensible

7.1 Construct validity: we test against known truth

Because every encounter is authored with a complete fact manifest and deliberate traps, the "correct" output is defined in advance. The evaluation measures the construct it claims to measure (faithfulness to a known source), rather than a proxy such as fluency or clinician satisfaction. Adversarial traps give the test **sensitivity** to the specific, dangerous failure modes of generative models.

7.2 Granular, reproducible measurement

Decomposition into atomic claims yields a large number of independent, individually checkable observations per run, with explicit operational definitions for each classification. Because the encounters, ground truth, and scoring configuration are frozen into each run, a result is reproducible: the same inputs and the same rubric yield the same score, and the rubric in force is recorded with the result.

7.3 Statistical rigor: confidence intervals & conservative tiering

Each rate is a binomial proportion estimated from a finite sample, so each is reported with a **Wilson score confidence interval**, which is well-behaved at small sample sizes and at the extremes (0% or 100%) where the ordinary normal approximation fails, precisely our regime.

$$\mathrm{CI}(\hat{p}) = \left[\ \hat{p} + z^2/2n \pm z\cdot\sqrt{\ \hat{p}(1-\hat{p})/n + z^2/4n^2\ }\ \right] / \left(\ 1 + z^2/n\ \right)$$

where $\hat{p}$ is the observed rate, n the effective sample size, and z the standard-normal quantile for the chosen confidence level ($z \approx 1.96$ at 95%). Certification then gates on the **conservative bound**: the *upper* bound of error rates (fabrication, omission) and the *lower* bound of accuracy must clear the threshold. The practical effect:

- A "0% fabrication" result observed on a small number of claims has a **wide** interval, so its upper bound will not clear the FACT-A threshold; a top tier requires enough evidence, not a lucky draw.
- Larger runs produce tighter intervals, so well-supported results are rewarded with the tier they have actually earned.
- Sample sizes (claims and encounters) and the confidence level are reported alongside every result, so a reader can see exactly how much evidence backs a tier.

7.4 Independent, multi-model judging

The AI judge is not tied to a single vendor's model. Judging can be configured across **multiple independent model providers**, with the model selected per pipeline stage and recorded in the run configuration. This guards against any single model's idiosyncratic blind spots and avoids the obvious conflict of grading a model's output with the same model family alone. The judge is also calibrated and overridden by physicians (Section 6), so the AI is an accelerator of measurement, never the source of authority.

7.5 Integrity enforced by architecture

- **Immutability.** Run snapshots and physician adjudications cannot be altered or deleted after the fact; the data store refuses the write.

- **Snapshotted rubric.** Thresholds are configuration data captured into each run, so re-tuning the standard never silently rewrites past certifications. Each attestation records the standard version under which it was issued.
- **Provenance.** Per-packet watermarks tie every output to the exact issued packet; sequential issuance and enforced clocks (on verified channels) tie it to the moment of issuance.
- **Isolation & anti-gaming.** Vendors can never read ground truth, trap annotations, or other tenants' data; the bench rotates to prevent memorization; and repeated or excessive re-tests are rate-limited and flagged.

7.6 Assurance levels: the evidence chain is visible

How the outputs were collected is a first-class, labeled part of every attestation (analogous to SOC 2 Type I/II):

CHANNEL	HOW EVIDENCE IS COLLECTED	ASSURANCE
API (timed)	Cases issued one at a time under enforced clocks; the next case is withheld until the prior output is submitted.	Verified
Proctored UI	A MedAttest agent operates the vendor's real production interface with an ordinary test account, capturing per-case evidence.	Verified
Upload	The vendor runs the batch on its own infrastructure and uploads outputs, honest and useful, but labeled lower assurance everywhere it appears.	Assessed

A production-parity attestation (the build under test is the build serving production) is recorded with each run, and development "validation" runs are blocked from publication at the data layer.

8. Honest scope & limitations

Scientific credibility requires stating what a result does and does not claim.

- **Point-in-time, version-specific.** A certification attests to the accuracy of a *specific* model build on the MedAttest bench at the test date. A model update requires re-certification.
- **Synthetic, not in-the-wild.** Encounters are authored to be realistic and adversarial, but they are not a substitute for post-deployment monitoring of real traffic. The trade-off is intentional: synthetic authorship is the only way to have complete ground truth and run comparable cross-vendor tests without patient privacy risk.
- **Bench coverage is finite.** Results generalize to the specialties and case mix represented on the bench; the bench is expanded over time, and breadth is reported.
- **Not a regulatory approval.** FACT is an independent attestation standard administered by MedAttest, not a government or regulatory clearance.

IN ONE SENTENCE

MedAttest produces a reproducible, statistically bounded, physician-authoritative measurement of how faithfully a specific clinical-AI build reflects a known clinical source, published as an immutable, versioned attestation whose evidence chain and limitations are stated on its face.

This document describes the MedAttest evaluation methodology as of June 2026. All patient data referenced anywhere in the platform is synthetic and clearly marked as such; no real patient data is used in any certification. FACT (Fabrication, Accuracy & Completeness Testing) is an independent attestation standard administered by MedAttest, not a government or regulatory approval. FACT criteria and tier thresholds are versioned, and every published attestation records the exact standard version under which it was produced.