EON AI Ventures

The Shift Test

An open benchmark for industrial safety judgment

Why the obvious way of measuring industrial AI does not work, what to do instead, and how a score produced by the party who benefits from it can nevertheless be trusted.

Table of Contents

PART ONE: Plain English

1. The thing nobody can measure

Every company selling artificial intelligence into heavy industry currently makes some version of the same claim: that its system understands industrial operations. Not one of them can demonstrate it, and no buyer can check it, because there is nothing to check it against.

This is not a small gap. It is the reason technical evaluations take twelve months, the reason procurement decisions come down to which vendor the operations director already trusts, and the reason an investor looking at any company in this category cannot distinguish real capability from a convincing description of one.

The gap, stated once

Industrial AI has no equivalent of a crash test, a fuel economy figure or a drug trial. It has demonstrations. A demonstration proves that something worked once, in conditions chosen by the person demonstrating it.

The Shift Test is an attempt to close that gap with a single narrow number that anybody can reproduce.

2. What a benchmark actually is

Car manufacturers publish one figure: the time taken to accelerate from a standstill to sixty miles an hour. Anyone can measure it. Anyone can dispute it. Everybody understands it.

They do not publish a figure for the ownership experience — which is the thing customers actually buy, and which is far more important. They do not publish it because it cannot be measured, and an unmeasurable claim persuades nobody.

The number makes people believe the engineering. The engineering is what they buy.

A benchmark is a single, narrow, checkable question. Its authority comes entirely from being narrow — and from the fact that its author could lose.

What the Shift Test deliberately does not measure

It does not measure Genesis, Universal, TASI, the Situation Room, or anything that could be described as the EON ecosystem. Those are the answers to “how did you achieve that score?”, and they are what a customer actually buys. Benchmarking them would produce a number nobody outside EON could verify, which is the same as producing no number at all.

3. Why the obvious test does not work

Ask anyone how to test this and you will get the same answer within about ten seconds: take real industrial incidents, hide what happened, and see which system works it out. It is the right instinct. It does not survive contact with how these systems are built.

The systems have already read the incidents

Large general-purpose models are trained on very large collections of text whose contents are not published. They are retrained on a continuing basis. This means an evaluator cannot establish that a given incident was absent from a given system's training material, and cannot verify any cutoff date a vendor states.

Every well-documented industrial accident of the last forty years is almost certainly in the training material of every serious system. Asking about one measures memory and reasoning together, in a proportion that cannot be determined.

Using recent incidents only postpones the problem

The natural repair is to use incidents from the last few months. But retraining happens on a cadence of weeks, the window keeps closing, and its width can never be confirmed. A benchmark whose validity depends on a margin you cannot measure is not a benchmark; it is a hope with a number attached.

This is worth being blunt about, because it was our own first design. We built a benchmark on a date split, presented it, and had it correctly taken apart. The version described here is the replacement.

4. What aviation worked out fifty years ago

Airlines do not certify pilots by examining them on historical crashes. They put them in a simulator and give them a situation they have never seen. Nobody has ever suggested this is unfair. It is the most trusted evaluation method in safety-critical work anywhere, and it has been for half a century.

We test AI the way airlines test pilots.

Real physics. New situations. No answer sheet to memorise.

Three properties make the simulator legitimate, and all three carry over exactly:

1. **Real physics.** The equipment behaves the way real equipment behaves. Nothing is invented about how the world works — only about which particular things are happening at once.
2. **New situations.** The specific combination has not occurred before, so there is no past case for anyone or anything to recall.
3. **No answer sheet.** Because it never happened, nothing was ever written about it. There is nothing to memorise. The answer has to be worked out.

And the reason it is us who does this

EON has built simulators of industrial work for twenty-five years. Testing an AI system inside one is not a clever new idea. It is the obvious use of the thing we already have, and it is the one part of this that a competitor cannot assemble quickly.

5. Why it is called the Shift Test

A shift is the unit of industrial work. Eight hours, twelve hours — the stretch of time during which one crew is responsible for a plant. Everything that goes wrong, goes wrong on somebody's shift.

And the question the benchmark scores is exactly the question a crew asks at handover:

Given only what we can see right now, what is most likely to hurt us before the next shift ends — and what check would confirm it?

So the test asks a model to do one shift's worth of thinking: the thinking a good engineer does at handover, which is the most valuable and least documented thinking in the industry.

The second meaning is intended. This is the test that shifts how industrial AI is measured.

Names considered and rejected

Name	Why not
The 48	Hard-codes the forty-eight hour horizon into the name. The horizon will change with the equipment class; the name would then be wrong.
FORESIGHT / PRECOG	Promises prediction. The benchmark does not predict — it ranks what is most likely developing. Overclaiming in the name invites the exact attack we are trying to survive.
SIM-1	Accurate, neutral, and forgettable.
IPB-1 (Industrial Precursor Benchmark)	Our own earlier name. “Precursor” requires a paragraph of explanation before anyone can use the word in a meeting, and a name that has to be defined does not spread.

6. Why you can trust a score produced by the company that benefits from it

You should not simply take it on trust, and we are not asking you to. There are three checks, and the useful property of all three is that none of them involves a model at all.

Check one — working engineers cannot tell our situations from real ones

We mix generated situations in with situations taken from the real incident record, and give the mixture to practising process safety engineers with one instruction: mark which are which.

If they cannot do better than guessing, then as far as expert human judgment can determine, our situations are indistinguishable from reality. If they can, the set is thrown out and the generator is rewritten. No AI is involved anywhere in this check — it is engineers, a list, and a coin-flip threshold.

Check two — humans write the answers, not the simulator

The answer key is authored by domain experts working from the situation alone. It is not generated, and it is not read off what the simulation did next.

Where the experts and the simulation disagree about what was developing, we publish the disagreement rate rather than quietly resolving it in favour of whichever we prefer. A simulation that frequently disagrees with expert judgment is not a suitable basis for generating anything, and the disagreement rate is how you find that out.

Check three — the ranking survives the switch to real cases

Every contestant is scored twice: once on generated situations and once on situations from the documented record. If the order of finish is the same in both, and the scores track each other, the two sets are measuring the same ability.

This check has a property worth stating plainly. Most of the contestants are systems we did not build, cannot configure and cannot tune. We are not able to arrange for this check to pass. That is the entire reason it is worth anything.

The criticism that is simply correct

There is no test whose author cannot influence it. Anybody who tells you otherwise is selling something.

What can be done is to make every place a hand could go visible to everyone else: publish the construction method, publish every score including our own losses, use answer keys written by people who do not work for us, release a public sample anyone can run, and have the real scoring set held and administered by a neutral party. Each of those removes one place a hand could go. That is the whole defence, and we offer it as sufficient rather than as perfect.

7. The claim we are making — and the one we are not

We are not claiming to be smarter than the best model in the world.

We are claiming that a model which studied industrial accidents beats one that studied the whole internet — at understanding industrial accidents.

That is the only claim we make. The benchmark either shows it or it does not.

A claim this modest is far harder to attack than a grand one, and it has the additional advantage of being the true one. It is also falsifiable in a single afternoon by anybody who wants to try, which is the property that makes it worth stating at all.

Losing narrowly would still be a good outcome

If a frontier model reasons more elegantly in general terms but cannot cite a real comparable case, cannot state honest confidence and cannot bring itself to say “I do not know”, then it has failed at the three things that determine whether an engineer on a night shift can act on what it said. We would report that as a loss on the headline and a win on the dimensions that matter, and we would be telling the truth.

8. Seven objections, and what to say

These are the objections we expect, in the order we expect them. The answers are written to be said out loud.

Objection	Answer
“You made up the questions, so of course you win.”	We did not write the answers — engineers did. And engineers cannot tell our questions apart from real ones. Both of those are checkable without involving us.
“Simulated situations are not real situations.”	The physics is real; only the combination is new. That is exactly what a flight simulator is, and no airline has ever accepted that objection.
“You are grading your own exam.”	The full scoring set is held and run by someone else. We publish every score, including the ones we lose.
“A frontier model will beat you.”	It might, on general reasoning. Our claim is about citing real cases, honest confidence and knowing when to stay quiet. Look at the dimensions, not the headline.
“Why should anyone care about your benchmark?”	You should not care about ours. You should care about having one at all. If somebody builds a better one we will run our model on it and publish that too.
“You could just tune your model to the test.”	The scoring set is not ours to see. A public sample exists precisely so you can check whether the sample and the real set behave the same way.
“There is no score yet.”	Correct. That is why this document makes no performance claim anywhere. Ask us again after it has been run.

9. What a good score is worth

Our own data rights work named the underlying problem precisely: an investor cannot verify that this company owns anything real, and that gap is the difference between one valuation and a much larger one.

A published benchmark is the instrument that closes exactly that gap. It converts “we have deep industrial knowledge” — true, and unverifiable — into “here is a number, here are the questions, run it yourself.” Four things follow.

What changes	Why
Category position	If buyers begin asking for a Shift Test score by name, every competitor is measured on terms we defined. That is structural rather than temporary.
Diligence	An externally reproducible score is the clearest available evidence that the asset is real. It converts “trust us” into “check it.”
Procurement	A supermajor's technical evaluation team can run the benchmark themselves, which turns a twelve-month evaluation into an afternoon.

What changes	Why
Pricing power	We stop selling training software against other training software and start selling the only measured system in a category we defined.

The honest caveat

A benchmark generates no revenue by itself. It removes the reason a buyer or an investor says no. Do not confuse the two — and do not underestimate how often that is the only thing standing in the way.

PART TWO: The Method

10. Formal statement of the contamination problem

Let S be a system under evaluation and let its training corpus be C_S. For an instance x drawn from the documented incident record, the observed score decomposes into a reasoning component and a recall component:

```
   score(S, x)  =  reason(S, x)  +  recall(S, x)

   where recall(S, x) > 0  whenever  x is represented in C_S

# The evaluator's problem is not that recall exists.
# It is that membership of x in C_S is NOT DECIDABLE by the evaluator:

     C_S is undisclosed
     C_S is revised on a continuing cadence
     any stated cutoff for C_S is unverifiable by a third party

# Therefore recall(S, x) is not merely unknown but UNBOUNDED,
# and score(S, x) does not identify reason(S, x) at all.
```

Selection procedures over historical material cannot repair this. Restricting to recent events reduces the prior probability of membership but does not bound it, and the bound that is being appealed to shrinks monotonically with retraining cadence. The defect is a property of the instance source, not of the selection criterion applied to it.

The only construction that eliminates the term is one in which membership is decidable by construction: an instance that did not exist in any corpus at the time C_S was assembled, because it did not exist at all.

11. Instance generation under temporal truncation

Instances are generated from a simulation of a physical asset, not selected from record. The generation procedure is as follows.

```
for each instance:

  1. initialise asset state from an equipment class configuration
  2. execute the simulation forward to time t
  3. TRUNCATE the observable record at t
  4. retain trajectory (t, t+delta] as ground truth,
     WITHHELD from the instance
  5. emit instance = observable record on [t-w, t]
```

```
# The constraint is applied AT GENERATION, not by redaction.
# Information established after t was never present in the instance,
# so its absence is a property of construction rather than a claim
# about the completeness of a removal step.
```

The distinction in the final comment matters more than it appears to. Redaction of a record in which post-cutoff information was present is verified by inspection and may be incomplete; contamination that survives redaction is invisible and improves every metric available at development time. Generation under truncation has no corresponding failure mode.

11.1 The horizon

The observation window w and the outcome horizon delta are configured per equipment class. The working default for rotating equipment and pressure systems is a horizon on the order of forty-eight hours, chosen because it is the interval over which a maintenance intervention can realistically be scheduled. The horizon is a parameter of the instance set and is published with the results; it is not part of the benchmark's identity, which is why it is not part of its name.

11.2 Instances derived from documented events by parameter alteration

A second construction is used for part of the set. A documented event is taken and specified parameters are systematically altered according to a published procedure, such that the correct response differs from the response documented for the original event.

This has a useful adversarial property. A system that recognises the original event and reproduces its documented conclusion scores worse than a system that does not recognise it at all. Recall becomes a penalty rather than an advantage, which is the inverse of its effect on a conventional historical benchmark and a direct probe for the contamination the benchmark exists to exclude.

12. Admissibility by blinded human discrimination

A generated instance set is not admitted on the assertion of its authors. It is admitted on the outcome of a procedure in which no model participates.

```
PROCEDURE  discrimination

  present to n practising process safety engineers a set of
  2m instances, of which m are generated and m are drawn from
  the documented record, interleaved in randomised order
```

```
  task: label each instance GENERATED or DOCUMENTED

  let a = mean labelling accuracy across raters

  ADMIT     the set if a is not distinguishable from 0.5
            at the configured significance level
  REJECT    otherwise; revise the generation procedure
            and repeat with a fresh panel

# raters are blind to which party produced which instance
# raters are compensated identically regardless of outcome
# no model is involved at any point in this procedure
```

12.1 Why this is not circular

The objection to any machine-generated evaluation is that it measures the agreement of correlated systems. The three admissibility signals used here are chosen specifically so that none of them derives from the generator:

Signal	Source	Can EON influence the outcome?
Human discrimination accuracy	Practising engineers, blind, no model involved	No — the panel is external and the threshold is fixed before the panel sits
Reference response	Authored by domain experts from the truncated instance alone	No — experts are not shown the withheld trajectory or the generator's output
Rank-order preservation	Computed over contestants including systems EON did not build	No — EON cannot configure or tune the comparison contestants

The strongest form of the circularity objection, and its answer

The strongest form is not “a model made the questions.” It is: the generator and the judge may share a systematic blind spot, so a defect invisible to one is invisible to the other, and the evaluation is silent about a whole class of error.

The discrimination check is the answer, and it is the answer because it is the one component with no model in it. A shared blind spot between generator and judge is not shared with a human process safety engineer, whose failure modes are different in kind. If the generated instances carry a systematic artefact, the engineers are the population most likely to notice it — that is what expertise is.

13. Reference responses

The reference response against which contestants are scored is authored by human domain experts working from the truncated instance alone. It is not produced by the generator, and it is not read off the withheld trajectory.

This is a deliberate cost. Reading the answer off the simulation would be free, instantaneous and perfectly consistent — and it would make the benchmark a test of agreement with our simulator rather than a test of engineering judgment. Experts are slower, more expensive and less consistent, and they are the only source of a key that means anything.

13.1 Recording divergence rather than resolving it

Where the expert key and the withheld trajectory disagree about the developing mechanism, the divergence is recorded as a property of the instance and published. It is not resolved in favour of either.

A high divergence rate is a finding, not an embarrassment. It indicates either that the simulation is not a suitable basis for generation in that equipment class, or that the situation is genuinely ambiguous to expert judgment — and the second of those is itself a result worth publishing, since it bounds what any system could achieve.

13.2 Inter-rater agreement

Agreement between expert authors is measured and published. Expert disagreement is expected and is not a defect in the benchmark; concealing it would be. A benchmark reporting perfect expert agreement on ambiguous industrial situations should be assumed to be measuring something other than what it claims.

14. Scoring

Scores are reported per dimension. They are not collapsed into a single figure, because the dimensions fail independently and a composite would conceal exactly the differences the benchmark exists to expose.

Dimension	What is scored	Graded by
Identification	Whether the developing mechanism is correctly named	Expert against reference key
Evidence	Whether a real, checkable comparable case is cited — and whether the cited case exists	Deterministic lookup, then expert
Actionability	Whether the recommended check would in fact confirm or exclude the hypothesis	Expert

Dimension	What is scored	Graded by
Calibration	Agreement between stated confidence and observed accuracy across the set	Deterministic
Abstention	Whether the system declines where evidence does not support a conclusion	Expert against a pre-marked ambiguous subset
False alarms	Rate of asserted hazards on the quiet track	Deterministic

The Evidence dimension includes a deterministic first pass: a cited case either exists in the public record or it does not. A system that fabricates a plausible citation fails that pass without a human ever having to read it, and fabricated citations are the single most common failure mode of general models on this task.

14.1 The quiet track, and why it dominates

A separate track of instances in which nothing is developing is scored alongside the main set. Performance on it is not a secondary consideration; on any realistic base rate it is the dominant term in whether a system is deployable at all.

```
Let  pi   = prior probability that a given asset-shift contains
            a developing serious hazard
     sens = sensitivity      spec = specificity

     precision = (sens * pi) / (sens * pi + (1 - spec) * (1 - pi))

Take a generous case:  pi = 1e-4,  sens = 0.99,  spec = 0.99

     precision = (0.99 * 1e-4) / (0.99 * 1e-4 + 0.01 * 0.9999)
               ~ 9.9e-5 / 1.0098e-2
               ~ 0.0098

# Roughly ONE PERCENT of alarms are real.
# About a hundred false alarms for every genuine catch,
# from a system that is 99% accurate on both axes.
```

This is not a criticism of any particular system; it is arithmetic that applies to all of them. It is also the reason a benchmark without a quiet track is actively misleading: it would rank systems by a quantity that has almost no bearing on whether the resulting alarms are worth reading. A warning system that cries wolf is switched off by the people it was built to protect, and then its sensitivity is irrelevant.

The practical consequence for scoring is that specificity is reported at a fixed operating point, and each system's own abstention behaviour is measured rather than tuned by the evaluator.

15. Contestants, blinding and administration

Contestant	Purpose
The EON model	The system whose claim is under test
A leading general-purpose model	The comparison that matters commercially. Run at its published best configuration, not handicapped
A second general-purpose model of a different family	Guards against a result that is an artefact of one vendor's idiosyncrasies
A panel of human process safety engineers	The anchor. Without it, all machine scores float — the number that makes the others interpretable

15.1 Blinding

Responses are stripped of formatting, characteristic phrasing and any self-identification before grading. Graders are not told which response came from which contestant, and the human panel's responses are blinded on the same terms as the machines'. Style is scored separately from substance so that a grader's preference for a particular register does not leak into the substantive dimensions.

15.2 Neutral administration, and why the scoring set is not published

The complete scoring set is held and executed by a party that is not EON. A public sample is released; the full set is not.

The reason is not secrecy. A published test set is absorbed into the next training cycle of every system it is used to evaluate, at which point it stops measuring anything. Holding the scoring set is what makes the benchmark repeatable in a year's time. The public sample exists so that anyone can verify the sample and the held set behave alike.

16. Cost and schedule

Item	Cost
Pilot: 20 instances, frontier baseline, tractability check	~$100
Build 300 instances via the agent panel, human-audited sample	~$400
Adversarial quiet track, 150 instances	~$150
Run four contestants across all 450 instances	~$300

Item	Cost
Total, excluding the expert panel	~$950
Expert panel — eight engineers for one day	Internal, or $10–15K external

Sequencing: the pilot comes first

The twenty-instance pilot costs about a hundred dollars and takes two days. It establishes the frontier baseline and whether the task is tractable at all before anything larger is committed. If the frontier models already score near the expert panel on the pilot, the benchmark as designed has no headroom and should be redesigned rather than built.

Run the small bet before the large one. The pilot is cheap enough that not running it is the more expensive decision.

The benchmark costs roughly what the model costs, and it is the more durable of the two. It survives every model change — ours and everyone else's — and continues to function as an asset long after the first Delta has been replaced twice over.

17. Risks and failure modes

Risk	Likelihood	Response
A frontier model wins overall	Real	Report per-dimension scores. Losing on general reasoning while winning on citation, calibration and abstention still establishes the claim.
Experts disagree with each other	Likely	Publish inter-rater agreement. Expert disagreement is a genuine finding about the difficulty of the task.
Generated instances look artificial	Moderate	The discrimination check catches this before publication. If engineers spot the generated ones, the set is not ready and nothing is published.
Expert key diverges from simulation	Moderate	Publish the divergence rate. A high rate disqualifies that equipment class from generation rather than being resolved away.
Someone builds a benchmark first	Low, rising	Speed. This is the reason the benchmark precedes the model.
We are accused of grading ourselves	Certain	Neutral administration, per-dimension publication, expert-authored keys, public sample. Design for this one, because it will be made regardless.

17.1 What we would accept as refutation

A benchmark that its author could not be shown to be wrong about is marketing. These are the outcomes we commit in advance to publishing as failures:

- Practicing engineers distinguish generated from documented instances at better than chance. The instance set is unusable and no score derived from it is published.
- Rank order is not preserved between generated and documented sets. The generated set is not measuring the same capability and the design is wrong.
- Expert inter-rater agreement is so low that the reference key cannot bear weight. The task as specified is not gradeable and must be narrowed.
- The EON model loses on identification, evidence, calibration and abstention simultaneously. The central claim is false and we will say so.

The sequencing decision, restated

The benchmark is built before the model is finished. If the model came first we would design a benchmark it passes — probably without noticing, which is what makes the risk serious rather than merely possible.

Design the exam while you still do not know the answer, and the number means something to you as well as to everybody else.

Appendix A — The twenty-instance pilot

Recommended before any further commitment. Two days, roughly one hundred dollars.

Step	What happens	Output
1	Generate 20 instances in a single equipment class under temporal truncation	20 truncated instances plus withheld trajectories
2	Two internal engineers author reference responses from the truncated instances alone	20 keys, plus an inter-rater agreement figure
3	Run two frontier models and the internal engineers on all 20	Baseline scores per dimension
4	Informal discrimination check: 20 generated mixed with 20 documented, shown to 3 engineers	A first read on whether the instances pass as real
5	Decision	Proceed, redesign, or abandon

What the pilot is actually testing

Not whether EON wins — there is no EON model yet. It tests three things: that the instances survive expert inspection, that experts can agree well enough for a key to exist, and that the frontier baseline leaves headroom.

If frontier models already perform at expert level on the pilot, the honest conclusion is that this benchmark has nothing to measure and the design should change. That would be an uncomfortable finding worth a hundred dollars.

No performance figure, score or comparison appears anywhere in this document. At the time of writing the Shift Test has not been run and EON makes no claim about how any system performs on it, including its own.