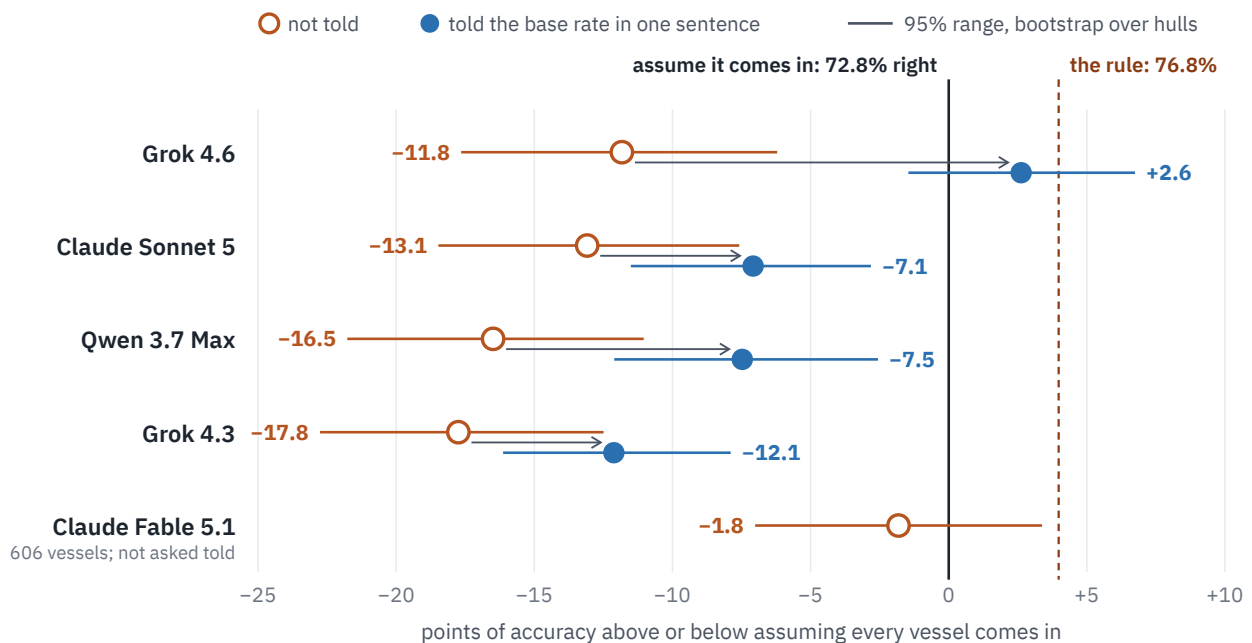


THE GAGE INSTITUTE · MARITIME NO. 1

# AI on the Night Watch

Frontier AI models were asked about 1,031 real night stops off U.S. coasts, ten years of them, and every answer was graded against what the vessel did next. Asked cold, every model leaned to suspicion. One sentence of scored history improved every model that received it. Behind the watch officer's own rule, the AI cut false alarms.



Each model's accuracy above or below assuming every vessel comes in, not told and then told the record's base rate in one sentence. Grok 4.6 moved from -11.8 to +2.6 points. Lines are 95% ranges. Sealed before scoring: journal head 8335d83abf1b734e.

**PUBLISHED BY**

The Gage Institute  
 A program of Gage Systems LLC  
 frontiercolosseum.com

**CONTACT**

Shane Gage, Founder  
 gagesystems2026@outlook.com

ABOUT THIS BRIEF

Document control and contents

|                |                                                                                                                              |
|----------------|------------------------------------------------------------------------------------------------------------------------------|
| Title          | AI on the Night Watch                                                                                                        |
| Series         | ARCS Study Brief · Maritime No. 1                                                                                            |
| Version        | 1.0 · 29 October 2026                                                                                                        |
| Distribution   | Public. Built only from public data. Names no vessel. Free to reuse under CC BY 4.0, with attribution to the Gage Institute. |
| Published by   | The Gage Institute, a program of Gage Systems LLC · Shane Gage, Founder · gagesystems2026@outlook.com                        |
| Pre-registered | 28 October 2026, 21:08 Eastern, before any model was asked                                                                   |
| Run            | 28 October 22:34 to 29 October 03:49 Eastern, 390 calls, sealed in a hash-chained journal                                    |
| Journal head   | 8335d83abf1b734edc2a327654c8473e19d26b043d0819c87ad1a8bfd2c98bc8                                                             |
| Record         | frontiercolosseum.com, free to reuse under CC BY 4.0                                                                         |

Key figures

|                                                                                                                                                                |                                                                                                                                         |                                                                                                                                       |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------|
| <div>1,031</div> <div>real night stops, ten years, each graded against what the vessel did next</div> <div>778 hulls off Texas, Louisiana and California</div> | <div>+14.4</div> <div>points gained by Grok 4.6 from one sentence of scored history</div> <div>every model told it: +5.6 to +14.4</div> | <div>17 of 22</div> <div>alerts cut by the AI veto that were false alarms</div> <div>+2.0 points over the watch officer's rule</div>  |
| <div>93%</div> <div>right when sure: Claude Fable 5.1, claiming 86%</div> <div>four others: 56 to 68% right, claiming 87 to 91%</div>                          | <div>1.9</div> <div>independent opinions from four AI models</div> <div>all four wrong together 67 times, 12 expected</div>             | <div>16 of 42</div> <div>batches of routine vessel data one model declined</div> <div>each in under three seconds, with no text</div> |

Contents

|    |                                                     |    |
|----|-----------------------------------------------------|----|
|    | Summary: what was tested, and what it found         | 3  |
| 1  | The test                                            | 4  |
| 2  | Asked cold, the models leaned to suspicion          | 5  |
| 3  | One scored number moved every model                 | 6  |
| 4  | Rule first, AI second                               | 7  |
| 5  | Score each model's confidence against what happened | 8  |
| 6  | Four models are not four opinions                   | 9  |
| 7  | Not answering is a result too                       | 10 |
| 8  | What a watch floor should do                        | 11 |
| 9  | Score every signal, and give the AI the scores      | 12 |
| 10 | Method and limits                                   | 13 |
| 11 | The record                                          | 14 |
|    | About the Gage Institute and Gage Systems           | 15 |

SUMMARY

# What was tested, and what it found

## Purpose

When a frontier AI model judges a vessel, how often is it right against what the watch already does, and what makes it better? Tested on 1,031 night stops, 2015–2024, under a design recorded, with hashes, before any model was asked.

## Where the question comes from

A real drug drop. On 3 February 2015 a patrol aircraft saw a coastal freighter “hovering outside U.S. territorial waters” off Puerto Rico and launching a small boat that ran for shore without lights; the boat was stopped with 989 kg of cocaine (U.S. Customs and Border Protection, 10 February 2015). Her public AIS track shows the stop this study asks about: under 3 knots for 12.8 minutes, 12.7 nautical miles off U.S. land, in full dark. Off Texas, western Louisiana and California over ten years, nearly three in four of the 1,031 vessels that made that stop were inside the line within six hours: the pattern is mostly ordinary, which is why an AI that assumes the worst is costly.

## Bottom line

| FINDING                                                                                                                                                                                                               | REMEDY                                                             | MEASURED                                                                                                                                   |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Asked cold, AI leans to suspicion.</b> The four models that answered every vessel called “stays out” on 58 to 60% of them; 27% stayed out. They did 11.8 to 17.8 points worse than assuming every vessel comes in. | <b>Give the AI the record’s scored base rate,</b> in one sentence. | <b>+5.6 to +14.4 points</b> for every model told it, each range above zero. Grok 4.6: −11.8 to +2.6, level with the bar and with the rule. |
| <b>Alone, the panel loses to the watch officer’s rule.</b> Its own alerts, added to the rule’s, did not help.                                                                                                         | <b>Rule first, AI second:</b> the panel vetoes the rule’s alerts.  | <b>+2.0 points over the rule</b> [+0.5, +3.7]; 22 fewer alerts, 17 of them false.                                                          |
| <b>Claimed confidence is not earned confidence.</b> At 7 or more, four models were right 56 to 68% of the time while claiming 87 to 91%.                                                                              | <b>Score each model’s confidence</b> against what happened.        | Claude Fable 5.1’s confident calls: <b>right 93%</b> , claiming 86%; assuming it comes in, 54%.                                            |
| <b>Four models are not four opinions.</b> Their errors move together, across labs.                                                                                                                                    | <b>Count independent opinions,</b> not models.                     | <b>1.9 independent forecasters</b> from four; all four wrong together 67 times, 12 expected.                                               |
| <b>Not answering is a result.</b> Claude Fable 5.1 declined 16 of its 42 batches, each in under three seconds, with no text.                                                                                          | <b>Score availability</b> beside accuracy.                         | <b>425 of 1,031</b> vessels unanswered by that model: 400 declined, 25 timed out.                                                          |

## Recommended next steps

- 1. Put the measured practices on any AI-assisted watch:** scored history first, the rule first, every model’s confidence and refusals scored (Section 8).
- 2. Run the registered follow-up, Maritime No. 1b:** each model given the scored record of earlier stops, tested head to head against the record alone (Section 9).
- 3. Name the next watch question.** The method runs on any question whose answer the record settles later; readers are invited to propose one.
- 4. Check the record.** Every case, prompt, reply and score is published with the code that re-derives it (Section 11). A walkthrough is available on request.

1 • THE TEST

# A real watch question, with the answer already in the record

## The question

*Will this vessel be inside the 12-nautical-mile line (U.S. territorial waters) within six hours after this stop? YES or NO, with a confidence from 1 to 10.*

## The stops

Every night stop in ten years of NOAA’s public AIS archive (MarineCadastre) whose outcome the record shows: 1,031 cargo ships, tankers and service craft that slowed below 3 knots for 5 to 30 minutes at night, 12 to 24 nautical miles off Texas and western Louisiana (419) and California (612), 2015–2024. The answer is each vessel’s own later track: 751 came inside within six hours (72.8%), 280 did not. The stops are 778 hulls; 120 stopped more than once.

## What the models saw

The vessel as a watch officer would see it when it stopped, one line each: coast, type and length, position and distances to land; month, local time and the sun’s elevation; how it stopped; who was within 1 km; the charted route, anchorage and platforms near it; where it came from in the 72 hours before; any tropical storm nearby. **Left out:** names, MMSI, IMO, call signs, dates and years, storm names, and anything the vessel did after the stop.

## Two arms

**Not told:** the October pilot’s question and condition.

**Told:** the same words and one more sentence: *“in this ten-year record, about 73 of every 100 vessels that made a stop like these were inside the 12-nautical-mile line within six hours after it.”*

## The bars

**Assume it comes in** is right 72.8% of the time. Every result is lift over it, in points, with a 95% range from a bootstrap over whole hulls. **The watch officer’s rule of thumb** is a harder bar: “stays out” only for a vessel that left a U.S. port or anchorage and crossed the 12-mile line less than six hours before it stopped. It is right 76.8% of the time.

## Sealed before scoring

The question, the cases, the answer key’s hash, the bars and the scoring rules were recorded at 21:08 Eastern on 28 October 2026, before any model was asked. Every prompt and reply went into a hash-chained journal as the run happened: 390 calls, 22:34 to 03:49 Eastern. Five dated amendments came before any answer was scored and two notes after; none changed a case, a prompt, the answer key or a primary measure.

## The panel

| MODEL            | LAB       | AS CALLED        | ROLE       | ANSWERED, NOT TOLD | ANSWERED, TOLD |
|------------------|-----------|------------------|------------|--------------------|----------------|
| Claude Fable 5.1 | Anthropic | claude-fable-5-1 | U.S. panel | 606                | not asked      |
| Claude Sonnet 5  | Anthropic | claude-sonnet-5  | U.S. panel | 1,031              | 1,031          |
| Grok 4.6         | xAI       | grok-4.6         | U.S. panel | 1,031              | 1,031          |
| Grok 4.3         | xAI       | grok-4.3         | U.S. panel | 1,031              | 1,031          |
| Qwen 3.7 Max     | Alibaba   | qwen3.7-max      | comparison | 1,031              | 1,031          |
| MiniMax M3       | MiniMax   | MiniMax-M3       | comparison | 0                  | not asked      |

Each model was called through Frontier Colosseum’s own caller with the board’s fixed settings, 25 vessels to a call, the same batches for every model. MiniMax M3’s replies exceeded the board’s length limit, so none of its answers was scored. Every call is counted in Section 7.

2 • FINDING

# Asked cold, the models leaned to suspicion

The October pilot suggested that AI models not given the base rate default to suspicion. Measured on all 1,031 stops, they did: every model treated a stop at sea as a warning sign. A watch floor taking their calls would have chased 1.1 to 1.9 false alarms for every real one; the watch officer’s rule raises 0.7.

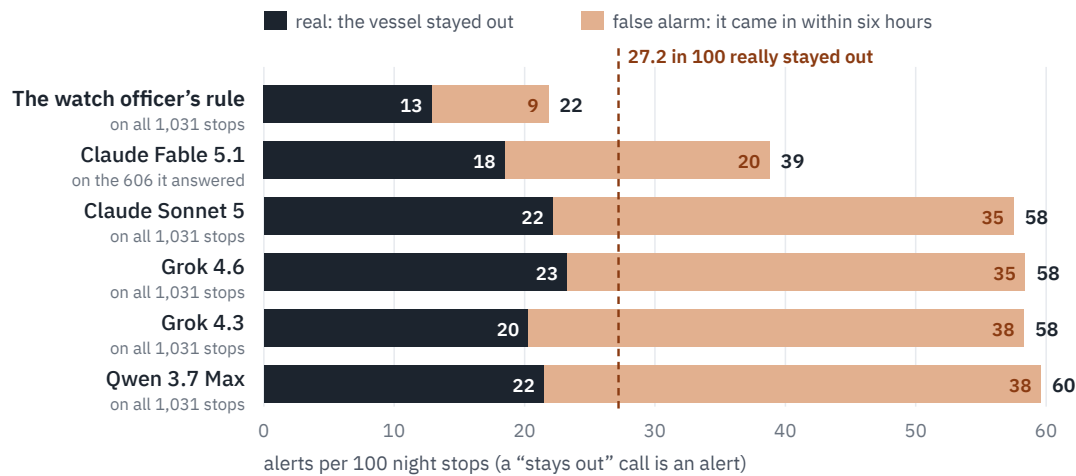


**Figure 1.** Alerts per 100 night stops, not told the base rate. Each “stays out” call is an alert: dark where the vessel stayed out, light where it came inside the line within six hours. Scored answers, not told; the rule on all 1,031 stops.

| MODEL               | VESSELS | LIFT OVER ASSUMING IT COMES IN | OVER THE RULE        | SAID “STAYS OUT” | OF THOSE, STAYED OUT | BRIER SKILL |
|---------------------|---------|--------------------------------|----------------------|------------------|----------------------|-------------|
| Claude Fable 5.1    | 606     | −1.8 [−7.0, +3.4]              | −5.0 [−10.1, +0.2]   | 39%              | 48%                  | +0.026      |
| Claude Sonnet 5     | 1,031   | −13.1 [−18.5, −7.6]            | −17.1 [−21.1, −13.1] | 58%              | 39%                  | −0.433      |
| Grok 4.6            | 1,031   | −11.8 [−17.6, −6.2]            | −15.8 [−20.2, −11.5] | 58%              | 40%                  | −0.431      |
| Grok 4.3            | 1,031   | −17.8 [−22.8, −12.5]           | −21.7 [−25.7, −17.8] | 58%              | 35%                  | −0.690      |
| Qwen 3.7 Max        | 1,031   | −16.5 [−21.8, −11.0]           | −20.5 [−24.4, −16.9] | 60%              | 36%                  | −0.816      |
| COMPARISON          |         |                                |                      |                  |                      |             |
| U.S. panel majority | 508     | −5.9 [−13.1, +1.2]             |                      |                  |                      |             |
| 98 ties left out    |         |                                |                      |                  |                      |             |

The four models that answered every vessel did 11.8 to 17.8 points worse than assuming every vessel comes in, every range below zero. Their “stays out” calls were right 35 to 40% of the time, against 59% for the rule’s own “stays out” group, and their stated confidence made their forecasts worse than the base rate alone (negative Brier skill).

Claude Fable 5.1 was the exception. On the 606 vessels it answered it was level with the bar (−1.8 [−7.0, +3.4]), said “stays out” least often (39%), and was the only model with positive Brier skill when not told (+0.026).

3 • REMEDY ONE

# One scored number moved every model

The told arm added one sentence of scored history to the same prompt. Every model that answered both arms improved, by +5.6 to +14.4 points, and every improvement’s range sits above zero.

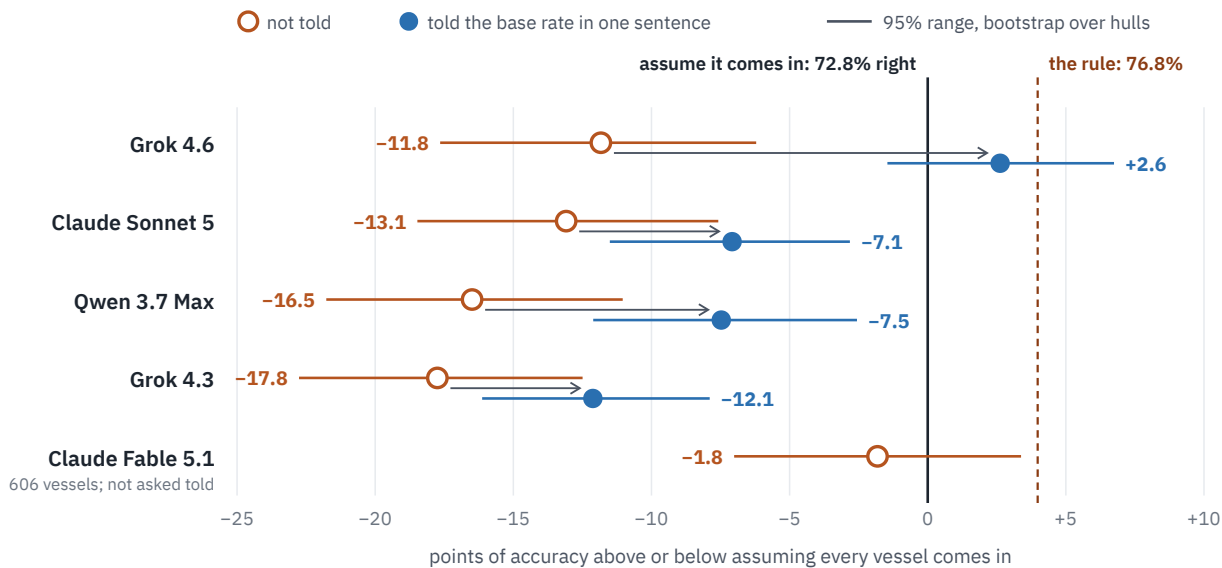


**Figure 2.** Each model’s lift over assuming every vessel comes in, not told and told the base rate, with 95% ranges from the bootstrap over hulls. Dashed: the watch officer’s rule of thumb. In the table, each change is computed before rounding. Scored answers, both arms.

| MODEL           | NOT TOLD | TOLD  | CHANGE, SAME VESSELS | SAID “STAYS OUT” | ANSWERS CHANGED |
|-----------------|----------|-------|----------------------|------------------|-----------------|
| Grok 4.6        | -11.8    | +2.6  | +14.4 [+11.2, +17.7] | 58% to 32%       | 285             |
| Qwen 3.7 Max    | -16.5    | -7.5  | +9.0 [+6.3, +11.8]   | 60% to 47%       | 225             |
| Claude Sonnet 5 | -13.1    | -7.1  | +6.0 [+3.0, +9.1]    | 58% to 45%       | 228             |
| Grok 4.3        | -17.8    | -12.1 | +5.6 [+1.8, +9.4]    | 58% to 39%       | 416             |

**Grok 4.6** moved from -11.8 to +2.6 [-1.5, +6.7]: level with the bar, and level with the watch officer’s rule (-1.4 [-4.2, +1.5] against it). Its Brier skill turned positive (+0.064), and its confident calls beat the bar: at 7 or more it was right 83% of the time where assuming the vessel comes in was right 71%; not told, 68% against 69%.

What changed was suspicion. Every model’s share of “stays out” calls fell toward the 27.2% that really stayed out.

### What this means

Much of the models’ suspicion came from what they were not told. One scored sentence of what really happened improved every model and brought one level with the watch officer’s own rule.

4 • REMEDY TWO

# Rule first, AI second

Asked to replace the watch officer’s rule, the panel lost to it. Placed behind the rule as a second filter, it improved it, by a margin whose range clears zero.

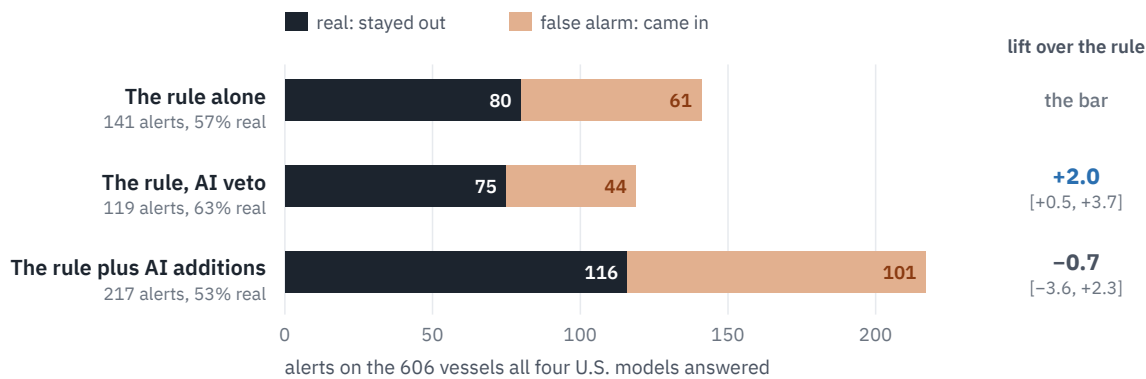


**Figure 3.** Alerts from the rule alone, from the rule with the AI panel as a veto, and from the rule plus the panel’s own alerts, on the vessels all four U.S. models answered. Secondary readouts, registered in Amendment 4 before any answer was scored.

|                            | ALERTS | REAL | FALSE | LIFT OVER ASSUMING IT COMES IN | OVER THE RULE     |
|----------------------------|--------|------|-------|--------------------------------|-------------------|
| The rule alone             | 141    | 80   | 61    | +3.1 [−1.0, +7.4]              | the bar           |
| The rule, AI veto          | 119    | 75   | 44    | +5.1 [+1.2, +9.2]              | +2.0 [+0.5, +3.7] |
| The rule plus AI additions | 217    | 116  | 101   | +2.5 [−2.9, +7.8]              | −0.7 [−3.6, +2.3] |

The rule flags a vessel just out of a U.S. port. On the 606 vessels all four U.S. models answered it raised 141 alerts, 80 of them real. Keeping only the alerts that most of the panel also called “stays out” left 119, 75 real: 22 fewer alerts, 17 of them false alarms, at the cost of 5 real ones. The veto beat the rule by +2.0 points [+0.5, +3.7].

Adding the panel’s own alerts (vessels where all four said “stays out”) raised 217 alerts and did not beat the rule (−0.7 [−3.6, +2.3]). Those unanimous calls still carry information: of the 149 vessels, 91 stayed out, where the base rate would expect 40.5 and the rule’s own groups 57.0. Their value is as a check on the rule, not as a second alarm.

The operating rule, measured

Let the rule raise the alert, and keep only the alerts most of the AI panel also calls “stays out.” Fewer alerts, more of them real, and a gain over the rule that a watch floor can check against its own record.

5 • REMEDY THREE

# Score each model’s confidence against what happened

Every answer came with a confidence from 1 to 10. On the record, a confident call from one model and a confident call from another meant very different things.

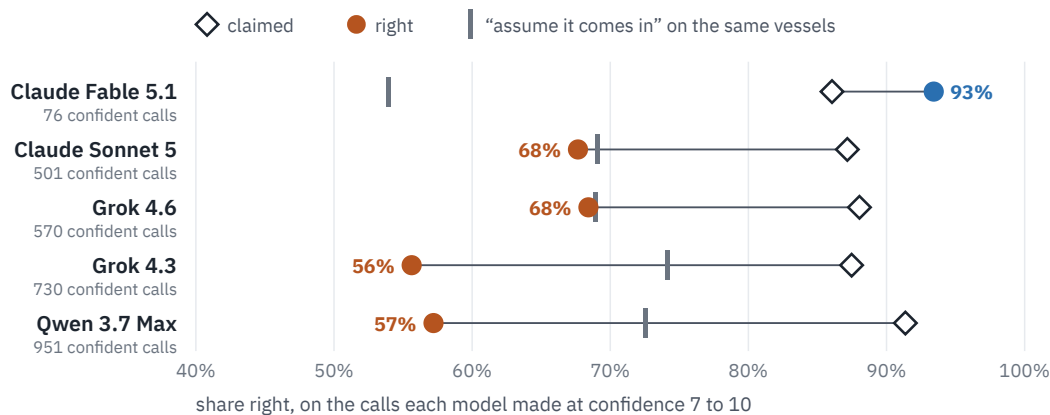


**Figure 4.** At a stated confidence of 7 or more: what each model claimed (on the Colosseum’s map, 7 = 85% sure), how often it was right, and how often assuming the vessel comes in was right on the same vessels. Scored answers, not told.

Claude Fable 5.1’s confident calls were right 93% of the time, against 86% claimed and 54% for assuming the vessel comes in on the same vessels. The other four were right 56 to 68% while claiming 87 to 91%: on the vessels where they were most sure, they did no better than assuming the vessel comes in. Told the base rate, confidence became more honest too: of Grok 4.6’s 181 confident “stays out” calls at 7 or more, 119 stayed out (66%), against 155 of 313 (50%) not told.

## Confident “stays out” calls, 7 or more

|                                         | CALLS | STAYED OUT | SHARE (95% RANGE) |
|-----------------------------------------|-------|------------|-------------------|
| Claude Fable 5.1                        | 34    | 32         | 94% (81–98%)      |
| Claude Sonnet 5                         | 261   | 127        | 49% (43–55%)      |
| Grok 4.6                                | 313   | 155        | 50% (44–55%)      |
| Grok 4.3                                | 429   | 147        | 34% (30–39%)      |
| The rule’s “stays out” list (all stops) | 225   | 133        | 59%               |

### What this means

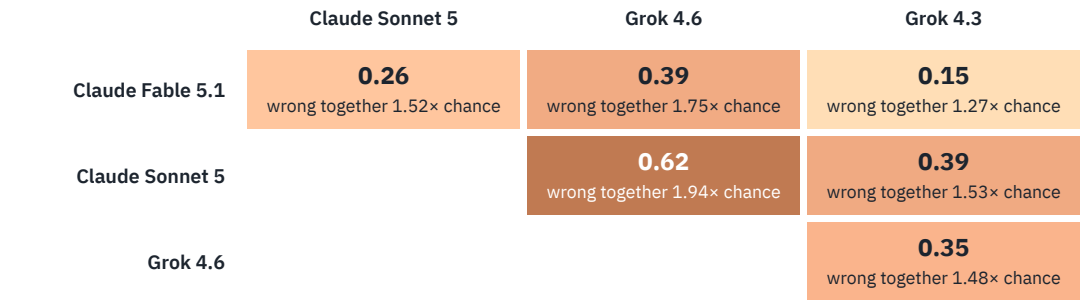
A model’s confidence is a signal like any other: it means what its record says it means, model by model. Route the confident calls of a model whose confidence has been earned, and discount the rest.

6 • FINDING

# Four models are not four opinions

A panel of AI models looks like a second, third and fourth opinion. On this record the four U.S. models were worth about two.

|                                                                                                       |                                                                     |                                                                                                                                   |                                                                       |
|-------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------|
| <div>0.36</div> <div>mean correlation of the four models' errors</div> <div>95% range 0.31–0.41</div> | <div>1.9</div> <div>independent forecasters, from four models</div> | <div>67 vs 12</div> <div>vessels all four got wrong, against what independence predicts</div> <div>of 606 all four answered</div> | <div>Mixed</div> <div>the verdict, by a test set before the run</div> |
|-------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------|



Each cell: the error correlation (phi) of the pair, from 0 (independent) to 1 (the same mistakes), and how often the two were wrong on the same vessel against what two independent models would give.

Figure 5. Error correlation of each pair of the four U.S. models, not told the base rate. Measure registered in Amendment 3, before the run began.

The tightest pair crossed labs: Claude Sonnet 5 and Grok 4.6 (0.62), wrong on the same vessel 1.94 times as often as chance. The two Grok models (0.35) and the two Claude models (0.26) were looser. Claude Fable 5.1 was the most independent of the four (0.15 to 0.39 with the others).

The test was set before the run. On Frontier Colosseum’s prediction-market record (14 models, 41,971 settled calls), the models’ errors correlate at 0.68, so 14 models are worth about 1.4 independent forecasters. That finding would hold on vessels if the correlation reached 0.5 with all four wrong together at least ten times as often as independence predicts, and fail under 0.3. At 0.36 and 5.6 times it is mixed: the models herd less on vessels than on markets, but they herd, and the herding is not a same-company effect.

### What this means for a panel

Count independent opinions, not models. A unanimous panel is not four confirmations, and a model added to a panel adds what its measured independence adds.

7 • FINDING

# Not answering is a result too

A model on a watch floor has to answer the routine traffic. Availability can be measured, and it belongs on the same scorecard as accuracy.

|                                                                                                            |                                                                                           |                                                                |                                                                                               |
|------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------|----------------------------------------------------------------|-----------------------------------------------------------------------------------------------|
| <div>16 of 42</div> <div>batches Claude Fable 5.1 declined, not told</div> <div>400 of 1,031 vessels</div> | <div>0.9 s</div> <div>median time to each refusal</div> <div>no text in any of them</div> | <div>4 of 4</div> <div>declined when asked a second time</div> | <div>0</div> <div>told-arm answers</div> <div>the run closed before it reached that arm</div> |
|------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------|----------------------------------------------------------------|-----------------------------------------------------------------------------------------------|

Claude Fable 5.1 declined 16 of its 42 batches of 25 vessels when not told. Each refusal came back in under three seconds (median 0.9 s) with no text: the API’s refusal stop, not an answer. Four batches were asked a second time, as the rules allow, and it declined all four. After four declines in a row the runner benched it for the rest of that run, and the run was closed as it stood (Amendment 5), so it never reached the told arm and 12 declined batches were not asked a second time. One more batch went unanswered when its call ran past the board’s 480-second limit. Where it answered, its confident calls were right more often than it claimed, the only model’s that were (Section 5).

Every call, counted

| MODEL            | ARM      | BATCHES | ANSWERED | NOT ANSWERED                        | ASKED A SECOND TIME                        | MEDIAN REPLY |
|------------------|----------|---------|----------|-------------------------------------|--------------------------------------------|--------------|
| Claude Fable 5.1 | not told | 42      | 25       | 16 declined, 1 timed out            | 4, all declined                            | 76 s         |
|                  | told     | 0       | 0        | not asked: the run closed first     | –                                          | –            |
| Claude Sonnet 5  | not told | 42      | 42       | –                                   | –                                          | 103 s        |
|                  | told     | 42      | 42       | –                                   | 1 timed out, then answered                 | 76 s         |
| Grok 4.6         | not told | 42      | 42       | –                                   | 1 timed out, then answered                 | 180 s        |
|                  | told     | 42      | 42       | –                                   | 2 timed out, then answered                 | 184 s        |
| Grok 4.3         | not told | 42      | 42       | –                                   | 1 refused at the lab’s door, then answered | 20 s         |
|                  | told     | 42      | 42       | –                                   | –                                          | 16 s         |
| Qwen 3.7 Max     | not told | 42      | 42       | –                                   | –                                          | 47 s         |
|                  | told     | 42      | 42       | –                                   | 1 timed out, then answered                 | 45 s         |
| MiniMax M3       | not told | 2       | 0        | 2 past the length limit             | –                                          | –            |
|                  | told     | 0       | 0        | not asked: sat out in the other arm | –                                          | –            |

A batch is 25 vessels (the last, 6). A call that timed out or broke off is asked once more at the end of its arm, a refusal at the lab’s door is retried, and a reply that answers nothing is asked once more. Median reply is the time to a full answer.

What this means

A model that declines four batches in ten of routine vessel data cannot carry a watch alone, however good its answers are. Score refusals and response times beside accuracy, and keep a second model on the watch.

## 8 · RECOMMENDATIONS

# What a watch floor should do

Five practices, each tied to what this record measured. Each uses what a watch floor already has: its own record, scored against what really happened.

## 1 Give the AI the scored history before it judges.

**MEASURED HERE** One sentence of base rate lifted every model told it by +5.6 to +14.4 points, each range above zero, and pulled its “stays out” calls toward the 27.2% that really stayed out.

**TO START** Attach the base rate for that kind of stop to every AI query. The registered follow-up tests the full record of earlier stops like this one and at this spot (Section 9).

## 2 Keep the rule first. Let the AI clear alerts, not raise them.

**MEASURED HERE** As a veto on the watch officer’s rule, the AI panel beat the rule by +2.0 points [+0.5, +3.7] and cut 17 false alarms on 606 stops. As a source of extra alerts it added nothing (−0.7 [−3.6, +2.3]).

**TO START** Route the rule’s alerts through the AI panel, keep the ones most of it also calls “stays out,” and check the cleared ones against the floor’s own record.

## 3 Trust a model’s confidence only as far as its record.

**MEASURED HERE** At their most confident, four models claimed 87 to 91% and were right 56 to 68%. One was right 93% while claiming 86%.

**TO START** Keep every model’s confident calls against what happened, and act on confident calls only from models whose record has earned it.

## 4 Count independent opinions, not models.

**MEASURED HERE** Four models were worth 1.9 independent forecasters, and all four were wrong together 5.6 times as often as independent models would be.

**TO START** Measure how often each pair of models errs together on the floor’s own record, and read a unanimous panel as one strong opinion, not four.

## 5 Score availability beside accuracy, and keep a backup on watch.

**MEASURED HERE** One model declined 16 of its 42 batches of routine vessel data. The four that answered every batch took a median of 16 to 184 seconds per batch of 25 vessels.

**TO START** Log every refusal, timeout and response time per model, and never let one model carry the watch alone.

### One practice underneath all five

Score every signal the watch relies on (each sensor, each rule of thumb, each AI model, each watchstander’s call) against what really happened, continuously, and let those scores decide how far each is trusted.

## 9 · WHAT FOLLOWS

# Score every signal, and give the AI the scores

The cheapest remedy in this study was the strongest: one settled number lifted every model told it, by +5.6 to +14.4 points.

---

## THE OPERATING IDEA

- 1 Give the system fresh data that has already been graded against outcomes.
  - 2 Let those grades decide how much weight each signal (or each model) carries.
  - 3 Do not let either the operator or the model rewrite the grades.
- 

That is the idea behind scoring every signal against settlement and feeding the scores forward, and it is what Gage Systems builds, in its PPLN protocol and its maritime overlay, Gage Maritime. A watch floor already relies on many signals: AIS, radar, the rules of thumb its officers carry, reports from other vessels, and now AI models. Each can be graded, automatically and continuously, against what really happened, the way this study graded five models against ten years of outcomes. This study measured how large the lift can be from the simplest version of that idea: one pre-computed base-rate sentence.

## The next step is registered: Maritime No. 1b

It tests whether the principle scales when the data is richer than a single number. Each model gets, on each vessel, two lines of scored record built only from stops that came before it: of the earlier stops like this one, how many came inside within six hours; of the earlier stops within 10 km of this spot, how many did. Three things are compared on the same vessels:

1. **the AI alone:** this study's answers, not told;
2. **the record alone:** "comes in" when at least half the earlier stops like this one did;
3. **the AI given the record.**

The thesis holds if the AI given the record beats both, each by a range that clears zero. Anything else is reported as it falls. The design was registered at 00:15 Eastern on 29 October 2026, while this study was running and before any of its answers was scored. Its prompts are built and checked, and it is ready to run.

## The next fields

The same method, one study at a time, each published on Frontier Colosseum with its full record: **the power grid** (grid operators' day-ahead demand forecasts beside the demand that came), then **weather** (the National Weather Service's daily climate reports, city by city).

## 10 · METHOD AND LIMITS

# What the numbers rest on

- **The answer.** Each vessel's first fix inside the 12-mile line, on the stop-at-sea sweeps' own clock. A second clock in the source lists disagrees on 19 vessels (most crossed within about an hour of the six-hour mark). Each model's lifts and the panel majority's were scored again on the 1,012 vessels where the clocks agree; no finding changes.
- **The line.** The 12-mile line is measured from Natural Earth's 1:10m coast, not the charted baseline; a vessel within about a mile of the line may sit on either side of the legal one. A vessel that crossed the line unheard and left again would count as staying out.
- **Hiding the date.** The month, clock time, sun's elevation and position together narrow the day of the year, never the year. No case carries a name, an identifier or a date. One field, other ships' short stops within 2 km, counts the spot's whole decade, before and after the stop: it describes the place, as a watch floor's history of it would. NOAA publishes no files for January–June 2024.
- **Ships repeat.** The 1,031 stops are 778 hulls. Lift ranges resample whole hulls. Rates shown with a Wilson range (confident calls, the panel's unanimous calls) treat each stop as its own and are narrower than a hull-level range would be.
- **Many measures.** With five models, two arms and several measures, some results will clear their range by chance. A single lift near its line is a lead, not a finding. The veto (Section 4) is a secondary readout, one of three rules registered together, and its range clears zero narrowly; the base-rate gains (Section 3) clear zero for every model.
- **Two arms, one sentence.** A model's change between the arms includes the variation of asking it twice; the board sends its own fixed settings, and no lab promises the same reply twice.
- **How the models were called.** Through Frontier Colosseum's own caller: each lab's own endpoint and model name, a 32,768-token limit and a 480-second timeout, with temperature and seed set where the lab accepts them. Results describe the named models on 28–29 October 2026 through those settings, and may differ for other versions or settings.
- **Who is on the panel.** The seats on Frontier Colosseum's roster that could be called on the run date (Amendment 2): the U.S. panel is Anthropic's and xAI's. Google's two seats could not be called, and the roster carried no OpenAI seat. Claude Fable 5.1 declined the smoke test and was asked under the refusal rule; its lifts are on the 606 vessels it answered, against the bar on those same vessels.
- **The panel measures** use vessels answered by four or more U.S. models: 606 not told. Told, only three U.S. models answered, so the panel's told-arm measures are reported over every model instead.

## Sources

U.S. Customs and Border Protection, "CBP, Coast Guard and HSI interdict \$29 million dollar cocaine shipment," news release, 10 February 2015: [cbp.gov/newsroom/local-media-release/cbp-coast-guard-and-hsi-interdict-29-million-dollar-cocaine-shipment](https://cbp.gov/newsroom/local-media-release/cbp-coast-guard-and-hsi-interdict-29-million-dollar-cocaine-shipment). NOAA Office for Coastal Management, [MarineCadastre.gov](https://marinecadastre.gov) AIS data, 2015–2024. Natural Earth, 1:10m land.

## Scoring, as registered

Lift over assuming every vessel comes in, never raw accuracy. Confidence becomes a probability by the Colosseum's declared map (YES  $p = 0.5 + c/20$ , NO  $p = 0.5 - c/20$ ), and Brier skill is against the base rate. Earned confidence at 7 or more is set against what was claimed and against the bar on the same vessels. Failed calls and refusals are counted and reported, never scored.

## 11 · THE RECORD

# Everything needed to check every number

The cases, the words of both arms, the runner, the bridge to the board, the scorer, the journal with every prompt and reply, the scored files and the answer key are published on [frontiercolosseum.com](http://frontiercolosseum.com), free to reuse under CC BY 4.0 with attribution to the Gage Institute.

| FILE                                            | SHA-256                                                          |
|-------------------------------------------------|------------------------------------------------------------------|
| Journal head (the run's seal)                   | 8335d83abf1b734edc2a327654c8473e19d26b043d0819c87ad1a8bfd2c98bc8 |
| sea_cases.jsonl (the cases the models saw)      | cd96db0bdf588c5218bce10ac9c9083ccad154668e1ee7675965e66f1f6ef3a  |
| sea_answer_key.csv (the answers, as registered) | a86eda57a3d6cd4ba335b081e8ca10bb3b28e3499fd8288e8b27b8fccdbd5fce |
| sea_protocol.py (the words of both arms)        | a68cafc223b2ed0cc9c34bc53adc0a73f5699ec7ef1b62270b6f88926e7d3c20 |
| study.py                                        | 5d42f7329e0726527a6ba8af3a847676f4936f662bee07c48c8c5eb4c133490b |
| answers.jsonl (every answer, as scored)         | 8b41c20192dd7373d9dec7e71dc1d4124bd704e7fa1cb925f6bffb43c2215d63 |
| scores.json                                     | 65bbf49970aa097b02989ff1f70d4d584214e60854a2d79654685e094f3e59bb |
| herding.json (Amendment 3)                      | e7a9fc3cdf624cbc4af74ba47ea32ef508d3de2051f9edbdb2020389480c21a1 |
| readouts.json (Amendment 4)                     | 04883f58fd076230bbe229e3ec75882835ed410197cb3f9802b7938f83d32edf |

The journal holds 404 lines, each chained to the one before by its hash: the smoke test's opening and its 11 calls, the run's opening, its 390 calls, and the close. The pre-registration, its five dated amendments and two later notes are published beside the results, with the hashes of every file they changed.

## To re-derive every number

`python3 run_panel.py --score` scores the journal against the published answer key, and `python3 herding.py` and `python3 readouts.py` compute the two registered additions: all four scored files come out byte for byte. With the named files, `python3 verify.py` rebuilds the cases from the source lists and checks them. Standard-library Python only: no keys, no network, the same numbers on any machine.

## Named vessel data

The published study materials deliberately omit vessel names, MMSI and IMO numbers, and the exact dates and times of the stops. This keeps the public record free of any implication about individual vessels. For researchers or organizations that need the named case list to check the cases and answers against the public AIS record, the named files are available only under a short data-use agreement. Their SHA-256 hashes match the ones registered before any model was asked, so anyone who signs can confirm they received the originals.

**To request them**, email [gagesystems2026@outlook.com](mailto:gagesystems2026@outlook.com) with your name and affiliation; a short statement of purpose (for example, "independent verification of the ARCS Maritime No. 1 results"); and confirmation that you will not redistribute the named list or use it for any commercial, enforcement or public reporting purpose. You will receive a short data-use agreement; once it is signed and returned, the named case file will be provided.

ABOUT

# About the Gage Institute and Gage Systems

The Gage Institute (ARCS: Auditable and Responsible AI in Critical Systems) studies how AI performs when its calls matter. It asks the question before the answer exists, seals what the AI said, and grades it against what really happened.

*“If AI has the signals, it will be right more often. If it isn’t given the context, how can it answer? So we score every signal, the AI’s and the human’s too, against what really happened, and give the AI those scores.”*

SHANE GAGE, FOUNDER

## What Gage Systems brings

### The record

Ten years of the public AIS record, swept stop by stop and graded against what each vessel did next. Earlier work on the same record: AIS silence and identity anomalies off Texas and off South Carolina, 2015–2024.

### The protocol

PPLN: every signal and every model scored automatically against settlement. Trust is earned, never declared, and the grades are computed in code that neither an operator nor a model can adjust.

## Independence

No AI lab pays for a placement or a result, and no lab funded, reviewed or endorses this work. The Gage Institute is a program of Gage Systems LLC.

## Contact

|              |                                                                  |
|--------------|------------------------------------------------------------------|
| Name         | Shane Gage, Founder                                              |
| Organization | The Gage Institute · Gage Systems LLC · Beaufort, South Carolina |
| Email        | gagesystems2026@outlook.com                                      |
| Web          | frontiercolosseum.com                                            |
| UEI          | VNENBH8UBC33                                                     |

### The harness

Frontier Colosseum: since August 2026 frontier AI models have answered sealed questions three times a day, every answer hash-chained before its outcome was known, with 41,971 settled calls graded by 27 October 2026. Gage Systems donated it to the Institute on 28 October 2026.

### The method, in the open

Every study is pre-registered, sealed and published with its code and its full record, so anyone can check every number.

**Disclaimer.** This brief reports how AI models answered questions about public AIS records and how those answers compared with what the record shows. It names no vessel and alleges nothing about any vessel, company or person: geometry, never intent. Results describe the named model versions on the dates asked, called with the stated settings.