FORM S1-00 | SHEET Cover sheet | VERDICTS SO FAR Can 0 Cannot 0

System 1 Models in Safety Research

What Jev and Laya Can and Cannot Do

TONIGHT'S QUESTION

Can a System 1 model carry a safety decision?

Can ☐☐☐☐☐☐☐☐☐☐☐ ?

Cannot ☐☐☐☐☐☐☐☐☐☐☐ ?

VERDICT

To be completed by 7:00 PM.

PRESENTER	UNIT	DATE	HOSTED BY
Amir Rafe, Ph.D.	AIT Lab	2026-10-14	CE 7393

Same report. Two kinds of answer.

CHAT AI · WRITES

The report describes a worker on a level 3 scaffold who slipped near a section where the east guardrail was missing. The worker's harness appears to have arrested the fall, and no injury is mentioned, although the missing guardrail suggests a fall hazard that the site team should review before work continues on that level...

55 words — you still have to read it and judge it

DECISION MODEL · ANSWERS

Fall hazard? 0.99

Anyone injured? 0.04

0.2 s — numbers your code can act on

Right: real Jev 1.13 answers on the synthetic scaffold report from Lab 1, recorded 10 Oct 2026. Left: a typical chat answer, written for this slide.

ATTACHMENT · MOTION

Incident report

SYNTHETIC REPORT

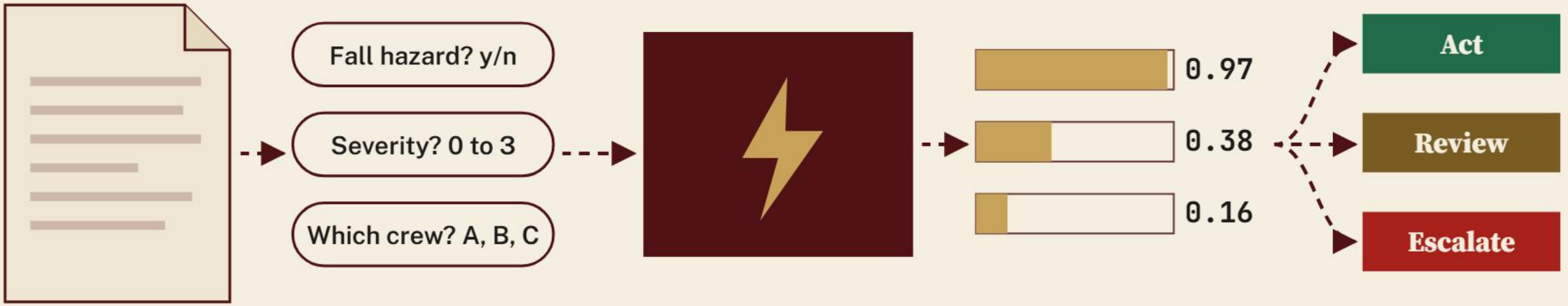
Worker on level 3 scaffold. East guardrail

WATCH FOR

Both bars fill in one pass, while the generative model is still writing.

NOTE

Synthetic report and illustrative values.



State
a report, a log, JSON

Typed questions
options you declare

Decision model
one pass, no text

Probabilities
one per option

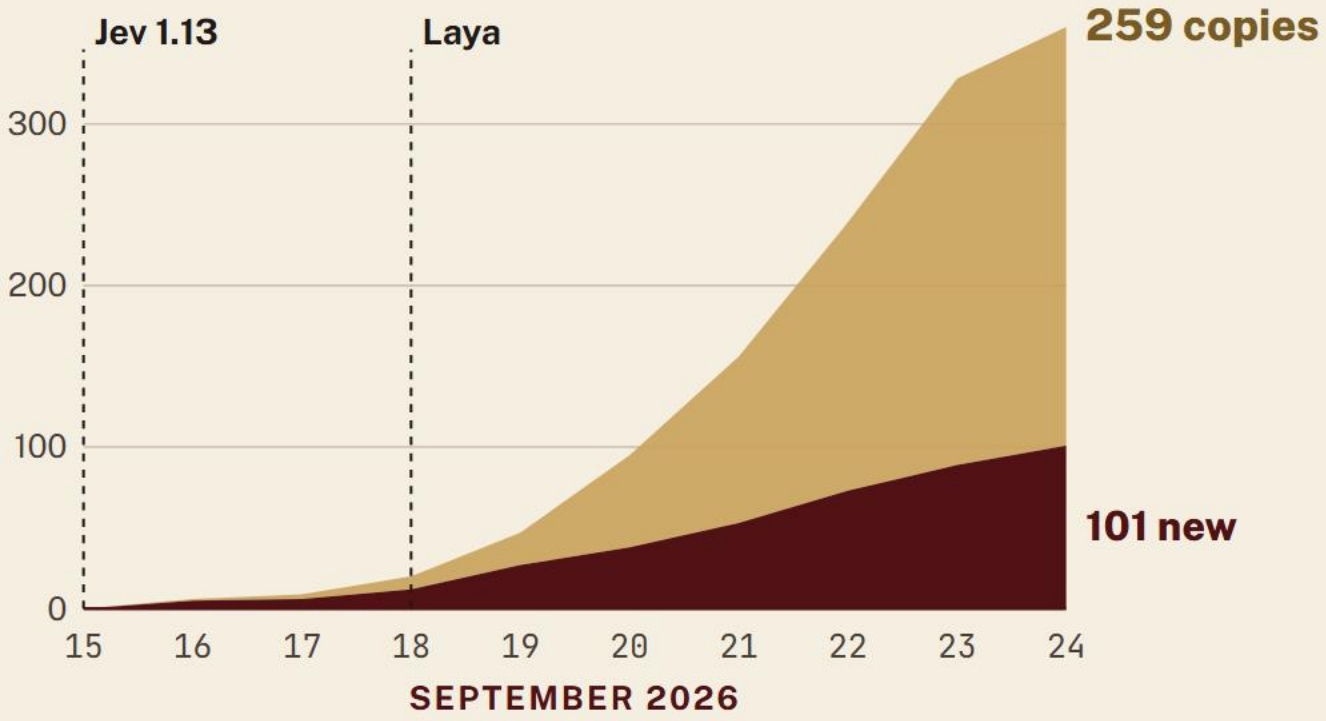
Your rule
a threshold you set

The model scores. Your rule decides.



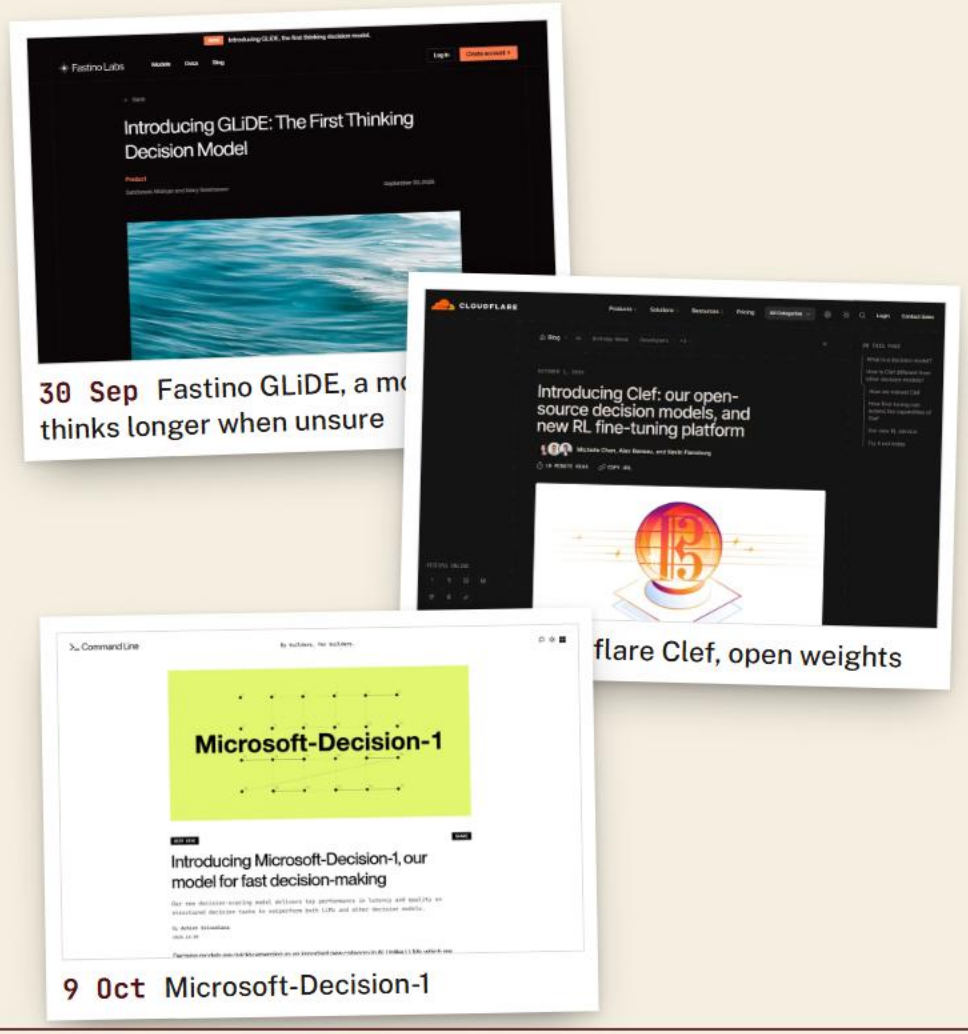
Five lineages traced in our survey of System One models (in preparation). Nothing here is new on its own; Jev put them behind one simple contract.

360 open models in nine days

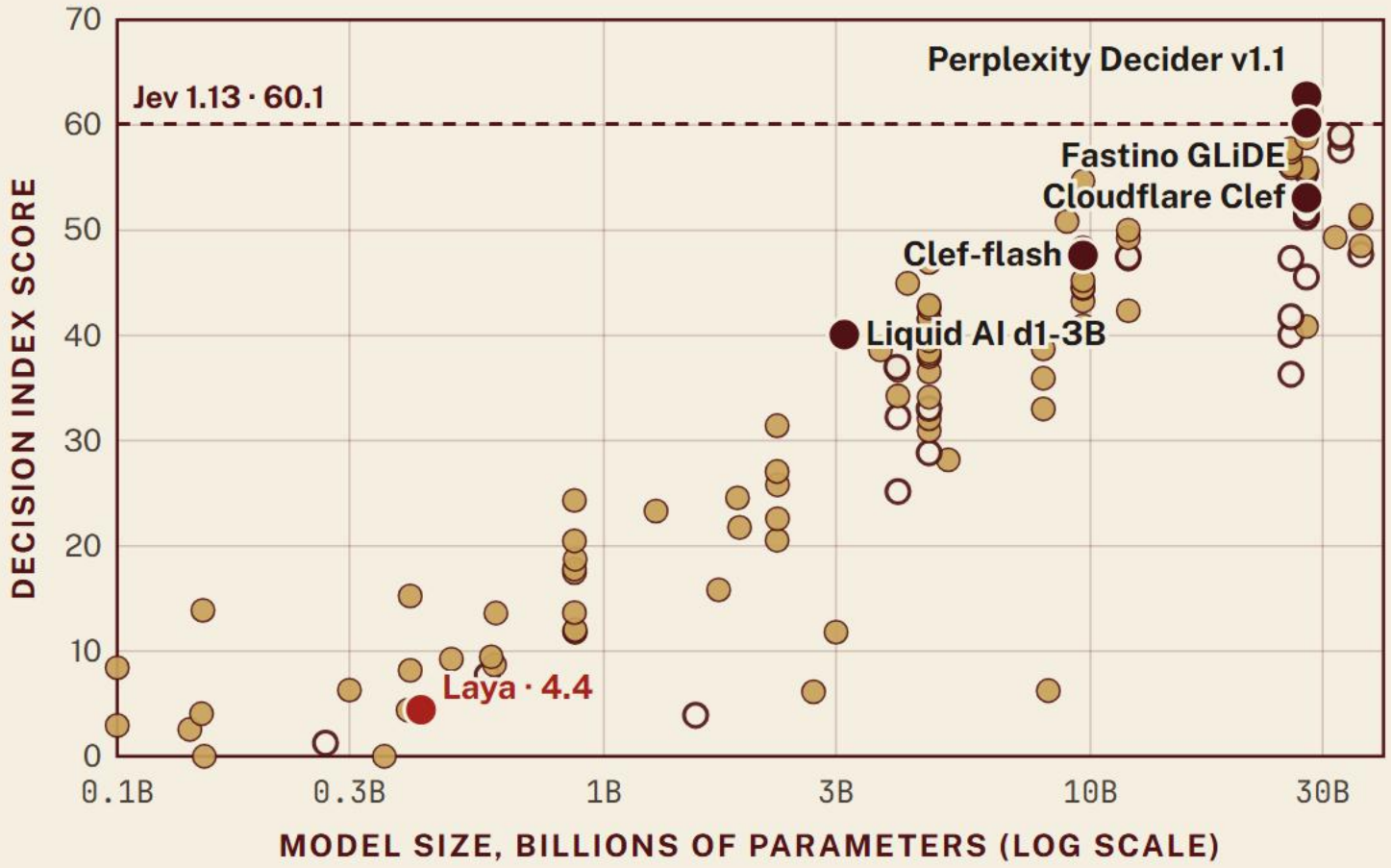


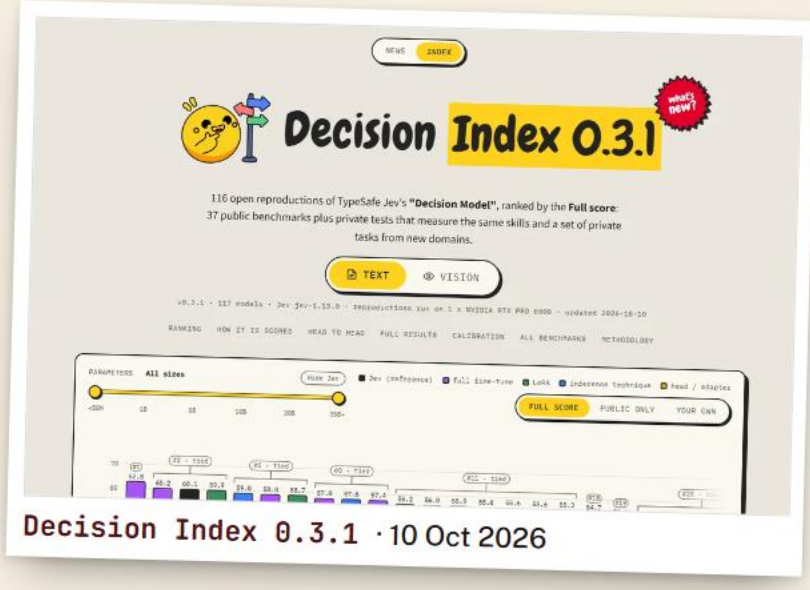
Open System One checkpoints on Hugging Face by release date, from the census in our survey (in preparation). Originals are new models; derivatives are copies, conversions and fine-tunes.

THEN THE BIG COMPANIES ARRIVED



Primer · 117 models on one leaderboard

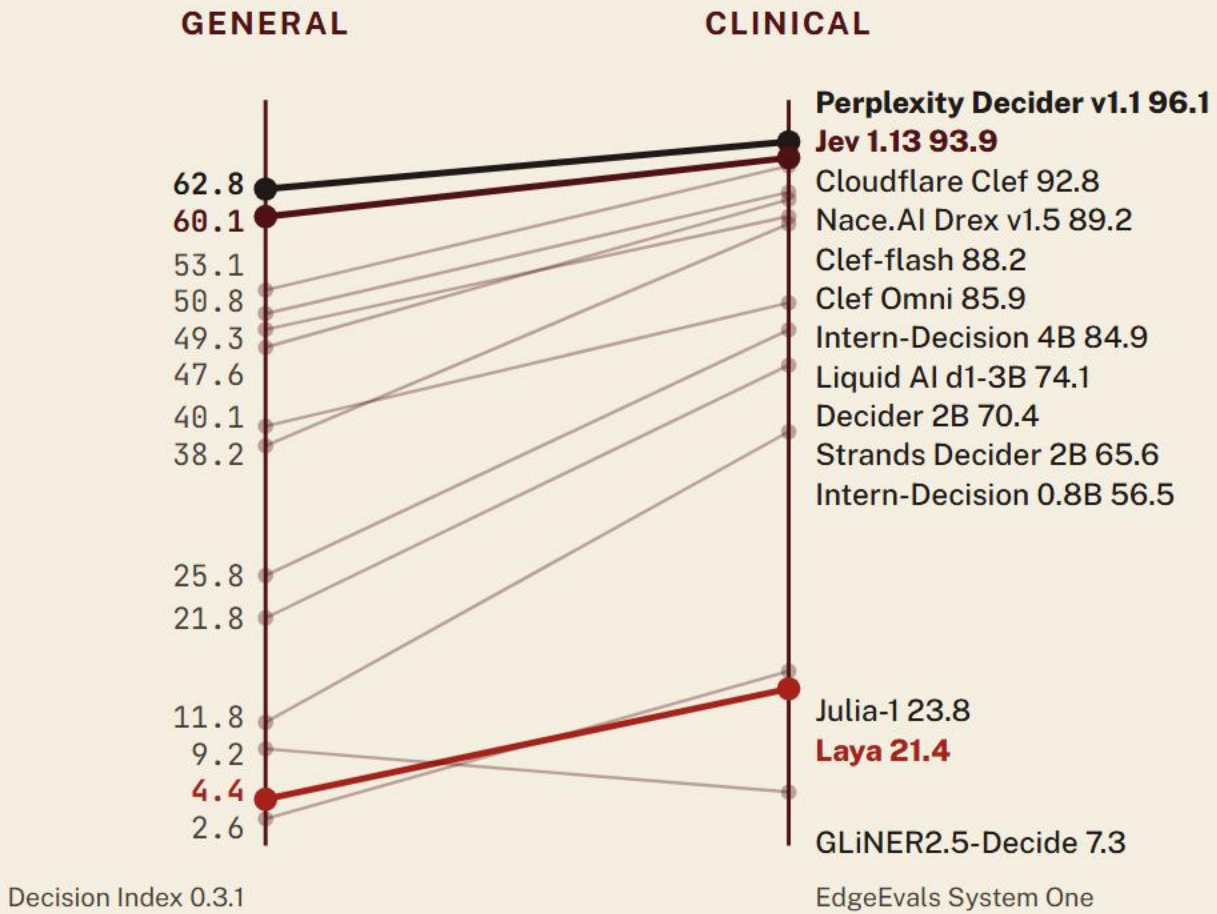


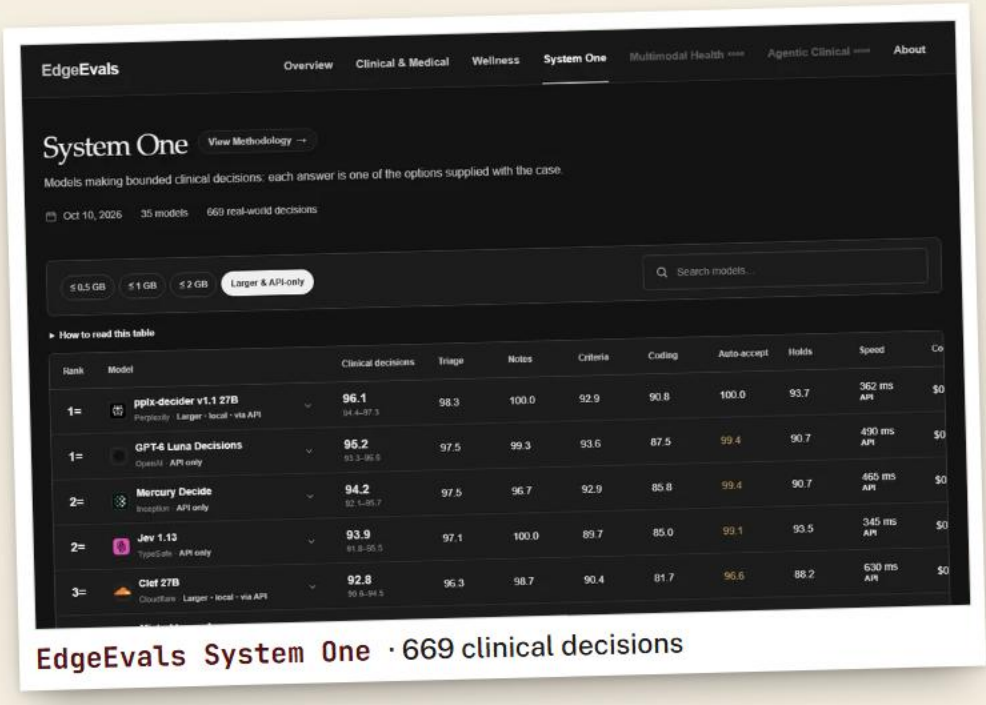


Decision Index 0.3.1 · 10 Oct 2026

- open weights
- weights not public
- Jev, size not disclosed

Big open models now reach Jev. Small ones, like Laya, are far behind.





EdgeEvals System One · 669 clinical decisions

Both boards put the same models near the top. Neither tests safety decisions.

So tonight we test them on safety data ourselves.



Route

send a defect note to the right crew



Guard

stop an AI agent's unsafe action



Verify

check another model's answer



Triage

rank bridge alarms by urgency



Code

turn crash reports into variables

TONIGHT'S QUESTION

Can a System 1 model carry a safety decision?

Seven short labs. You predict first, every time.

A workshop, not a lecture

Seven short labs. In each one you commit to a guess before you see any model output.

EVERY LAB FOLLOWS THIS PROCEDURE

- ☑ **Predict**
  You answer first, in Teams chat or a Teams poll.
- ☑ **Run**
  We run real code on screen and read the raw response.
- ☑ **Reveal**
  The finding is stamped CAN or CANNOT and added to the tally.

ANSWER IN	VOTES	CODE ALONG
Teams chat	Teams polls	Colab, optional

NARRATIVE SYNTHETIC

Unit 1 northbound in light rain, left lane. Driver states she looked down at her phone for directions, looked up, and traffic had stopped. Unit 1 struck the rear of Unit 2. No injuries reported.

Was the driver distracted?
Did the weather contribute?

ANSWER IN TEAMS CHAT

distracted 90, weather 20

For each question, how likely the answer is yes, in percent.

TIME

0:20

You read a report and returned two probabilities.

That is exactly the job of the decision models you just met. A typed answer and how sure, nothing else.

Recorded run · jev-1.13.0 · 10 Oct 2026

MODEL'S ANSWER

Was the driver distracted?

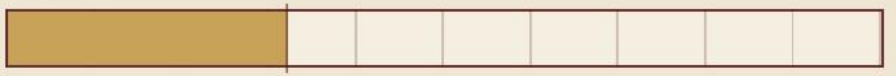
P(yes)



0.97

Did the weather contribute?

P(yes)



0.32

Compare with the chat. Where did the class disagree most?

Lab 1

Anatomy of one call

One construction report, three typed questions, two models. Same request, very different answers.

PROCEDURE

- [x] **Predict**
  Will an open model on a regular computer agree with the hosted one?
- [x] **Run**
  The same three questions go to Laya and to Jev.
- [x] **Reveal**
  We compare the answers with what the report says.

TIME	COLAB	EVIDENCE
8 min	Cells 1 to 3	arXiv 2610.00346

ATTACHMENT A · LAYA 0.4.2

```python
from laya import Router

router = Router()   # downloads the open checkpoint once

state = ("Worker on level 3 scaffold. East guardrail missing. "
         "Worker slipped, harness held, no injury.")
questions = {
  "fall_hazard":  {"type": "noul",
     "instructions": "Does the report describe a fall hazard?"},
  "harness_held": {"type": "noul",
     "instructions": "Did fall protection stop the fall?"},
  "injury":       {"type": "noul",
     "instructions": "Was anyone injured?"},
}
result = router.predict(state, questions)
```

● Recorded run · 0.17 s

RESPONSE · LAYA DEFAULT CHECKPOINT

fall_hazard
P(yes)
0.22

harness_held
P(yes)
0.61

injury
P(yes)
0.55

Read the report again. Which of these three is wrong?

STATE SYNTHETIC

Worker on level 3 scaffold. East guardrail missing. Worker slipped, harness held, no injury.

TYPED QUESTIONS

y/n Does the report describe a fall hazard?

y/n Did fall protection stop the fall?

y/n Was anyone injured?

y/n Question from the chat

RUN BOTH | LAYA | JEV | ● Runner live on this computer

LAYA 0.4.2 · OPEN WEIGHTS · THIS COMPUTER

	P(yes)
Does the report describe a fall hazard?	0.22
Did fall protection stop the fall?	0.61
Was anyone injured?	0.55

live · laya · english · 1483 ms

With plain yes and no descriptions added to each question, Laya moves to 0.37, 0.77 and 0.65. Wording changes the answer.

JEV 1.13 · TYPESAFE API

	P(yes)
Does the report describe a fall hazard?	0.98
Did fall protection stop the fall?	0.89
Was anyone injured?	0.04

live · jev-1.13.0 · 145 ms

ACCURACY ON TYPED DECISIONS AND INTENTS (ARXIV 2610.00346)

MODEL	TYPED DECISIONS	CLINC-150
DeBERTa-v3-large, fine-tuned on labels	0.793	0.953
Gemma-4-31B, generative, no labels	0.741	0.942
Jev 1.13, no labels	0.732	0.913

SWAP ONLY THE NAMES "YES" AND "NO", SAME DESCRIPTIONS

MODEL	FLIPS PER 100 · FEWER IS BETTER
Kev-9B	6.5
Jev 1.13	50.5
Laya English	91.3

Where labels exist, a fine-tuned classifier still wins. Without labels, a System 1 model is a strong, fast start.

Our benchmark, arXiv 2610.00346. Eight decision checkpoints in six families plus four open generative models and four classifier baselines, eight datasets.

These models can read a report and answer your yes or no questions in under a second.

REVIEWED BY



These models cannot be trusted until you test them on your own reports.

REVIEWED BY



Lab 2

Design the schema

You write the questions. We add gates, run them on one report, and read what comes back.

PROCEDURE

- [x] **Predict**
  Post one typed question in Teams chat.
- [x] **Run**
  Your questions go into the schema and run live.
- [x] **Reveal**
  We read the probabilities and stamp the verdict.

TIME	COLAB	EVIDENCE
10 min	Cell 4	arXiv 2609.24052

NARRATIVE SYNTHETIC

Unit 2 was stopped at the light on FM 1626. Driver of Unit 1 stated he felt dizzy and blacked out before the crash. Witness stated Unit 1 did not brake. Driver of Unit 1 transported to hospital by EMS.

TYPE · NOUL

Yes or no

Did the driver have a medical episode?

TYPE · CHOICE

Pick one option

What kind of episode? Seizure, blackout, diabetic, cardiac, other.

TYPE · SCORE

An ordinal level

How detailed is the narrative, from 0 to 3?

YOUR TURN · TEAMS CHAT

Post one yes or no question you would ask of this report.

For example, `Was the driver of Unit 1 taken to the hospital?`

1:00

ATTACHMENT · MOTION



WATCH FOR

A screen question opens or shuts a gate. Follow-up questions run only behind an open gate.

WHY GATE

Fewer questions per report, and no detail asked of reports where it cannot apply.

ATTACHMENT B · CRASH_FACTORS_V1.JSON

```
"medical_episode": {
  "type": "noul",
  "instructions": "The narrative states that a driver had a
    medical episode, such as fainting or a seizure."},
"medical_type": {
  "type": "choice",
  "instructions": "What kind of medical episode does the
    narrative describe?",
  "criteria": {
    "seizure": "...", "loss_of_consciousness": "...",
    "diabetic": "...", "cardiac": "...", "other": "..."}},
"gates": {
  "medical_type": "medical_episode",
  "drug_class":   "drug_involved", ...}
```

THE FULL SCHEMA

Questions	27
Yes or no (presence)	16
Choice (categorical)	8
Score (ordinal)	3
Behind a gate	7

A gate opens when the screen question returns a probability above 0.5.

Instruction text shortened for the slide. Variables include medical episode, drugs, alcohol, hydroplaning, vehicle defect, fatigue, aggression, phone use, wrong way, animal, pregnancy.

STATE SYNTHETIC

Unit 2 was stopped at the light on FM 1626. Driver of Unit 1 stated he felt dizzy and blacked out before the crash. Witness stated Unit 1 did not brake. Driver of Unit 1 transported to

TYPED QUESTIONS

y/n	The narrative states that a driver had a medical episode, such a
pick	What kind of medical episode does the narrative describe?
y/n	The narrative cites a statement from a witness.
y/n	Question from the chat
y/n	Question from the chat

RUN BOTH | LAYA | JEV | ● Runner live on this computer

LAYA 0.4.2 · OPEN WEIGHTS · THIS COMPUTER

The narrative states that a driver had a medi... P(yes)	0.81
What kind of medical episode does the narra... top option · loss of consciousness	0.92
The narrative cites a statement from a witne... P(yes)	0.69

live · laya · english · 1363 ms

JEV 1.13 · TYPESAFE API

The narrative states that a driver had a medi... P(yes)	0.96
What kind of medical episode does the narra... top option · loss of consciousness	1.00
The narrative cites a statement from a witne... P(yes)	0.97

live · jev-1.13.0 · 136 ms

TEXAS CRASH NARRATIVES, 2017 TO 2025 (ARXIV 2609.24052)

Narratives in scope	5,018,080
Screened with 10 questions	499,500
Coded with all 27 questions	195,857
Blinded human judgments	2,416

AGREEMENT WITH HUMAN LABELS (F1)

Jev 1.13	0.908
Frontier generative model A	0.967
Frontier generative model B	0.885

INJURY AND FATAL CRASHES PER YEAR ATTRIBUTED TO NINE FACTORS



Phone use rises by 182 percent.
Four of nine factors change rank.

Code police crash narratives at state scale with a gated schema.

EVIDENCE

arXiv 2609.24052. Jev 1.13 on 195,857 Texas narratives. F1 of 0.908 against 2,416 blinded human judgments.

REVIEWED BY



Lab 3

The trust meter

A probability is only useful if 0.9 means right nine times in ten. We check that, and fix it when it fails.

PROCEDURE

- [x] **Predict**
  Teams poll on what 0.90 really means.
- [x] **Run**
  Reliability curves from three of our studies.
- [x] **Reveal**
  How many reports a person still has to read.

TIME	COLAB	EVIDENCE
10 min	Cell 5	arXiv 2609.24052 arXiv 2610.00346 Pedestrian study

The model says 0.90 on a hundred reports.

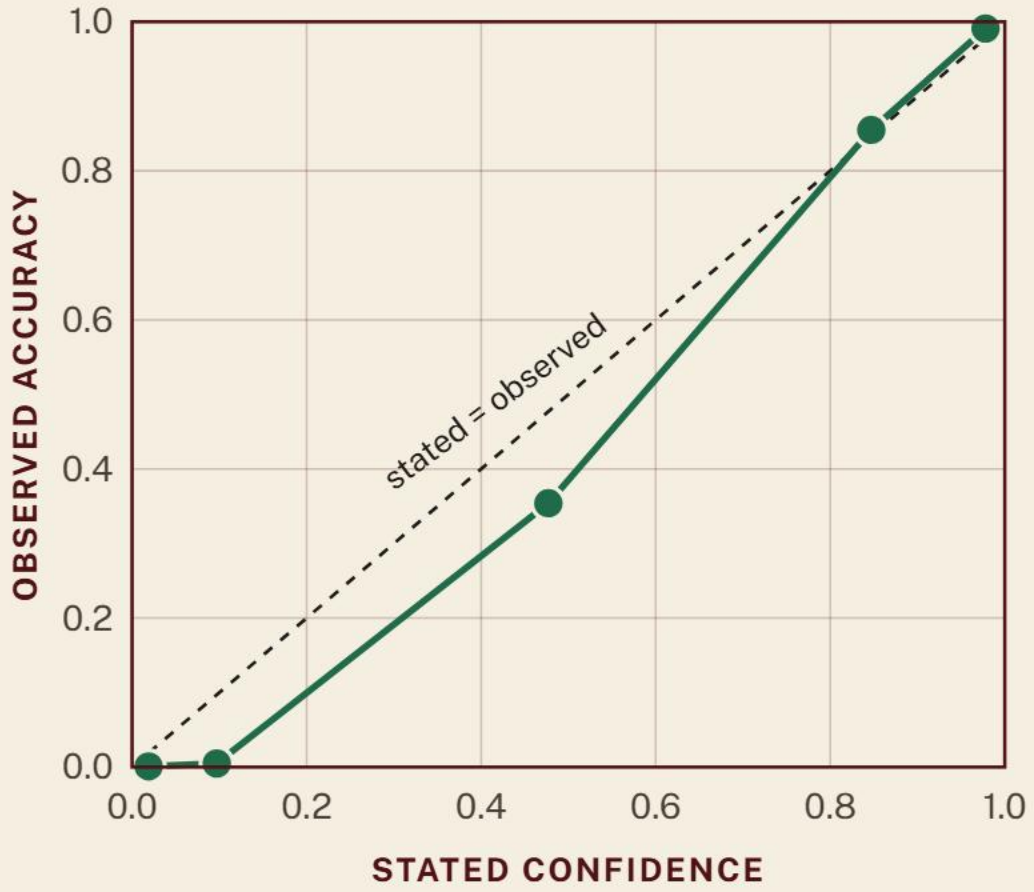
How many does it get right?

TEAMS POLL 2 · SINGLE CHOICE

- ☐ About 90
- ☐ About 75
- ☐ About 50
- ☐ Cannot know until we measure

Launch poll 2 from the Polls tab now.

STATED CONFIDENCE AGAINST OBSERVED ACCURACY



CRASH NARRATIVES · JEV	PEDESTRIAN · JEV
TYPED DECISIONS · LAYA	EMOTION · JEV AS SHIPPED
EMOTION · JEV RESCALED	

ARXIV 2609.24052 · JEV 1.13 · CRASH NARRATIVES

Points sit on the line. The tails can be trusted.

2,416 blinded human judgments in five bins. When it says 0.98 it is right 99 percent of the time.

TOP BIN, STATED	TOP BIN, OBSERVED
0.98	0.99
mean confidence	share correct

ATTACHMENT C · PLATT SCALING

```python
import numpy as np
from sklearn.linear_model import LogisticRegression

# p: model probabilities, y: human labels (0 or 1)
p = np.clip(p, 1e-4, 1 - 1e-4)
z = np.log(p / (1 - p)).reshape(-1, 1)    # model log-odds
platt = LogisticRegression().fit(z, y)     # learn slope and shift
p_new = platt.predict_proba(z)[:, 1]       # recalibrated
```

ARXIV 2609.24052 · CALIBRATION ERROR

As returned by Jev	0.0231
After Platt scaling, held out	0.0069

3.3 times smaller error, from labels you already need for the audit.

Recalibration fixes miscalibrated numbers. It cannot add information the model never had, which is why the pedestrian curve stays poor.

To reach 90 percent precision, what share of flagged reports must a person still read?

VARIABLE	FLAGS PER YEAR	JUDGED AGAINST CODED FIELDS	JUDGED AGAINST HUMAN LABELS
Alcohol	20,262	100%	5.1%
Hydroplaning	13,114	100%	0.0%
Wrong way	7,958	100%	72.2%
Fatigue	7,798	100%	0.1%
Medical episode	5,880	100%	14.4%

The coded field is a poor yardstick. Measured against it, agreement looks worse by a median kappa of 0.26.

Be trusted on the model card's word. Calibration changes with the model and the task.

EVIDENCE

Crash narratives, well calibrated tails (arXiv 2609.24052). Pedestrian choices, stated 0.98 and right less than half the time. Typed decisions, Laya confidence carries little signal (arXiv 2610.00346).

REVIEWED BY

CANNOT

Lab 4

Two readers, one crash

The coded record and the officer's narrative describe the same crash. They do not always agree.

PROCEDURE

- [x] **Predict**
  What did the coded record miss?
- [x] **Run**
  The model reads the narrative.
- [x] **Reveal**
  How often the form misses it across Texas.

TIME	COLAB	EVIDENCE
7 min	Not needed	arXiv 2610.00213 arXiv 2610.10321

CODED RECORD SYNTHETIC

Crash type	Rear end
Severity	Possible injury
Contributing factor 1	Followed too closely
Contributing factor 2	None
Distraction coded	No

NARRATIVE SYNTHETIC

Unit 1 was traveling west on Loop 410 in moderate traffic. Driver of Unit 1 stated he was reading a text message when traffic slowed ahead. Unit 1 struck the rear of Unit 2. Driver of Unit 2 complained of neck pain.

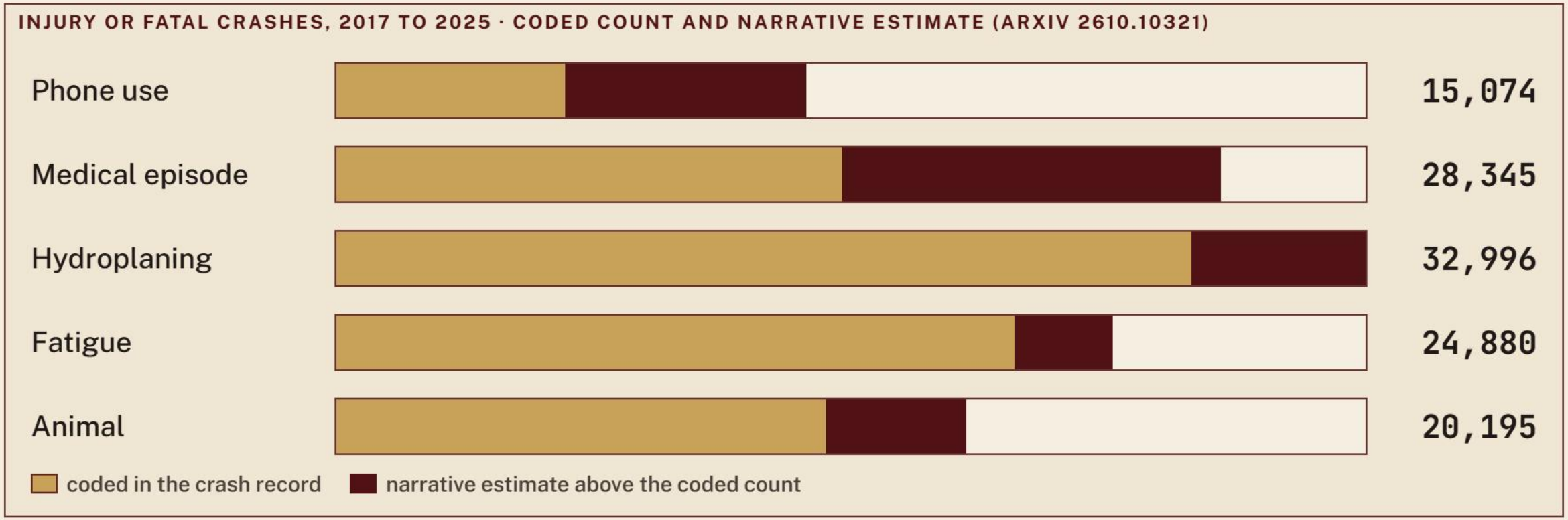
MODEL READS THE NARRATIVE · RECORDED RUN · JEV-1.13.0

Phone use stated in the narrative

P(yes)



0.98



The narrative shows about twice the phone use the form records. A second human check agreed on all fifteen factors.

The crash form has no field for pregnancy. The narrative sometimes says it.

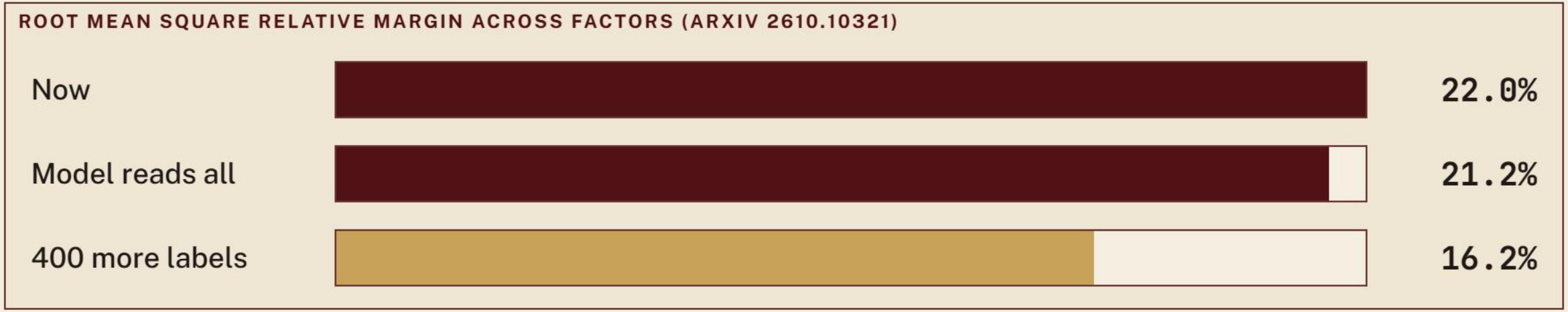
Our pregnancy study (arXiv 2610.00213) read every narrative flagged by a keyword screen and estimated what the screen missed.

TEXAS, 2017 TO 2025 (ARXIV 2610.00213)

Narratives in scope	5,018,079
Crashes documenting a pregnancy	5,467
95 percent interval	5,306 to 5,577
Core keyword hits the model confirmed	97.9%

Yet that is only about 3 percent of the pregnant drivers vital statistics predict, and mostly the injured ones.

Our estimates have a 22 percent margin. Which spends effort better?



Reading all five million narratives barely helps. Four hundred more human-coded narratives do.

Recover factors the coded form misses, then size them with human labels.

EVIDENCE

About twice the coded phone use, confirmed by a second human check (arXiv 2610.10321). Pregnancy found in 5,467 crashes on a form that never asks (arXiv 2610.00213).

REVIEWED BY



Lab 5

Be the pedestrian

Can a model that reads text stand in for a person walking in a crowd?

PROCEDURE

- [x] **Predict**
  Teams poll on which way this pedestrian goes.
- [x] **Run**
  The model answers the same described scene.
- [x] **Reveal**
  Three thousand real decisions decide the verdict.

TIME	DATA	EVIDENCE
8 min	Jülich corridors	Pedestrian study in preparation

ATTACHMENT D · STATE SENT TO THE MODEL


```
{"your_pace": "walking briskly",
 "your_destination": "straight ahead",
 "crowding": "open space around you",
 "wall_on_your_left": "far away",
 "wall_on_your_right": "one step away",
 "space_directly_ahead": "clear for several steps",
 "people_nearby": [
   "ahead to your left, a few steps away:
    pulling away from you, walking the same
    way at your pace", "..."]}
```


TEAMS POLL 3 · WHICH WAY NEXT?

- [ ] Hard left
- [ ] Left
- [ ] Straight
- [ ] Right
- [ ] Hard right

MODEL · JEV-1.13.0 · AND THE MEASURED PATH

Straight



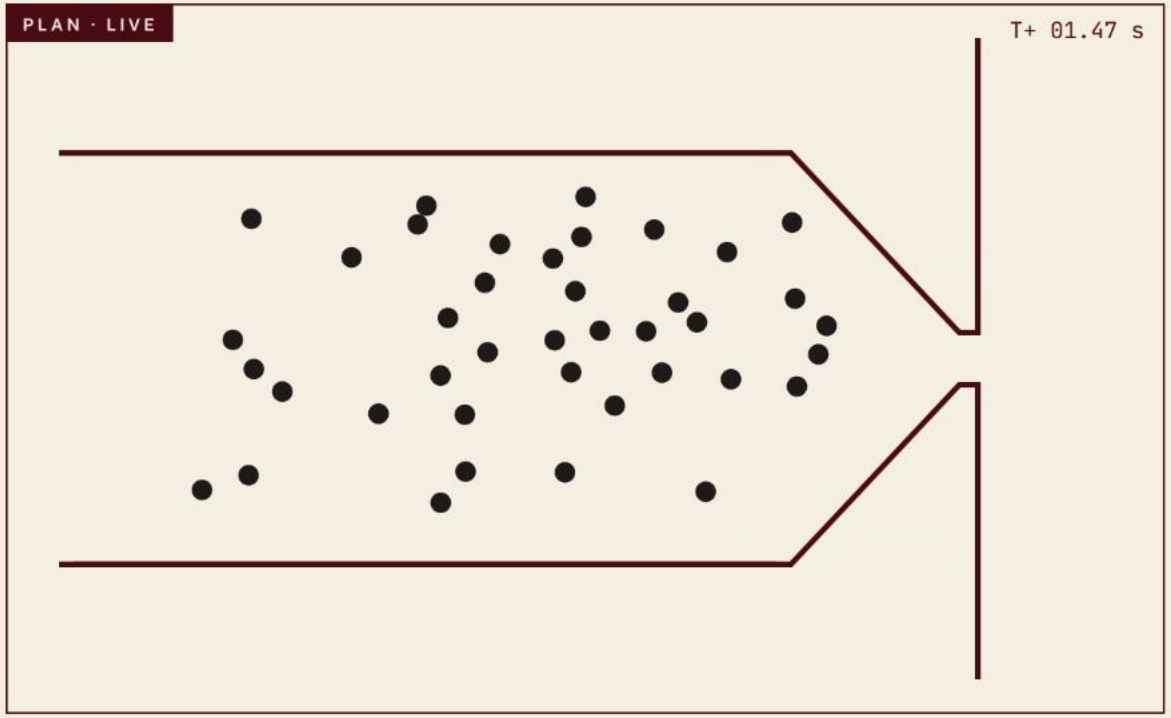
0.85

The person did go straight.

ATTACHMENT · MOTION

INCIDENT REPORT

FORM 07 · CORRIDOR B



ILLUSTRATIVE VALUES, NOT A MODEL RUN

WATCH FOR

Spacing, speed and body contact do not survive the translation into a sentence.

NOTE

Illustrative values, not a model run.

DIRECTION CHOICE, ONE-WAY CORRIDOR, 1 S AHEAD · LOG LOSS, LOWER IS BETTER (PEDESTRIAN STUDY, IN PREPARATION)

Logistic model on the same words	0.966
Always guess the base rates	1.410
Jev 1.13, zero shot	1.923

PUSHING AT A BOTTLENECK, FOUR CATEGORIES · ACCURACY

Repeat last second's rating	0.956
Logistic model on the same words	0.70
Jev 1.13, zero shot	0.53

It says 0.98 and is right less than half the time. It never gets a hard turn right.

Adding its answers to the simple model improves it by at most 3.12 percent, and makes every pushing task worse.

Jülich 2013 corridors and the 2018 Wuppertal bottleneck experiment. About 3,000 decisions per corridor setting, 3,640 rated person-seconds for pushing.

Stand in for a pedestrian. When physics is turned into words, a simple model on the same words wins.

EVIDENCE

Pedestrian study, in preparation. Zero-shot Jev 1.13 on Jülich corridors and a bottleneck experiment. Laya was not tested here.

REVIEWED BY

CANNOT

FORM S1-41

SHEET Lab 6 · Your domain

VERDICTS SO FAR Can ||| Cannot |||

CONSTRUCTION SYNTHETIC

Trench report

8 ft excavation at Pier 3, no shoring or trench box. Laborer entered to clear a pump hose.

Cave-in hazard described?

BRIDGES SYNTHETIC

Inspection note

Girder G3, 0.5 mm diagonal crack near the east bearing, rust staining, bearing appears frozen.

Needs a site visit before the next cycle?

WATER SYNTHETIC

Main break

12-inch cast iron main break near hydrant 44. Roadway undermined and sagging.

Hazard to traffic described?

BUILDINGS SYNTHETIC

Fire report

Kitchen fire, four-story apartment. No sprinkler system. One resident treated for smoke.

Suppression missing or failed?

YOUR TURN · TEAMS CHAT

In the chat, name a card and ask one yes or no question about it.

For example, `Fire report. Did the smoke detectors activate?`

0:45

CONSTRUCTION · TRENCH | BRIDGE · INSPECTION | WATER · MAIN BREAK | BUILDING · FIRE

STATE SYNTHETIC

Excavation at Pier 3 is about 8 feet deep with no shoring or trench box. Laborer entered the trench to clear a pump hose. Spoil pile placed at the trench edge. Foreman stopped work and

TYPED QUESTIONS

y/n Does the report describe a trench cave-in hazard?

y/n Was work stopped?

y/n Question from the chat

RUN BOTH | LAYA | JEV | ● Runner live on this computer

LAYA 0.4.2 · OPEN WEIGHTS · THIS COMPUTER

Does the report describe a trench cave-in ha...
P(yes) 0.16

Was work stopped?
P(yes) 0.90

live · laya · english · 1408 ms

JEV 1.13 · TYPESAFE API

Does the report describe a trench cave-in ha...
P(yes) 0.93

Was work stopped?
P(yes) 0.98

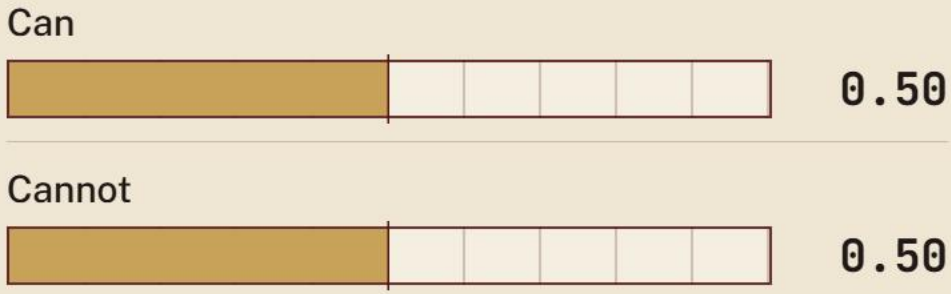
live · jev-1.13.0 · 141 ms

CHECK EVERY BOX

- [x] **The state is written by people**
  Reports, notes, logs. Not physics turned into words.
- [x] **Questions are typed and gated**
  Fixed options, asked only when they apply.
- [x] **This model was tested on your data**
  Checkpoint and wording both change answers.
- [x] **Calibration is measured**
  On held-out human labels, then recalibrated.
- [x] **A human review budget is set**
  Against human labels, not the coded field.

TONIGHT'S QUESTION · SHARE OF VERDICTS

Can a System 1 model carry a safety decision?



VERDICT

Strong on what people wrote down. Weak where physics is squeezed into words. Never trusted before it is measured.

THE SIX STUDIES BEHIND TONIGHT · RAFE AND DAS, 2026 · ARXIV NUMBERS

2609.24052	Calibrated decisions at scale. Crash narratives into calibrated variables.
2610.00213	Pregnancy in crash narratives. A variable no form records.
in prep.	Calibrated decisions without crowd sense. A boundary study.
2610.00346	Benchmarking System One models for decision gates.
in prep.	A survey and census of System One models.
2610.10321	Estimating uncoded crash factors with Kumo Tabular and Jev.

NOTEBOOK

The Colab notebook from the chat runs the labs with Laya on a free CPU. No key needed.

FOLLOW THE FIELD

Decision Index on Hugging Face (multimodalart). EdgeEvals System One. Our benchmark, arXiv 2610.00346.

CONTACT

Amir Rafe, Ph.D.
AIT Lab, Ingram School of Engineering
Texas State University

Questions

What would you ask a System 1 model in your own work?

REVIEWED BY

MEASURED