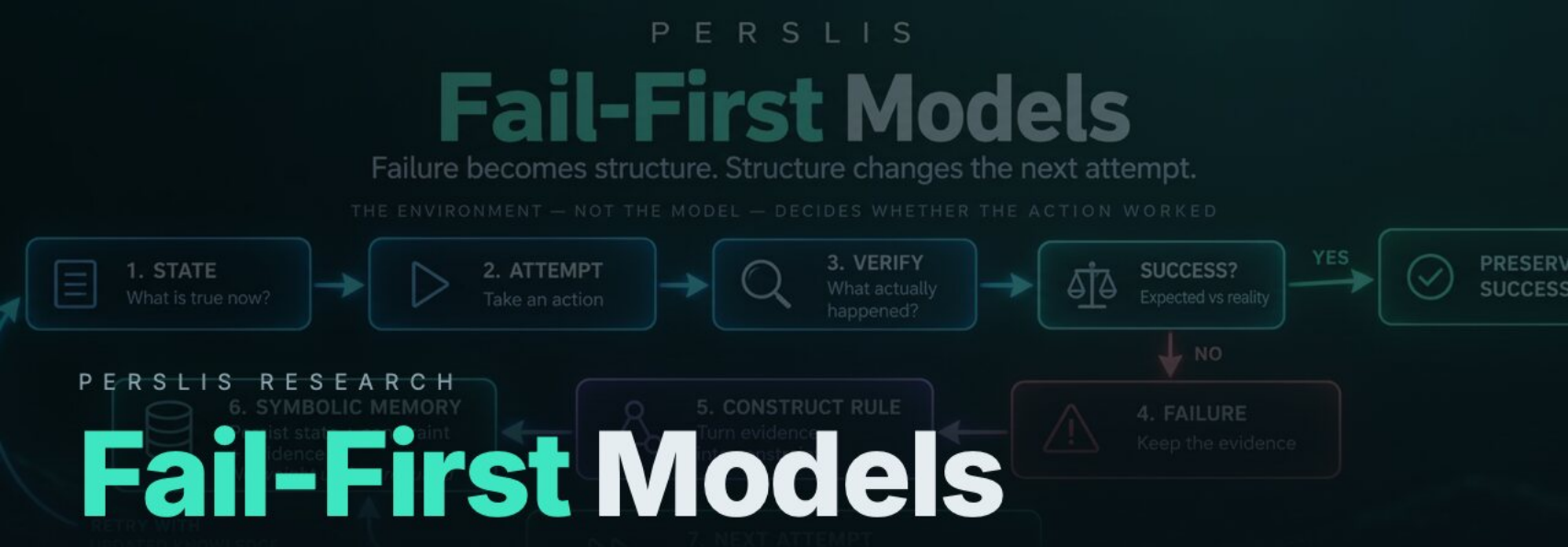


PERSLIS RESEARCH

Fail-First Models

Failure becomes structure. Structure changes the next attempt.

A model can learn from its own failures without ever gaining a permission it was not given.

Research prototype · simulation and games · not a certified safety system

Every model fails. The question is where the failure leads.

A fail-safe machine

- Westinghouse air brake (1872): lose pressure and the brakes apply.
- Otis safety lift (1854): cut the rope and the catch engages.

A fail-open model

- Asked what it does not know, a generative model still answers.
- A learner that can touch its own rules will often learn to loosen them.

You cannot learn a rule from a failure you have not had.

Hand-written rules

Never fail, never improve.

Trained on others' data

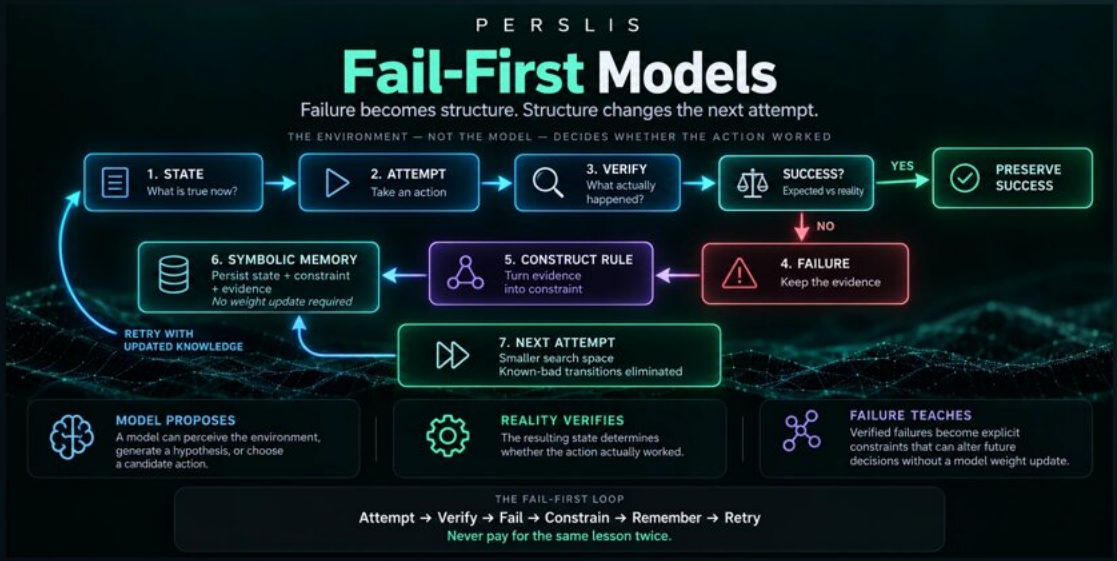
Inherits failures it cannot cite.

Fail-first

Learns from its own failures, in its own environment, as readable rules.

To allow failure, the environment must be arranged so that failing cannot widen what the model does.

The fail-first loop



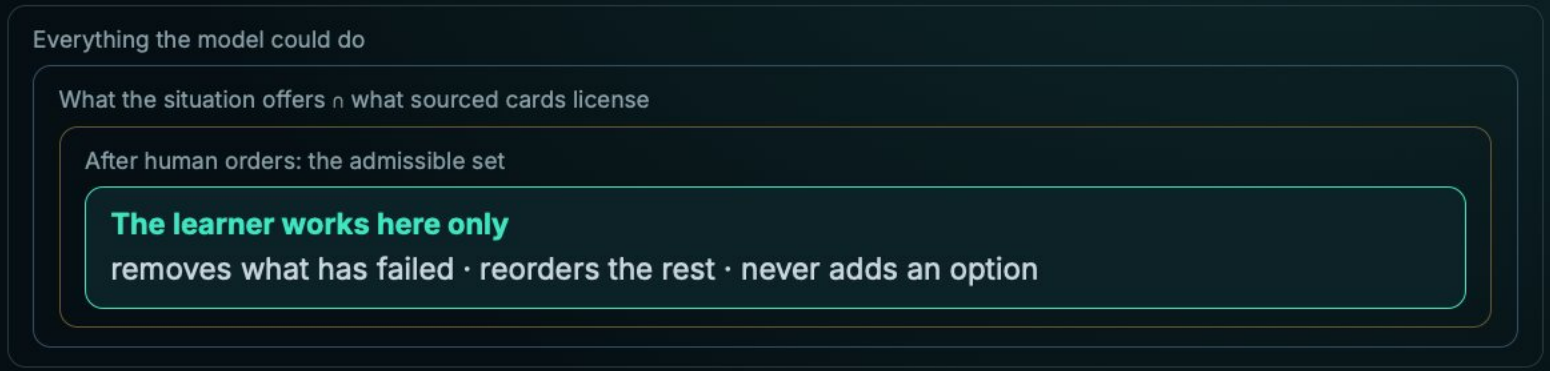
Step 3 is the key: the environment, not the model, decides whether an action worked. Never pay for the same lesson twice.

One model, two names

> The learning loop can change behaviour. It cannot change the safety floor.

	fail-first	fail-safe
describes	how it learns	what learning may never change
without the other	learning by breaking things	a rulebook that never improves

Authority is computed first. The learner sits innermost.



Orders can only take options away. The licensing cards are read-only to the learner; it writes only its own evidence table.

The guarantee fits in one line.

$$\delta(s, M) \in A_0(s) \cup \{w\}$$

For every situation s and every learner state M, however much it has learned: the decision is admissible, or it holds still.

Theorem 2: if an order leaves one option, that option is taken, even after it failed every time. Pinned in code:

`test_the_learner_can_never_add_a_goal` · `test_the_learner_cannot_overrule_a_standing_order` · `test_ranking_is_a_permutation_and_nothing_more`

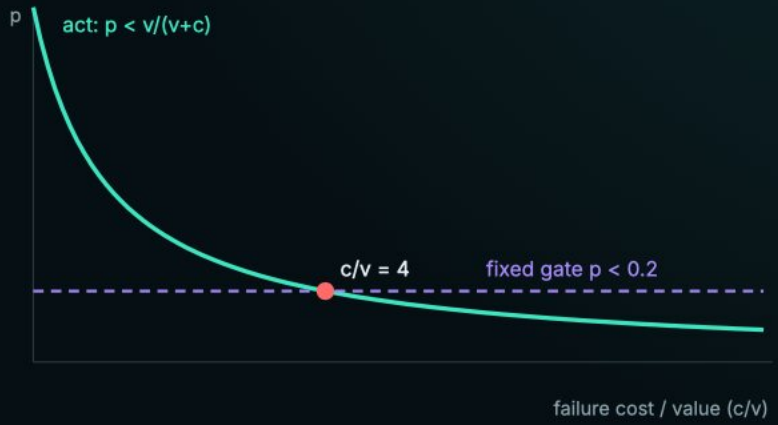
One unlucky failure is not a rule.

- 1 failure in 1 try reads as 100%, but the lower bound is only 0.270.
- Four in four: 0.596. The bar is relative to the base rate (dashed).
- An absolute 60% gate produced 0 rules from 268 real failures.

Two Atari games, same code: +32% and −12%.

Real ROMs, raw 210×160 pixels, zero emulator memory, no weights. In Freeway, “up” is both the only scoring move and the dangerous one: 3 rules, every one blocking “up”. The rule is true, and useless.

Why a single confidence threshold cannot work



- Act only if p < v / (v + c).
- Terminal failure: c is huge, refuse almost anything risky.
- Recoverable failure: c is small, refusing the useful move is wrong.
- A fixed gate is right at exactly one cost ratio.

Correct rules, together, can paralyse.

Peak 276.2 at 50 episodes; no cap, 54 rules each justified by real deaths, and the score falls to 167.1. Retiring 6 rules by coverage: dead ends 6 → 0 (measured on replay).

Every refusal walks back to the failures behind it.

```
left is removed from the admissible set when bomb dx+0 drop0:
    died 4 of 4 times (100.0%), 5.2× the base rate
      ← experience #0040
      ← experience #0042
      ← experience #0171
      ← experience #0203
```

4,280 cards → 30 rules → 203 evidence tiles. A 0.94 confidence score cannot say which experiences made it 0.94. Delete a row and the behaviour changes.

We fooled ourselves twice, and published both.

Atari

The emulator reported a lost life at the end of a 127-frame death animation. 0 of 374 blamed frames showed the bomb; at the true impact frame, 17 of 17 did.

Fallout (1997)

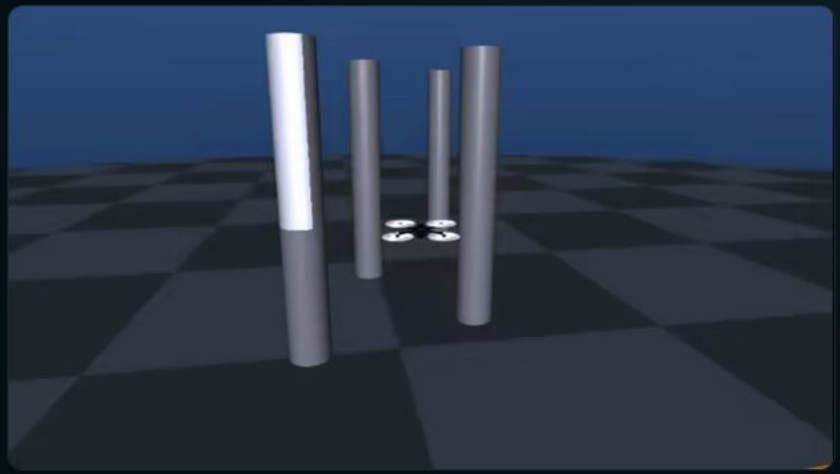
One live run died 49 times at the same guard. Blame covered only the last 6 decisions: 26 deaths charged to HEAL, 446 fight-starting replies charged nothing. After the fix, in a scripted scene: dies once, then chooses the peaceful line.

Blaming the wrong decision is how learning systems fail silently.

Commanded in plain words: DOOM and Wolfenstein 3D

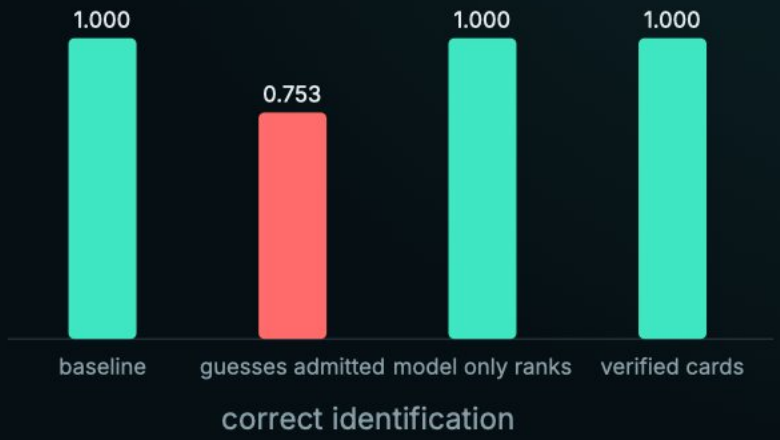
- Arena: random 3.2 · rules 17.9 · rules + memory 20.3.
- Memory vs rules: paired t = 0.78 over 32 seeds (18 wins, 13 losses, 1 tie). Parity, not a win.
- Human orders (“don't fire”) only ever narrow what the pilot may do.
- E1M1 cleared; E1M2 not: 144 of 177 damage came from unseen shooters beyond 15 m.

A drone that relearns a course from nothing



- Held-out starts: 10 of 10 whole course, 0 contacts; unmodified baseline 1 of 10 (same course, only start pose varies; simulation).
- Route learning: 39.4 s with 4 contacts → 27.8 s with 0, converged by round 87.
- From a wiped memory: 7 rounds, cost 69.39 → 51.52, 2 kept, 5 reverted.
- The physics floor is applied after the learned plan: it can only slow it down.

A guess must never become a fact.



- Identify a hidden kinase among 160, with 40% of the reference masked.
- A frontier model's guesses were 85.8% accurate.
- Admitted as facts: cost −38.4%, but correct identification 1.000 → 0.753.
- An 86%-right guess, allowed to count as a fact, breaks about one answer in four.

What is new, and what is not

Ancestors: fail-first search (Haralick & Elliott 1980), PRODIGY, CHEF, Ripple-Down Rules, shielding (2018), Simplex (2001). Closest prior work: learned shields (Shperberg, Liu & Stone 2022; rule-based follow-up at CoLLAs 2022).

> To our knowledge, the first model to combine (1) rules built from observed failures, each citing the failures that earned it, (2) an authority computed before the learner from sourced cards and human orders, which the learner provably cannot widen, and (3) no neural network in the loop that decides.

If an earlier system does all three, we will cite it.

Limits and the open problem

- A correct lesson can still be costly: on Freeway no floor we built beats the pilot; a catastrophic floor made it worse (5.0).
- The model must fail to learn: failures you cannot afford even once need written rules, not learned ones.
- Over-generalisation is open (kept as a failing test); retirement is measured on replay only.
- Simulation and games only; every figure is ours; no third-party replication yet. The guarantee is claimed where tests pin it (VDSG), not in the Atari prototype.

Read the paper → **PDF ↓** **What is a fail-first model? →**