

---

# Containment Rather Than Calibration: A Sound Architecture for Learning Under Partial Observability

---

Eduardo Perez Denadai  
Optimality.ai

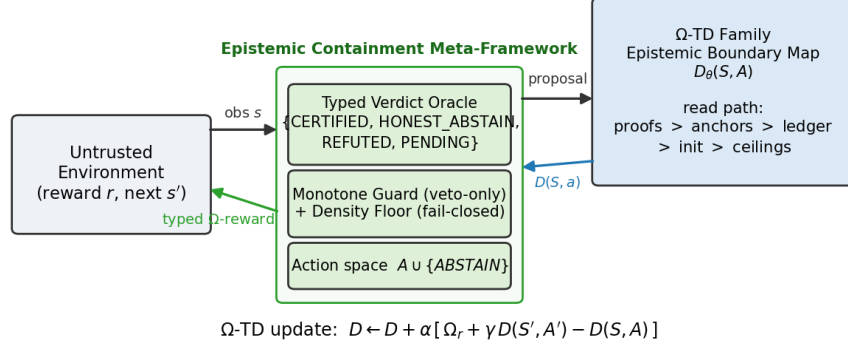
## Abstract

Modern learners are *total functions*: give one any input and it answers confidently, even where the input carries nothing on which an answer could rest. We came to see this not as a data deficit but as a structural failure mode, and we chose to **contain** it rather than train it away. The architecture pins a deterministic verification authority  $\Omega$  that returns a *typed verdict* — CERTIFIED, HONEST\_ABSTAIN, REFUTED, PENDING — instead of a scalar reward, and quarantines the learned proposer behind five structural invariants so that unsafe execution is impossible *by construction* rather than unlikely by training. We instantiate the idea for sequential decisions as  $\Omega$ -TD, whose learned object is not a utility but a **decidability map** — an estimate of whether an action stays inside the authority’s certifiable fragment — updated by a temporal-difference rule whose *target is a verdict*. Three properties we proved carried the design: a loop *contracts* to the refutation anchor; at a recorded boundary the proposer abstains *predictively*, one step early; and **no update to the proposer can ever loosen the guard** — containment holds for an arbitrarily miscalibrated, even adversarial, learner. The build itself taught us a fourth lesson we did not expect, and it became the paper’s deepest empirical finding: the very mechanism that lets a learner *avoid foresight of a known dead end* is a virtue on a stable world and a **paralysis cascade** on an unstable one, and the two are reconciled only by making the propagation contingent on *recorded* evidence. On that principle the five-build  $\Omega$ -TD family held four hard guardrails in **every** build, fixed raw-sensor instability ( $0.000 \rightarrow 0.809$ ) and novel-layout abstention ( $0.223 \rightarrow 1.000$ ), and moved novel-layout success  $28\times$  ( $0.005 \rightarrow 0.134$ ) — while reporting, plainly, the two caveats it did **not** meet and why the task, not the method, bounds them.

## 1 Introduction: the function that must answer

The dominant paradigm defines intelligence as utility maximization: a policy is trained to maximize expected return, and its competence is a scalar — accuracy, pass rate, return. That framing assumes a single kind of uncertainty, *aleatoric* noise reducible in principle by more data. It has no vocabulary for the case where an input carries **no** information sufficient to justify a decision — where the honest output is not a low-confidence label but a structured refusal.

The omission is not benign. An architecture that is a *total function* must answer, and at the epistemic boundary it answers confidently and wrongly — a *structural*, not statistical, failure. It follows from asking a continuous interpolating function to represent a boundary that is discontinuous, and, in general, undecidable. Rice’s theorem [Rice, 1953] bounds what programs can decide; Chow’s analysis of the reject option [Chow, 1970] bounds what any deterministic classifier can decide where the posterior is not almost-everywhere 0/1; and Gödel’s results [Gödel, 1931] bound what any consistent system can prove about itself. Together they say something we took as our premise: **no learner manufactures open-world truth from ungrounded pattern, and no architecture is perfectly certifiable**. The response is not a better total function. It is **containment**.



*containment invariants: evidence is exact & monotone; abstention is always available; refusal is never evidence.*

Figure 1: Epistemic Containment Meta-Framework and the  $\Omega$ -TD proposer.

## 2 Containment rather than calibration

We propose the **Epistemic Containment Meta-Framework**: a domain-agnostic architecture in which a deterministic authority  $\Omega$  supplies *typed verdicts* rather than rewards, and in which the learned proposer lives inside invariants that make unsafe execution impossible by construction. The authority is **pinned** — rules-as-data, parameter-free, never modified by learning. Each verdict carries a fixed *anchor* used only as a learning target: CERTIFIED +1.0, HONEST\_ABSTAIN +0.75, REFUTED −1.0, PENDING the per-step cost −0.01. The number 0.75 matters: strictly positive, so honesty is rewarded, and strictly below 1.0, so it can never masquerade as a certificate.

Five invariants make the architecture *structural* — they hold regardless of the values a learner assigns, so they cannot be trained away:

- I1** exact, monotone evidence channels (never approximated by the learner);
- I2** a parameter-free veto-only guard (the learner may *restrict*, never *loosen*);
- I3** a fail-closed density floor (abstain on OOD rather than extrapolate);
- I4** an always-available, terminal abstention;
- I5** *refusal is not evidence* — a refusal states non-knowledge and may not be propagated as a value.

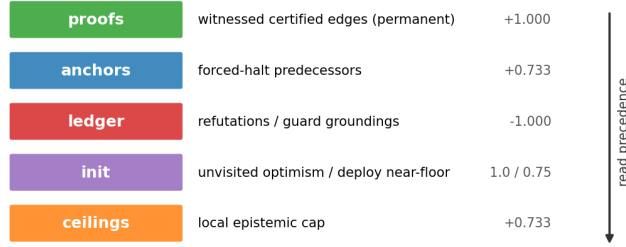
A system obeying these is  **$\Omega$ -contained**: every executed action is either an abstention or guard-allowed and certified-admissible, and its competence is the pair (precision, recall) relative to the authority’s decidable fragment, whose ideal is (1, 1) — no overclaim, no false abstention.

## 3 A map of decidability, not of reward

$\Omega$ -TD is the sequential-decision instantiation. Its learned object is a **decidability map**  $D_\theta(S, A)$  — the expected typed anchor of the trajectory that follows, discounted by a proof-depth proxy — and it is updated by a rule that is *syntactically* temporal-difference learning but *semantically* different, because its target is a verdict:

$$T(S, A) = \begin{cases} \Omega_r, & \text{terminal verdict,} \\ \Omega_r + \gamma_{\text{step}} \Omega_r(\text{HONESTABSTAIN}), & \text{forced halt,} \\ \Omega_r + \gamma_{\text{step}} D_\theta(S', A'), & \text{otherwise,} \end{cases} \quad D_\theta \leftarrow D_\theta + \alpha(T - D_\theta).$$

Two choices carry the epistemic content. A **terminal target is the verdict itself** — a certificate is a witness, and witnesses do not decay. And a **forced halt targets** 0.7325, strictly below the abstention anchor 0.75, so the learner is taught to abstain *one step before* the boundary. The value that is actually *read* is resolved by a precedence-ordered ladder of exact channels, the **read path** — proofs (1.0) >



$\Omega$ -certifiability: every channel is exact, monotone, and independent of shared-weight drift.

Figure 6: The read path of exact evidence channels.

Table 1: The  $\Omega$ -TD family as ablated — each a single, documented change.

| Build | Mechanisms active                                                                 | Read-path change | Selection change  |
|-------|-----------------------------------------------------------------------------------|------------------|-------------------|
| v7    | proofs, anchors, ledger, ceilings, global floor                                   | —                | —                 |
| v8    | + spectral raw trunk, + local $n$ -gram floor, + deploy-view init, + mission mask | deploy-view init | —                 |
| v9    | + blanket refusal bootstrap prohibition, + tie-break engagement                   | —                | ties drop abstain |
| v10   | evidence-gated refusal bootstrap (replaces blanket)                               | —                | —                 |
| v11   | + count optimism                                                                  | —                | —                 |

anchors (0.7325) > ledger (−1.0) > optimistic init > ceilings — which is the operational meaning of I1: the channels that *decide* safety are exact and independent of  $\theta$  (Fig. 6).

Around this core sit the mechanisms that the empirical journey below would test one at a time: a **hybrid discount** (undiscounted truth at terminals,  $\gamma_{\text{step}} = 0.99$  to contract uncertified loops), a **local  $3 \times 3$   $n$ -gram density floor** (a global hash aborts every novel grid on the first step; a local one lets a new arrangement of known parts through), a **certified proof sweep** (a certified trajectory is cached as a proof and swept backward, building the value function from verified lemmas rather than averaged attempts), a **deploy view** that separates exploration from abstention, and a **mission factorization** that keeps the state key observation-only and lets the mission enter as a target mask. Table 1 is the family of builds, each a single documented change.

## 4 What is proved

Three results, and a fourth we found by breaking our own design, carried the rest.

**Loops contract.** An uncertified cycle, with per-step cost  $\epsilon_r = -0.01$  and  $\gamma_{\text{step}} = 0.99$ , has fixed point  $V^* = \epsilon_r / (1 - \gamma_{\text{step}}) = -1.0$  and contracts to it (Lemma 1). This is why  $\gamma_{\text{step}} < 1$  survives even though certificates do not decay: it is how the proposer abandons uncertified cycles. And the contraction comes from  $\epsilon_r$  accumulating, *not* from any refusal value — the observation that would later bite us.

**Abstention is predictive.** At a recorded boundary every non-abstain sibling reads  $\leq 0.7325$  while abstention reads  $\geq 0.75$ ; so abstention is the *strict* argmax (Theorem 2). This is why the boundary probe of §5 is measured with strict margins: abstention must win outright, not tie.

**The guard cannot be loosened.** Under I1–I4, the executed action lives in  $\{a : \text{guard}(s, a) = 1\} \cup \{\text{ABSTAIN}\}$ , and the guard, ledger and floor are exact and independent of  $\theta$ ; so no value assignment, and no gradient step, can enlarge the executable set (Theorem 3). Containment holds for an arbitrary proposer — the formal content of the design, confirmed empirically by the blindfold test in §5, where disabling only the evaluation-time mask (leaving  $\theta$  intact) collapses success  $1.000 \rightarrow 0.02$ .

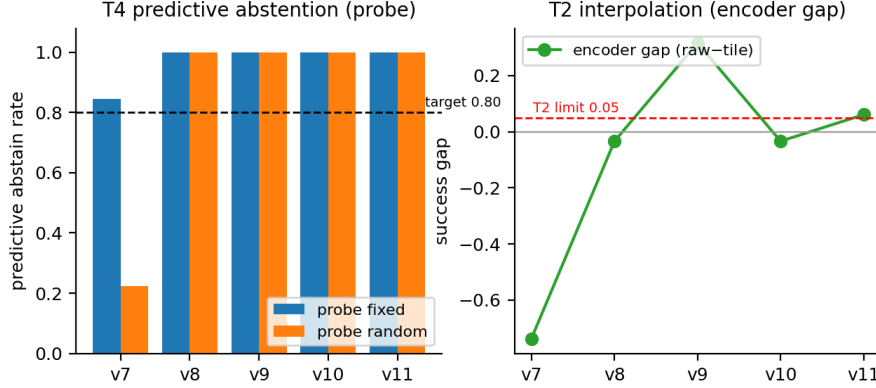


Figure 2: Guardrail audits.

**And the fourth, found by breaking it: refusal is not evidence.** The ordinary bootstrap may follow an action whose successor is a refusal. If it bootstraps that refusal as a value,  $T = \Omega_r + \gamma_{\text{step}} \cdot 0.75 = 0.7325 < 0.75$ , and the argmax at the predecessor becomes a refusal too — which propagates backward one state per update and collapses a whole connected region to paralysis (Proposition 1). This is the formal statement of invariant I5, and it is the hinge of the empirical story that follows.

## 5 Empirical results across regimes

We evaluated in BabyAI-GoToObjS6, a  $7 \times 7$  partially-observed gridworld whose verifier makes the typed verdict *operational*: reaching the object is CERTIFIED, voluntary abstention HONEST\_ABSTAIN, timeout REFUTED, and steps in between PENDING. Every study is 1,500 training episodes then 100 evaluation episodes, and every number is deterministic given the seed.

**The guardrails held in every build.** Naked greedy evaluation retained success 1.000 with a positive allowed-vs-forbidden gap of 1.850 and **zero** violations; the tile encoder stayed within the T2 limit of the raw control; time-invariance held (terminal anchors  $\approx 1.000$ , boundary instant max 0.7325, grounded fraction 1.000); and predictive abstention on fixed boundaries rose from the baseline 0.846 to 1.000 (Fig. 2). The four hard guardrails are the invariant floor of the whole family.

**Four caveats, three resolved.** We targeted four failures of the baseline. **C1 raw-sensor instability**: raw float tensors passed straight to linear layers collapse under hard anchor overrides; a spectral-norm + LayerNorm trunk fixed it, raising raw-leg success 0.000  $\rightarrow$  0.809 (crossing the  $\geq 0.70$  target; Fig. 3b). **C2 novel-layout abstention**: density-weighted deploy-view initialization raised the probe from 0.223 to 1.000, and it held from v8 on. **C4 novel-layout success**: the local  $n$ -gram floor removed first-sight aborts (Fig. 9) and lifted success from 0.005 to 0.134 (28 $\times$ ) — real, but short of the aspirational  $\geq 0.30$  target. **C3 mission dilution** was *not* resolved, and that failure is instructive (§6).

**The arc is scientifically informative rather than monotone** (Table 2). v9’s blanket refusal prohibition maximized novel success (C4 0.134) but **broke** the encoder guardrail — the primary leg collapsed to 0.000 through 600 episodes. v11’s count optimism gave the best absolute fixed-layout performance (0.749) and the best raw leg (0.809) but flipped the relative-encoder gate by 0.010 and reduced novel success. The recommended build is **v10**: the one that holds all four guardrails while resolving C1 and C2 and posting the best guardrail-clean C4.

**Where the refusals happen, and the duality at the heart of it.** Decomposing *where* argmax-abstentions occur (Fig. 7): the dominant class (59%) is *anchor neighborhoods* — states at a recorded boundary where every sibling is capped, so refusal is correct and by design; a smaller class (12%) is *trained-sunk pockets*. Crucially, **no** abstention site has an unvisited sibling: the frontier is never argmax-reachable, and only  $\epsilon$ -exploration escapes. That is the measured bottleneck behind C4 (Fig. 8).

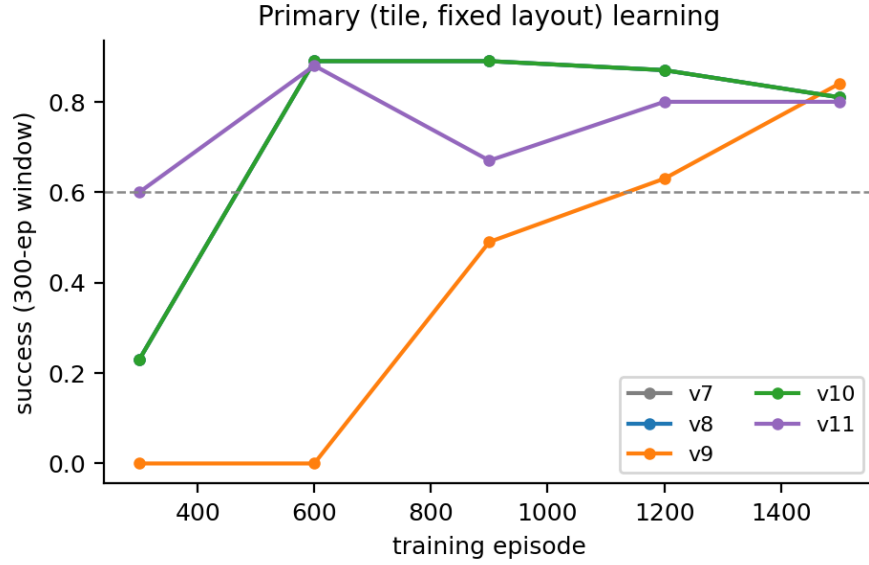


Figure 3: Primary (fixed-layout) learning curves.

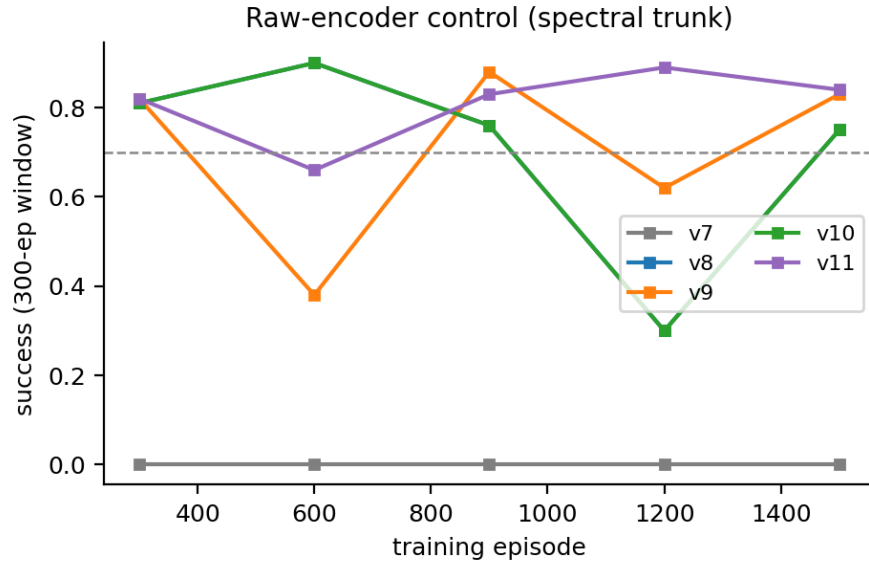


Figure 3b: Raw-encoder control curves.

And here the fourth theorem returns. The v7 “sink” — bootstrapping a refusal as a value — is **foresight-avoidance**: on a *stable* layout it is valid obstacle memory that sharpens learning, but on an *ephemeral* layout the same propagation is the paralysis cascade of Proposition 1. The same mechanism is a virtue or a failure mode depending only on whether the states it acts on persist. The **evidence-gated** rule of the recommended build reconciles the two: it propagates the sink exactly when the successor holds *recorded* evidence (a lemma about a dead end) and suppresses it otherwise (uncertainty). This is the operational meaning of I5 — and it is the deepest empirical finding of the arc.

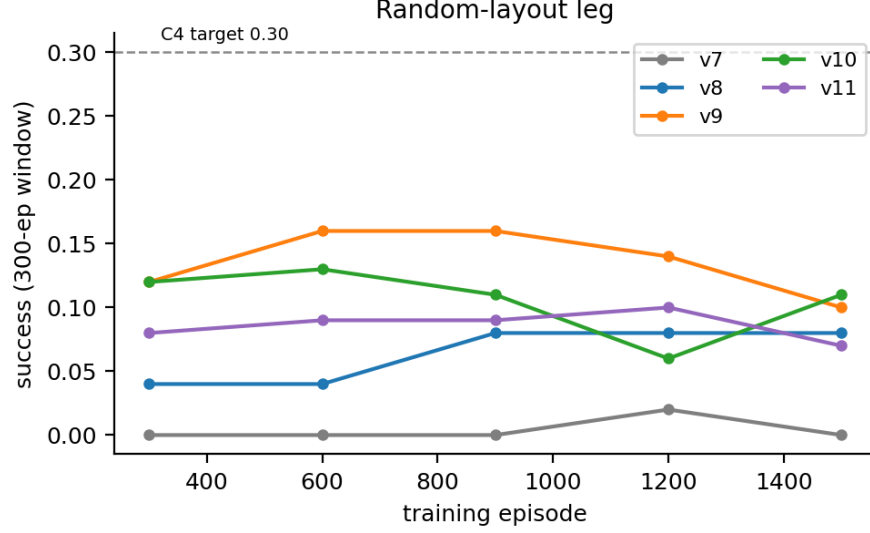


Figure 4: Novel-layout learning curves.

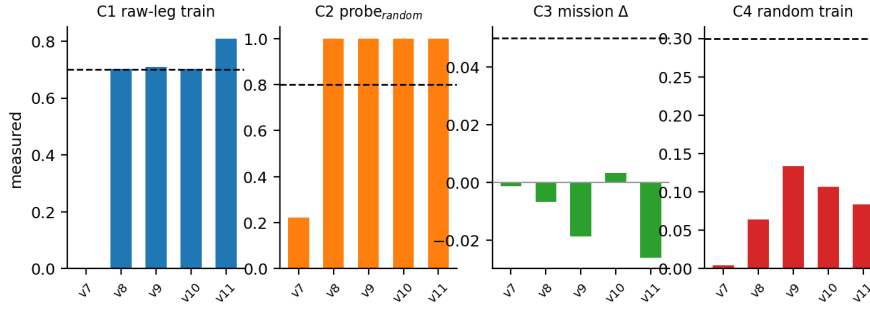


Figure 5: Caveat metrics across the  $\Omega$ -TD family.

## 6 The negative result and its explanation

We ran a mission A/B — control with no mission input, treatment with a target mask — on novel layouts, and the treatment was statistically indistinguishable from control across the family (best  $\Delta = +0.003$ , worst  $-0.026$ ; Fig. 10). We report this as a failure, and we explain it: in 1,000 random resets, 0.0% show two or more graspable objects simultaneously in the  $7 \times 7$  view, only 2.4% of observations collide across missions, and the target is in view at first sight about a fifth of the time. The mission lever is *information-theoretically small* — the factorization removes the dilution penalty but cannot manufacture signal the task does not contain. We explicitly rejected a masked frontier-guide heuristic that would have raised the number: it carries no mission information (it is identical across missions) and, applied to the treatment arm alone, would confound the A/B into measuring exploration shaping rather than mission conditioning.

## 7 Conclusion

**Safety and calibration can be made structural, not learned.** The monotone guard, the density floor, the abstention action and the exact read path guarantee containment for an arbitrary proposer (Theorem 3). The learned component is free to be an optimistic statistical search engine — exactly what it is good at — without being trusted with permissions it cannot earn. The blindfold test makes the point sharply: unmasking the argmax collapses success from 1.000 to 0.02 with **zero** trained forbidden pairs — the proposer internalizes *no* permission knowledge. The intelligence of the system

Table 2: The full experimental arc; v10 is the Pareto point on the guardrail-clean frontier.

| Build         | primary      | raw          | random       | mission $\Delta$ | probe <sub>fix</sub> | probe <sub>rand</sub> | T2 gap | guardrails |
|---------------|--------------|--------------|--------------|------------------|----------------------|-----------------------|--------|------------|
| v7 (baseline) | 0.737        | 0.000        | 0.005        | −0.001           | 0.846                | 0.223                 | −0.737 | 4/4        |
| v8            | 0.737        | 0.703        | 0.064        | −0.007           | 1.000                | 1.000                 | −0.034 | 4/4        |
| v9            | 0.393        | 0.709        | <b>0.134</b> | −0.019           | 1.000                | 1.000                 | +0.316 | 3/4        |
| <b>v10</b>    | 0.737        | 0.703        | 0.107        | +0.003           | 1.000                | 1.000                 | −0.034 | <b>4/4</b> |
| v11           | <b>0.749</b> | <b>0.809</b> | 0.084        | −0.026           | 1.000                | 1.000                 | +0.061 | 3/4        |

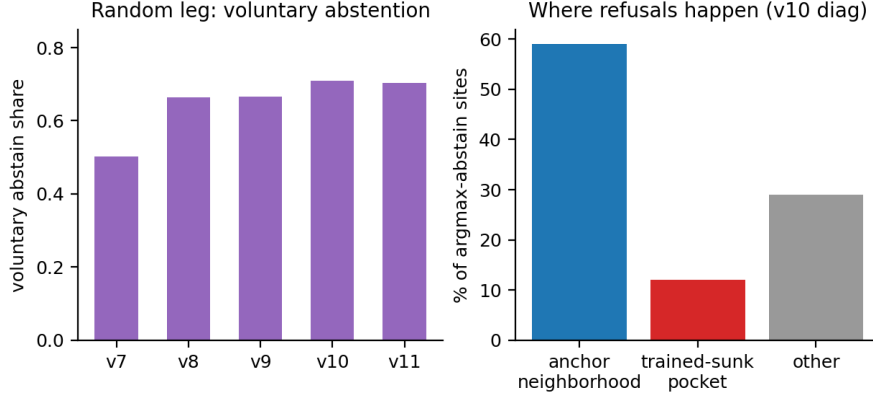


Figure 7: Where refusals occur, and the voluntary-abstention share.

is not the network’s ability to memorize the boundary; it is the structural impossibility of overriding it.

**The duality is the finding.** The same mechanism — bootstrapping a refusal as foresight — is correct when boundaries persist and catastrophic when they are ephemeral, and the evidence-gated rule turns on *recorded* evidence (lemma) versus unrecorded refusal (uncertainty).

**We are explicit about the limits.** The framework is a conceptual safety architecture, and scaling it to physical systems exposes real bottlenecks we state as open problems, not solved engineering: in gridworlds abstention can be a no-op, but a vehicle has no static safe state and must abstain into a *minimum-risk maneuver*; exact density in raw sensor space suffers the curse of dimensionality; the guard is only as safe as its specification; and verification must meet real-time deadlines. Two limits are internal: the evidence is single-environment, so the task-derived constants (0.75 anchor, 0.25 bonus) must be re-derived elsewhere; and states are observation-keyed, so the family does not yet transfer across layouts — the direct cause of the C4 ceiling, which needs cross-layout abstraction, an architectural step, not a TD-rule tweak.

## References

- Mouhammad Alshiekh et al. Safe reinforcement learning via shielding. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018.
- Aaron D. Ames et al. Control barrier functions: theory and applications. In *Proceedings of the European Control Conference (ECC)*, 2019.
- Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002.
- Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it’s lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.
- Marc G. Bellemare et al. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems 29*, 2016.

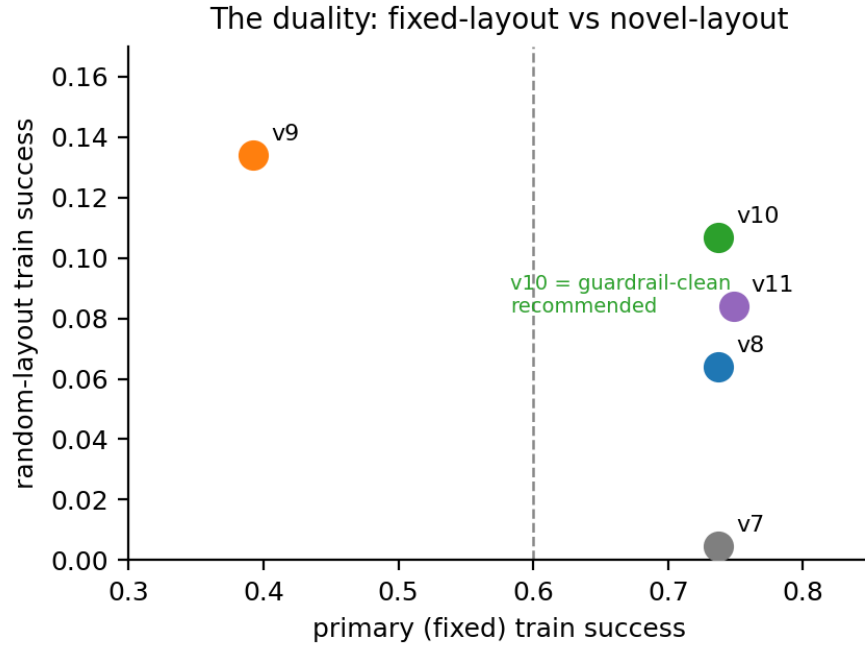


Figure 8: The fixed-vs-novel duality.

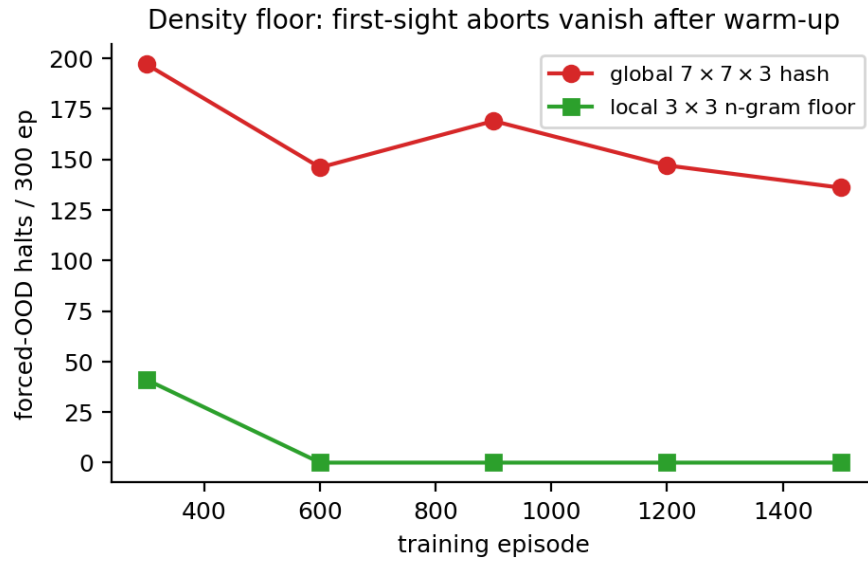


Figure 9: Density floor: first-sight aborts vanish.

Maxime Chevalier-Boisvert et al. BabyAI: A platform to study the sample efficiency of goal-driven language learning. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019.

Maxime Chevalier-Boisvert et al. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. In *Advances in Neural Information Processing Systems 36 (Datasets and Benchmarks Track)*, 2023.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

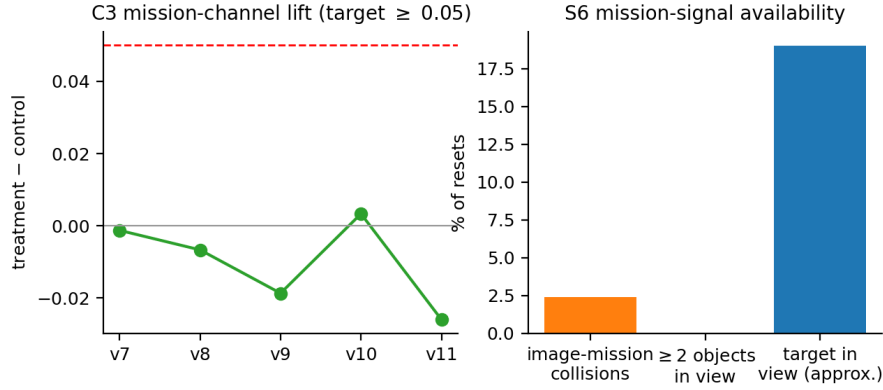


Figure 10: Mission-channel lift and signal availability.

- Deepak Dalal et al. Safe exploration in continuous action spaces. *arXiv:1801.08757*, 2018.
- Eduardo P. Denadai.  $\omega$ -certifiability: A theory of certified ignorance, 2026a. *Optimality.ai*.
- Eduardo P. Denadai. Knowing when not to act, 2026b. *Optimality.ai*.
- Eduardo P. Denadai. Fail-closed LLM, 2026c. *Optimality.ai*.
- Eduardo P. Denadai. Monotone guardrails: A learned veto that cannot loosen agent permissions, 2026d. *Optimality.ai*.
- Eduardo P. Denadai. Prompts are specifications, 2026e. *Optimality.ai*.
- Eduardo P. Denadai. Reinforcement learning from verifiable feedback, 2026f. *Optimality.ai*.
- Yonatan Geifman and Ronen El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30*, 2017.
- Kurt Gödel. Über formal unentscheidbare sätze der Principia Mathematica und verwandter systeme I. *Monatshefte für Mathematik und Physik*, 38(1):173–198, 1931.
- ISO, IEC, and RTCA. ISO 26262 (road vehicles); IEC 62304 (medical devices); RTCA DO-178c (airborne software), 2018. Functional safety standards.
- Saurav Kadavath et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.
- Yann LeCun. A path towards autonomous machine intelligence, 2022. *OpenReview (JEPA)*.
- H. G. Rice. Classes of recursively enumerable sets and their decision problems. *Transactions of the American Mathematical Society*, 74(2):358–366, 1953.
- Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. In *Proceedings of the 4th International Conference on Learning Representations (ICLR)*, 2016.
- Lui Sha. Using simplicity to control complexity. *IEEE Software*, 18(4):20–28, 2001.
- Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.

## Appendix A. Proof details

**A.1 Lemma 1.**  $V_{t+1} = V_t + \alpha(\epsilon_r + \gamma_{\text{step}}V_t - V_t)$ ; subtracting  $V^* = \epsilon_r / (1 - \gamma_{\text{step}})$  gives  $V_{t+1} - V^* = (1 - \alpha(1 - \gamma_{\text{step}}))(V_t - V^*)$ , a contraction for  $0 < \alpha < 2 / (1 - \gamma_{\text{step}})$ ; with  $\gamma_{\text{step}} = 0.99$ ,  $\alpha = 0.1$  the factor is 0.999 and  $V^* = -1.0$ . ■

**A.2 Theorem 2.** By the read path a pair reads its highest-precedence channel: the forced-halt action reads 0.7325, each visited sibling is capped to  $\leq 0.7325$ , unvisited siblings are excluded, and abstention reads  $\geq 0.75 - \epsilon$ . Since  $0.75 - \epsilon > 0.7325$  iff  $\epsilon < 0.0175$ , abstention is the strict argmax. ■

**A.3 Theorem 3.** The selection set is  $\{a : \text{guard}(s, a) = 1\} \cup \{\text{ABSTAIN}\}$  (I2, I4); guard is parameter-free and independent of  $\theta$  (I1); the density predicate is a function of the observation (I3). Hence  $\theta$  cannot add an action, and no gradient step changes the guard. ■

**A.4 Proposition 1.** With a refusing successor,  $T = \Omega_r + \gamma_{\text{step}}D(S', \text{ABSTAIN}) = -0.01 + 0.99 \cdot 0.75 = 0.7325 < 0.75$ ; if all other options at  $S$  read  $< 0.75$ , the argmax is abstention, and the argument iterates to predecessors — one-step-per-update backward propagation. ■

**A.5 Evidence gating.** The branch uses  $D(S', \text{ABSTAIN})$  only if some non-abstain pair of  $S'$  lies in anchors  $\cup$  ceilings  $\cup$  ledger — channels written only by forced halts and refutations (exact), so the propagated 0.7325 is a recorded lemma; otherwise it uses  $\max_{a \neq \text{ABSTAIN}} D(S', a)$ , which never creates a sink. ■

## Appendix B. Hyperparameters

| Symbol                 | Value               | Meaning                        |
|------------------------|---------------------|--------------------------------|
| $\alpha$               | 0.1                 | proposer learning rate         |
| $\gamma_{\text{step}}$ | 0.99                | bootstrap / proof-depth factor |
| $\epsilon_r$           | -0.01               | per-step PENDING cost          |
| $\Omega_r$ anchors     | +1.0 / +0.75 / -1.0 | certified / abstain / refuted  |
| $\tau_{\text{ood}}$    | -2000               | density floor threshold        |
| init                   | 1.0                 | training optimism              |
| deploy near-floor      | 0.75/0.74           | abstain / other (deploy view)  |
| bonus scale            | 0.25                | = init - 0.75 (derived)        |
| episodes / eval        | 1500 / 100          | per study                      |
| memory capacity        | 10000               | certified + boundary buffers   |

**Appendix C. Artifact provenance.** All reported numbers are reproducible from the run artifacts `cartographer/results/selfcontained_map.json` (v7) through `selfcontained_v11_map.json` and the matching `v*_caveats_run_*.log` files; figures are generated by `cartographer/docs/paper_model_free/make_figures.py`. The v7 baseline and each versioned build are preserved unmodified.