
When May a Robot Claim Completion? Evidence Certificates for Symbolic Tasks under Partial Observation

Moon Hwan Lee
RISU Institute, Washington, DC, USA

MOON@RISUINSTITUTE.ORG

In the 2026 BEHAVIOR Challenge, a policy receives RGB, depth, and proprioception, while the task evaluator can inspect symbolic simulator state that the policy cannot access. The evaluator can therefore mark a household task complete before the robot has evidence that supports the same claim. We ask when an embodied agent has enough evidence to say DONE.

We distinguish oracle completion from warranted completion. Let H_t be the robot's evidence and $\Gamma(H_t)$ the world histories compatible with it. A claim for task T is warranted only when its symbolic goal G_T is true in every compatible world. If compatible worlds disagree, the verdict remains UNKNOWN. A missing observation cannot establish a negative fact, and a final scene cannot establish facts that depend on hidden state or earlier events. This follows established treatments of partial observability and epistemic uncertainty (Bolander and Andersen, 2011; Cimatti et al., 2026).

We implemented a compact certificate calculus over the existing BDDL goal tree. Each atomic predicate contributes an obligation derived from its OmniGibson semantics. Spatial predicates use geometry or contact, open uses joint state, material predicates use particles, real requires transformed-entity existence, and cooked uses historical MaxTemperature while frozen uses current temperature. BDDL operators compose these obligations. Conjunction requires all children. Disjunction and existential conditions need one sufficient branch or witness. Universal conditions require grounded coverage. Negation requires evidence for the negative claim. Exact-count conditions need enough positive and negative evidence to close the count.

The compiler summarizes a minimal sufficient certificate as $B(T) = (A, F, Q, N, H, I)$, covering atomic obligations, verifier families, quantified grounding, negative obligations, temporal scope, and transformed-entity identity. The dimensions remain separate because they capture different evidence requirements.

We applied the calculus to all 100 tasks in the current 2026 Challenge set. Their goals contain 358 atomic predicate occurrences spanning 15 predicate types and seven verifier families. Sixty-five tasks require quantified grounding. Forty require at least one explicit negative certificate. Thirty-five require more than one verifier family. Seventeen require evidence beyond a final snapshot, and six contain genuine history dependence through cooked. Nine require evidence that transformed symbolic entities exist. These counts are computed directly from the goal structure, without a learned model or prompt-based classifier.

For 84 tasks, the vector is exact on the declared scope. Sixteen definitions contain scene-dependent wildcards, so their vectors are conservative lower bounds. The compiler follows the BDDL operators directly and uses no heuristic weighting. With sound atomic verifiers, a TRUE root certificate is sound by structural induction.

The vectors separate distinct sources of evidence burden. `assembling_gift_baskets` has one main verifier family but $Q = 32$ because many pairings require coverage. `canning_food` has $A = 10$, $F = 4$, $N = 4$, and $I = 2$ across material quantity, wrong-material exclusion, container state, and transformed identity. `cook_brussels_sprouts` reaches $B = (26, 4, 8, 2, 2, 8)$. Its eight transformed halves must exist, be cooked, and remain inside a closed oven. The cooking requirement carries history because the oracle reads stored `MaxTemperature`.

We next tested whether omitting these dimensions can change a completion verdict. A small witness generator built from the predicate semantics constructed five paired worlds while excluding privileged simulator state. Each pair fixes the available observation, changes one latent fact, and flips task truth. In `attach_a_camera_to_a_tripod`, visible geometry stays fixed while the attachment relation changes. In `cook_brussels_sprouts`, the same closed-oven observation is compatible with `MaxTemperature` of 80 °C or 60 °C, and only the first satisfies `cooked`. In `canning_food`, the same closed-cabinet observation is compatible with zero or one wrong-material particle, which flips a required negative contains condition. All five pairs produced opposite oracle completion values under the same permitted-observation abstraction.

The five witnesses cover structural state, history, transformed identity, negative material evidence, and a fill threshold at 0.20. The construction is deterministic and uses the predicate semantics directly, with no learned completion model.

These paired worlds give a direct limitation result. If two compatible worlds produce the same observation but have different task truth values, a Boolean rule using only that observation must answer both worlds alike. A sound agent therefore needs UNKNOWN until added evidence separates them. The logical goal supplies the proof structure, and the certificate layer identifies which observation, history, binding, or verification step is still missing before DONE.

McShane et al. (2025) ask when an interpretation is sufficient for action. We ask the post-action question. When is the evidence sufficient for DONE? That signal can trigger the next plan step or inform a supervisor. Assurance-case research likewise treats evidence sufficiency as an epistemic question (Graydon and Holloway, 2015). The 100-task audit and paired-world witnesses show that the required evidence can be derived mechanically from task semantics.

References

- BEHAVIOR Challenge. 2026. Evaluation and Rules. Stanford Vision and Learning Lab.
<https://behavior.stanford.edu/challenge/evaluation.html>
- Bolander, T., and Andersen, M. B. 2011. Epistemic planning for single- and multi-agent systems. *Journal of Applied Non-Classical Logics* 21(1): 9-34. <https://doi.org/10.3166/jancl.21.9-34>
- Cimatti, A., Grosen, T. M., Larsen, K. G., Tonetta, S., and Zimmermann, M. 2026. Exploiting Assumptions for Effective Monitoring of Real-Time Properties under Partial Observability. *Software and Systems Modeling*. <https://doi.org/10.1007/s10270-026-01394-6>
- Graydon, P. J., and Holloway, C. M. 2015. "Evidence" Under a Magnifying Glass: Thoughts on Safety Argument Epistemology. *Proceedings of the IET System Safety and Cyber Security Conference*, 6-11. <https://doi.org/10.1049/cp.2015.0277>
- Li, C., Zhang, R., Wong, J., et al. 2023. BEHAVIOR-1K: A Benchmark for Embodied AI with 1,000 Everyday Activities and Realistic Simulation. *Proceedings of the 6th Conference on Robot Learning*, PMLR 205: 80-93.
- McShane, M., Nirenburg, S., English, J., and Oruganti, S. 2025. Pursuing Actionable Perception Interpretation in Cognitive Robotic Systems. *Proceedings of the Twelfth Annual Conference on Advances in Cognitive Systems*, 369-388.