SAFE PROBING CERTIFIES WHAT A ROBOT CAN DO, NOT WHAT IT CANNOT

Anonymous authors
Paper under double-blind review

ABSTRACT

Safe physical probing can establish that a robot meets a task demand, while a safety envelope can prevent it from ruling out feasibility. We formalize this asymmetry using a capability signature $\theta^\star$, the supremum level an embodiment reaches along a task's binding capability axis in a scene, compared with a demand computed from scene geometry alone. Probes are restricted to $[0, A_{\text{safe}}]$, whereas $\theta^\star$ is defined over $[0, a_{\max}]$ with $A_{\text{safe}} < a_{\max}$. Within the stated response class, feasibility can be certified with finitely many probes whenever the true response at some safe grid amplitude exceeds c_{feas}, without a regularity constant. In the scalar-response observation model, for demands below the trivial upper bound, infeasibility has no uniformly valid test with non-trivial power: an infeasible response admits an indistinguishable extension beyond the envelope that changes the verdict. The upper identification endpoint remains at its trivial bound, even under smoothness or Lipschitzness with an unknown constant. SOMA-Probe combines a variance-adaptive, anytime-valid lower bound with four structural witnesses, each based on a stated, falsifiable, dimensionless order property. It selects probe amplitudes by risk-penalised information gain and returns CERTIFIED-FEASIBLE, CERTIFIED-INFEASIBLE, or UNCERTIFIED. A companion result bounds coverage for probing policies using the prescribed certification rules. The benchmark comprises four manipulation tasks and sixty embodiment perturbations, eighteen outside the parameter vector exposed to the identification baselines. On the 36 blind perturbations, certification coverage is 0.61, compared with 0.43 for the strongest safe-exploration baseline. The reported mis-certification rate is 0.011 at a nominal level of 0.05, excluding W_4^{cert} on D5a and the optional regularity branch. In a tier analysis across all sixty perturbations, the end-to-end advantage over ASID is concentrated in the eighteen perturbations outside its fitted parameter vector.

1 INTRODUCTION

Robots need to reassess capability before acting. Wear, bent links, and replacement grippers can change whether a robot can complete a task. Before attempting a lift or insertion, the robot must compare the scene's demand with what its current body can achieve. We describe that ability by a capability signature along the task's binding physical axis. The demand comes from scene geometry and physics independently of the body model; the signature must be inferred from physical observations.

Body identification can omit degradation that matters to the task. Latent adaptation summarizes interaction history within a training distribution (Kumar et al., 2021; Rakelly et al., 2019). Reactive correction observes joint behavior during degraded execution (Jo et al., 2026). Active identification estimates a chosen parameter vector (Memmel et al., 2024; Sobanbabu et al., 2025; Bohlinger & Peters, 2026), whose expressiveness determines which changes it can represent. In our benchmark, the fitted parameter vector omits load-dependent transmission efficiency and loading-history dependence. These distinctions motivate asking whether safe observations support the task decision when the degradation falls outside the fitted parameter vector.

Safe observations provide different evidence for feasibility and infeasibility. A safe response above the feasibility threshold can support a lower-bound certificate. An infeasibility certificate

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

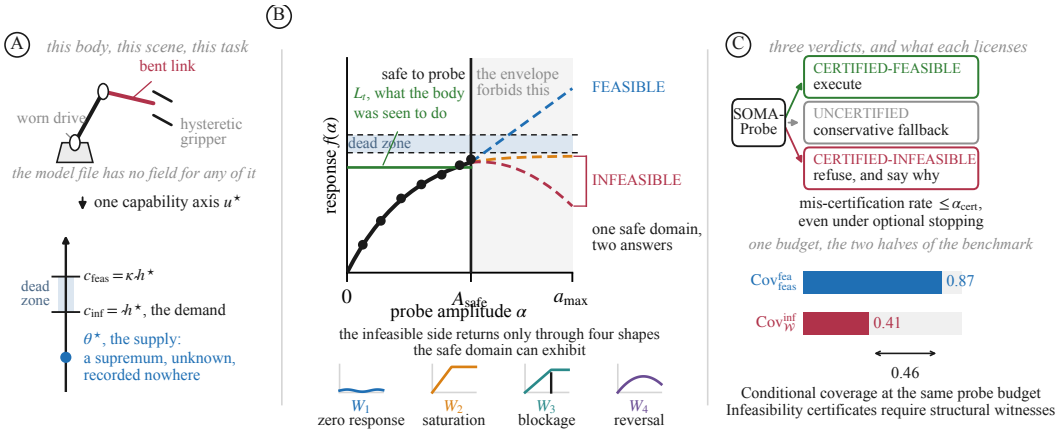


Figure 1: **Capability certification under a safety envelope.** (A) A task instance is represented by a capability signature $\theta^\star$ along one binding axis and a scene-derived demand $\hat{h}^\star$. The margin κ defines a dead zone in which neither certificate applies. (B) Probing is restricted to $A_{\text{safe}} < a_{\max}$. Compatible extensions beyond the envelope can differ in feasibility. The green curve shows the anytime-valid lower bound L_t used for feasibility certification without a regularity constant. Four structural witnesses provide upper bounds for infeasibility certification under their stated shape assumptions. (C) The three verdicts and their downstream actions. At the operating point, the same algorithm and budget certify 0.87 of feasible instances and 0.41 of infeasible instances.

must also rule out responses that reach the task demand beyond the safety envelope. Our scalar-observation analysis formalizes this obstruction when responses that agree on the safe interval induce the same observation law. Safe Bayesian optimization extrapolates using a known regularity bound (Sui et al., 2015; Berkenkamp et al., 2023), which can be difficult to obtain from prior knowledge (Fiedler et al., 2024). We study the certification rules available when no such bound is supplied.

SOMA-Probe uses this asymmetry to make three-way decisions. The method combines an anytime-valid lower confidence bound (Howard et al., 2021; Waudby-Smith & Ramdas, 2024) with four structural witnesses whose upper bounds require explicit shape assumptions. It adaptively selects safe probe amplitudes using risk-penalized information gain, certifies feasible or infeasible instances when the corresponding bound supports that verdict, and returns UNCERTIFIED otherwise. An unresolved verdict means that neither certification rule supports a decision. Limited safe amplitudes or an exhausted budget can leave a feasible task uncertified. Keeping these cases separate from certified infeasibility allows us to assess whether three-way decisions reduce the abandonment of achievable goals and improve downstream utility under different costs assigned to damage. The end-to-end success metric counts probing-induced safety aborts as failures; we also report certification coverage and probing risk.

Contributions.

- **Certification limits under a safety envelope.** With a positive safe-grid margin, feasibility is certifiable from finitely many probes at a prescribed confidence level. When the safety envelope excludes part of the amplitude domain, the common-kernel scalar-observation model admits no uniformly valid infeasibility test with non-trivial power over the stated response class for demands below the trivial bound (Section 3.1).
- **A three-way certification procedure with explicit guarantees.** We develop SOMA-Probe with false-certification control under the stated execution and shape assumptions, and derive a coverage bound for policies that retain its prescribed certificate family (Sections 3.2–3.4).
- **Capability-based decisions under model mismatch.** We evaluate four manipulation tasks and sixty body perturbations. Across all sixty perturbations, the observed end-to-end advantage over ASID is 0.207 on the eighteen perturbations outside its fitted parameter vector, versus 0.024 on D1–D4. Ablations using the same probe data examine the role of the capability target (Sections 4.2 and 4.5).

2 PROBLEM FORMULATION

Fix an embodiment e, scene s and task τ. A *capability axis* u is a pre-registered scalar physical direction with a declared unit and measurement protocol, and the *capability signature* of e in s along u is $h_{e,s}(u) := \sup\{\lambda \geq 0 : e \text{ in } s \textit{ achieves} \text{ level } \lambda \text{ along } u\}$. We use a declared quasi-static achievement criterion: level λ is achieved if some trajectory ends in a window of duration at least T_{hold} throughout which the reading on u is at least λ and $\|\dot{q}\|_\infty \leq v_{\text{tol}}$. This criterion excludes transient momentum from the capability measure. The supremum leads to the certification asymmetry studied below. Because h depends on the scene, probes are performed in the task scene. Every observable carries a registered (unit, origin, frame) triple and the task's one *certification channel* must match the demand's, so the map from reading to capability is the identity. Writing $u^\star$ for the binding axis, the quantity certified is $\theta^\star := h_{e,s}(u^\star)$; the demand $\hat{h}^\star$ comes from scene geometry alone, and with a pre-registered margin $\kappa \geq 1$ we fix two thresholds for the two verdicts, $c_{\text{feas}} := \kappa\hat{h}^\star$ and $c_{\text{inf}} := \hat{h}^\star$, with $\kappa = 1.15$ throughout.

A probe primitive a is a quasi-static application along $u^\star$ with scalar *amplitude* $\alpha := \text{amp}(a)$, expressed as an absolute target projection in the same origin as the certification reading, so $f(0) = 0$. Amplitudes lie on a fixed grid $\mathcal{A} \subset [0, a_{\max}]$ of $J = 24$ points, densified geometrically in the lower half, $a_{\max}$ being a nominal ceiling read coarsely off the model file. The grid is built once and never per safety level, which keeps bands and difference quotients comparable across levels. The *safety envelope* is the largest grid amplitude at which every quasi-static criterion holds on the nominal model,

$$A_{\text{safe}}(\zeta) := \max\{\alpha \in \mathcal{A} : \|\tau_j\| \leq \bar{\tau}(\zeta),\ m_{\text{tip}} \geq \underline{m}(\zeta),\ d_{\text{self}} \geq \underline{d}(\zeta),\ F_c \leq \bar{F}(\zeta)\}, \tag{1}$$

where joint torque, tip-over margin, self-collision distance and contact force each meet their own ζ-indexed limit, a fifth criterion on quasi-static velocity and acceleration being monitored online instead. The safety parameter $\zeta \in \{0.30, 0.50, 0.70, 0.85, 1.00\}$ runs from cautious to permissive, with 0.70 as the main level. We pre-register $A_{\text{safe}}(\zeta) < a_{\max}$ at every level so that each setting restricts the probe domain. Letting $f : [0, a_{\max}] \to \mathbb{R}$ be the response function, at step t we observe $y_t = f(\alpha_t) + \varepsilon_t$ with $\alpha_t \in \mathcal{A}$, $\alpha_t \leq A_{\text{safe}}$, $\mathbb{E}[\varepsilon_t \mid \mathcal{H}_{t-1}] = 0$ and observations bounded; the noise scale ς may be unknown and need not be constant in α, so the bands of §3 are variance-adaptive. By the achievement criterion, $\theta^\star = \sup_{[0,a_{\max}]} f$. We use the response class $\mathcal{F} := \{f : [0, a_{\max}] \to [0, B] \mid f \text{ Borel measurable},\ f(\alpha) \leq \alpha\}$, where $f(\alpha) \leq \alpha$ says only that a command is not exceeded; no continuity, monotonicity or Lipschitz condition is imposed. With $\theta(f) := \sup_{[0,a_{\max}]} f$, its largest value on $\mathcal{F}$ is the *trivial bound* $\widetilde{B} := \min(B, a_{\max})$. Writing $\mathcal{F}(\mathcal{D}_T) \subseteq \mathcal{F}$ for the functions consistent with the data after T probes, the *identification interval* is the image of that set under θ, with endpoints $L_T := \inf_{\mathcal{F}(\mathcal{D}_T)} \theta$ and $U_T := \sup_{\mathcal{F}(\mathcal{D}_T)} \theta$. Consistency means exact agreement at probed points when observations are noiseless and membership in the band of §3.2 when they are not, which puts $f^\star$ inside for all T at once with probability at least $1 - \alpha^{\text{CS}}$; Appendix B is precise about it. This is the partial-identification view (Manski, 2003; Tamer, 2010), which separates what the data imply about the two endpoints.

At each step the algorithm holds L_t and a set $\mathcal{W}$ of witness flags, each triggered W_k carrying an implied bound $\bar{\theta}_k$. The verdict is CERTIFIED-FEASIBLE if $L_t \geq c_{\text{feas}}$; CERTIFIED-INFEASIBLE if some triggered W_k has $\bar{\theta}_k < c_{\text{inf}}$; and UNCERTIFIED otherwise, including on budget exhaustion. Because the thresholds differ, $[c_{\text{inf}}, c_{\text{feas}})$ is a *structural dead zone* in which neither branch can fire, so an instance landing there comes back UNCERTIFIED however good the data are. The branches are mutually exclusive whenever the underlying assumptions hold, since a dual trigger would give $\bar{\theta}_k < c_{\text{inf}} \leq c_{\text{feas}} \leq L_t \leq \bar{\theta}_k$; any such event is audited and downgraded to UNCERTIFIED.

An *instance* is a tuple (e, s, τ) with its induced response $f^\star$ and demand $\hat{h}^\star$, and $\mathcal{P}$ is the distribution over instances the benchmark defines by sampling a task, a perturbation and a seed from fixed catalogues. All population quantities below are with respect to $\mathcal{P}$, whose infeasible mass $\pi_{\text{inf}} := \mathbb{P}_{\mathcal{P}}[\theta^\star < \hat{h}^\star]$ is a property of the benchmark and of no algorithm, and caps end-to-end success at $1 - \pi_{\text{inf}}$. At fixed budget and envelope we report $\text{Cov}(A_{\text{safe}}) := \mathbb{P}[\text{dec}_T \in \{\text{CERTIFIED-FEASIBLE}, \text{CERTIFIED-INFEASIBLE}\}]$, decomposed into Cov_{feas} and $\text{Cov}_{\mathcal{W}}$; the probing damage risk $\text{Risk}(A_{\text{safe}}) := \mathbb{P}[\exists t \leq T : a_t \text{ trips any safety criterion}]$; and the mis-certification rate MCR, the probability of CERTIFIED-FEASIBLE when $\theta^\star < c_{\text{feas}}$ or CERTIFIED-INFEASIBLE when $\theta^\star \geq c_{\text{inf}}$. An episode aborted for safety counts as UNCERTIFIED in the Cov

denominator, in the Risk numerator and as zero end-to-end, so Cov need not be monotone in A_{safe}. The target is $\text{MCR} \leq \alpha_{\text{cert}} = 0.05$, since a certificate is a statistical object with a declared error rate (Ren et al., 2023).

The safe-grid margin determines the certification scope. With $\theta^\star_{\text{safe}} := \max_{j:\alpha_j \leq A_{\text{safe}}} f(\alpha_j)$ the largest response on the safe grid, the *certifiable gap* is $\Delta_+ := \theta^\star_{\text{safe}} - c_{\text{feas}}$. For a positive gap, the sample cost scales as $1/\Delta_+^2$. We exclude feasible instances with gaps in $(0, 1.0\varsigma)$ from the stated scope. Instances with $\Delta_+ \leq 0$ remain in scope: they include infeasible instances, instances in the dead zone, and globally feasible instances whose safe-grid responses do not reach the feasibility threshold. These instances may receive an UNCERTIFIED verdict.

3 SOMA-PROBE

3.1 WHAT SAFE OBSERVATIONS CAN CERTIFY

Safe observations can leave global capability unidentified. Responses with different global capabilities can agree throughout the safe interval, even under continuity, smoothness, Lipschitzness or monotonicity. Lemma B.1 constructs the extensions used below.

Proposition 3.1 (Certification asymmetry)**.** *Let all probes satisfy $\alpha_t \leq A_{\text{safe}} \leq a_{\max}$, and let the data-consistent response class be non-empty.* (a) The two endpoints*: $U_T = \widetilde{B}$ for every T, and this survives revealing $f^\star$ exactly on all of $[0, A_{\text{safe}}]$, so $U_\infty = \widetilde{B}$ as well. The lower endpoint does respond to data: noiselessly it is $L_T = \max_{t \leq T} y_t$, non-decreasing in T and equal to $\theta^\star_{\text{safe}}$ once the safe grid is exhausted, while under the bands of §3.2 it is $L_t = \max_j(\hat{\mu}_j(t) - r_j(t))^+$ with $\mathbb{P}[\forall t \geq 1 : L_t \leq \theta^\star] \geq 1 - \alpha^{\text{CS}}$.* (b) Consequences*:* (b.1) *if $\theta^\star_{\text{safe}} > c_{\text{feas}}$, then $\{\theta^\star \geq c_{\text{feas}}\}$ is certifiable in finitely many safe probes with probability at least $1 - \alpha^{\text{CS}}$;* (b.2) *consider the scalar-observation experiment in which, conditional on the history and next amplitude, the observation kernel is common to all responses with the same value $f(\alpha_t)$; initial information is also common to responses agreeing on the safe interval. Let $\mathcal{A}$ be confined to $\alpha_t \leq A_{\text{safe}}$ and have false-*CERTIFIED-INFEASIBLE *probability $\sup\{\mathbb{P}_f[\mathcal{A}$ outputs* CERTIFIED-INFEASIBLE*$] : f \in \mathcal{F}, \theta(f) \geq c_{\text{inf}}\} \leq \alpha_{\text{cert}}$ for some $c_{\text{inf}} < \widetilde{B}$; then for every $f_0 \in \mathcal{F}$, $\mathbb{P}_{f_0}[\mathcal{A}$ outputs* CERTIFIED-INFEASIBLE*$] \leq \alpha_{\text{cert}}$, so no uniformly valid test of the infeasible side has non-trivial power;* (b.3) *if $c_{\text{feas}} > A_{\text{safe}}$ then $\theta^\star_{\text{safe}} \leq A_{\text{safe}} < c_{\text{feas}}$ by $f \leq \alpha$, so on the band-coverage event the* CERTIFIED-FEASIBLE *branch cannot fire; only a witness can certify the instance as* CERTIFIED-INFEASIBLE*.*

The testing conclusion depends on the observation law. Exact knowledge of $f^\star$ on $[0, A_{\text{safe}}]$ leaves the upper endpoint unchanged. Part (b.2) also requires identical complete scalar-observation laws; conditional mean-zero noise alone does not suffice. Auxiliary measurements need their own observation model and are outside (b.2). Corollary B.4 extends the upper-endpoint and testing conclusions to continuous, monotone and Lipschitz functions with unspecified constants. Under unimodality, an observed decline constrains later responses; W_4 uses this information. SafeOpt-type guarantees (Sui et al., 2015; Berkenkamp et al., 2023) require a known regularity constant. In-envelope data can lower-bound slopes but cannot establish the outside-envelope upper bound needed for extrapolation.

3.2 CONSTRUCTING FEASIBLE AND INFEASIBLE CERTIFICATES

A simultaneous confidence band supplies the feasible certificate. For each grid point with $n_j(t)$ repeats, we maintain the empirical variance $\hat{V}_j(t) = \sum_{i \leq t:\alpha_i = \alpha_j}(y_i - \hat{\mu}_j(i-1))^2$ and radius

$$r_j(t) = \frac{1}{n_j(t)}\,\mathcal{B}_{\text{eB}}\left(\hat{V}_j(t);\, \frac{\alpha^{\text{CS}}}{2JN_{\text{ch}}},\, \rho_{\text{eB}}\right), \qquad L_t = \max_j\big(\hat{\mu}_j(t) - r_j(t)\big)^+, \tag{2}$$

with $\mathcal{B}_{\text{eB}}$ an empirical-Bernstein mixture boundary (Howard et al., 2021; Waudby-Smith & Ramdas, 2024), adaptive to an unknown, amplitude-dependent noise scale ς. A boundary using a fixed assumed scale can be conservative or invalid when that scale differs from the true noise level. The lower edge $\underline{\mu}_j := \hat{\mu}_j - r_j$ gives L_t; the upper edge $\bar{\mu}_j := \hat{\mu}_j + r_j$ enters the witnesses. The allocation $\alpha^{\text{CS}}/(2JN_{\text{ch}})$ to each edge covers all $J = 24$ grid points, including those outside the current envelope, independently of ζ. The feasible certificate fires when $L_t \geq c_{\text{feas}}$.

Infeasibility requires an upper bound on every admissible extension. The unrestricted scalar-observation class of Proposition 3.1(b.2) provides no informative infeasibility test. Structural witnesses add explicit shape assumptions with observable in-envelope failure modes. Their tolerances η_1, η_2, η_4 and width $w_{\min}$ enter the bounds and are fixed before evaluation in units of ς, $\hat{h}^\star$ and $a_{\max}$ (Appendix D). W_1 detects a noise-floor response despite verified actuator effort; (S1) excludes delayed breakout. W_2 detects a plateau at least $w_{\min}$ wide at the torque limit under irreversible saturation (S2). Under the static-scene assumption (S3), W_3 detects persistent contact at a repeatable position that does not advance as push force increases. W_4 detects a reversal $\underline{\mu}_{j_1} - \bar{\mu}_{j_2} > \eta_4$ that persists to the envelope boundary and uses unimodality (S4). Table 3 gives the full triggers and implied bounds $\bar{\theta}_k$. A witness certifies CERTIFIED-INFEASIBLE only when $\bar{\theta}_k < c_{\inf}$.

Witness states distinguish missing tests from negative evidence. Each witness reports NOT-EVALUATED, NOT-TRIGGERED or TRIGGERED. Conflating the first two converts a budget shortfall into evidence of feasibility. The hand-enumerated set can miss witnesses, lowering $\mathrm{Cov}_{\mathcal{W}}$ without creating a false certificate. W_4 has a certifying form W_4^{cert}, requiring decline to the envelope boundary, and an observational form W_4^{obs}, which implies no bound.

D5a deliberately violates unimodality. Its transmission efficiency falls and partly recovers, creating a second response peak. The later rise usually blocks W_4^{cert} despite frequent W_4^{obs} triggers. Its W_4^{cert} mis-certification rate exceeds α_{cert} and is reported separately to test the assumption-failure term in Proposition 3.2.

3.3 ADAPTIVE PROBING AND THREE-WAY DECISIONS

The loop stops at a certificate, a safety violation or the budget. An episode snapshots the state, installs the weld anchor and virtual wall, and uses $n_0 = 8$ calibration steps at $0.1A_{\mathrm{safe}}$ for initialisation. Equation (2) reads observations at their actual safe-grid amplitudes; off-grid calibration readings never enter L_t. For at most $T_{\max} = 96$ steps, the loop returns CERTIFIED-FEASIBLE if $L \geq c_{\mathrm{feas}}$ or CERTIFIED-INFEASIBLE if a triggered witness gives $\bar{\theta}_k < c_{\inf}$. Otherwise it probes at $\alpha_t = \arg\max_{\alpha \in \mathcal{A} \cap [0, A_{\mathrm{safe}}]} [\mathrm{IG}(\alpha) - \lambda \mathrm{pen}(\alpha)]$, aborts on an online safety violation, and updates the band, witnesses and posterior. At exhaustion it repeats both checks, returning UNCERTIFIED with $[L_T, \widetilde{B}]$ if neither fires. Finally, it releases isolation, checks zero task progress and the anchor residual, and restores the snapshot (Algorithm 1).

The posterior selects probes; the band and witnesses certify. The GP posterior and risk penalty determine the next amplitude; Proposition 3.2 permits arbitrary action selection. Appendix D.2 specifies the information-gain model and risk penalty. Probes remain in the task scene, with the weld anchor and virtual wall holding task progress at zero. Each primitive restores the integration state.

3.4 CERTIFICATION VALIDITY AND COVERAGE

Simultaneous bounds control statistical certification errors. The budget $\alpha^{\mathrm{CS}} = 0.04$ covers the band and its functionals L_t, W_1, W_2 and W_4; $\alpha^{(3)} = 0.01$ covers W_3's contact-position band and repeatability test. Their union bound gives $\alpha_{\mathrm{cert}} = \alpha^{\mathrm{CS}} + \alpha^{(3)}$. The optional regularity branch has a separate $\alpha^{(R)} = 0.05$ budget (Appendix D.1).

Proposition 3.2 (Anytime-valid certification)**.** *For Algorithm 1 run with probes in $\mathcal{A} \cap [0, A_{\mathrm{safe}}]$ and stopped at an arbitrary stopping time, and for* every *$f^\star \in \mathcal{F}$, the probability that a false certification is ever issued is at most $\alpha_{\mathrm{cert}} + \mathbf{1}[\exists k : W_k$ triggerable and (S_k) fails on $f^\star]$. In particular, if the shape assumptions in force hold, the false-certification probability is at most $\alpha_{\mathrm{cert}} = 0.05$ under optional stopping, and marginalising over the part of $\mathcal{P}$ on which they hold gives $\mathrm{MCR} \leq \alpha_{\mathrm{cert}}$ there.*

The indicator separates shape-assumption failures from statistical errors. Each witness specifies the order property needed for its upper bound and an audit for observable in-envelope violations. Passing the audit does not establish the property outside the probed region. CERTIFIED-FEASIBLE also relies on execution conditions (A1) and (A7)–(A10). Violating four of these can produce unsafe decisions: three can inflate L_t, and one can lower its threshold. The fallback ordering prioritises the route that does not use L_t. Appendix C lists all conditions and audits.

Coverage depends on the prescribed certificate family. Consider policies using the CF and W_1–W_4 rules. Fix their grid, budget, calibration, safety checks and auxiliary measurement protocol. Let $\mathcal{E}_{\mathrm{feas}} := \{\theta^\star_{\mathrm{safe}} \geq c_{\mathrm{feas}}\}$, with CF reading only the safe-grid lower bounds. Let $\mathcal{E}_{\mathcal{W}}$ contain the instances for which some allowed history, compatible with that complete instance and with correct confidence bounds, makes the prescribed CI gate fire. The events must be measurable; a verified measurable superset may be used for $\mathcal{E}_{\mathcal{W}}$. Proposition B.5 gives $\mathrm{Cov}(A_{\mathrm{safe}}) \leq \min\{1, \mathbb{P}[\mathcal{E}_{\mathrm{feas}}] + \mathbb{P}[\mathcal{E}_{\mathcal{W}}] + 2\alpha_{\mathrm{cert}}\}$ when each branch's simultaneous confidence event fails with probability at most α_{cert} on each instance. The bound applies across selection and stopping policies that retain these rules. It requires neither an exact identification endpoint nor attainment of its supremum. Its event probabilities depend on the fixed protocol and must be established for that protocol before using the bound numerically.

The feasible-side probe budget depends on the safe-grid margin. For $\Delta_+ = \theta^\star_{\mathrm{safe}} - c_{\mathrm{feas}} > 0$, the requirement scales as $1/\Delta_+^2$, with a time-uniform iterated-logarithm factor (Proposition B.11). Appendix D.1 relates this bound to the $T_{\max} = 96$ quota and its $\Delta_+ = 1.0\varsigma$ scope boundary. Proposition 3.1(b.2) rules out a corresponding infeasibility guarantee over its stated scalar-observation class.

4 Experiments

4.1 Experimental setup

The benchmark separates represented and omitted body changes. All experiments use a Franka Emika Panda in MuJoCo (Todorov et al., 2012; Zakka et al., 2022) at a pinned engine version. Four tasks cover reach (high reach), clearance (narrow insertion), and load (heavy transport and heavy push). Each spans 11 bottleneck levels $\times$ 3 configurations, varying demand and supply across feasible, infeasible, and dead-zone cells. Of sixty body perturbations, D1–D4 contain 42 changes within the fitted parameter vector and randomisation space: joint damage, torque decay, payload and centre-of-mass shift, and morphology. D5 contains eighteen changes outside that vector: load-dependent torque transfer that falls and partly recovers (D5a), Bouc–Wen hysteresis (Bouc, 1968; Wen, 1976) on an added degree of freedom (D5b), a permanent 3–8 mm link bend (D5c), and anisotropic friction (D5d). D5c is representable by kinematic parameters and has a separate kinematic-calibration control (Table 15). The split is 24 development and 36 blind perturbations. Thresholds, budgets, and analyses were fixed before unsealing the blind split (Appendix E). Under the pre-specified interval-selection criteria, the retained benchmark has $\pi_{\mathrm{inf}} = 0.34$. The resulting quasi-static reference success ceiling is 0.66 for all methods, including the oracle; execution can relax (A1) (Appendix F).

Comparisons match probing opportunities. Ten baselines cover passive adaptation, active parameter identification, behaviour search, safe optimisation, and oracle or nominal policies (Table 7). Active methods share the probe primitives. A two-pass sweep matches B1–B6 to SOMA-Probe's per-episode probe allowance; B1–B3 may use free interaction. B10, SafeOpt-MC (Berkenkamp et al., 2023), shares the $J = 24$ amplitude grid and produces three-way verdicts using its own budget. Total steps are unrestricted because conservative routes add task-execution steps. Appendix E gives the baseline catalogue and configurations.

Metrics separate certification from task completion. Alongside Cov, Risk, and MCR (§2), we report $\mathrm{Cov}^{\mathrm{inf}}_{\mathcal{W}} = \mathbb{P}[\textsc{Certified-Infeasible} \mid \theta^\star < \hat{h}^\star]$ and $\mathrm{Cov}^{\mathrm{fea}}_{\mathrm{feas}} = \mathbb{P}[\textsc{Certified-Feasible} \mid \theta^\star \geq \kappa\hat{h}^\star]$. Binary end-to-end success measures original-task completion. Correct Certified-Infeasible abstentions, safety aborts, and successful conservative fallbacks receive zero, accounting for roughly two-fifths of our episodes. Paired utility ΔU compares three-way decisions with a policy merging Uncertified into Certified-Infeasible, assigning task value 1 and damage cost $\rho \in \{3, 10, 30\}$. Methods share seeds for each (perturbation, task, episode) triple. Appendix E specifies the mixed model, comparison families, and cluster bootstrap.

4.2 Task performance under body degradation

We test whether safe certification improves task performance and which body changes account for the advantage.

Table 1: Primary endpoints on the 36 blind perturbations × 4 tasks at $\zeta = 0.70$. End-to-end success includes probe risk and scores abstentions, aborts and successful conservative routes as 0. T_{probe} is the mean probe-step count, with $T_{\max} = 96$. B1–B6 receive allowances matched to SOMA-Probe; B10 uses its own budget. Total steps include the unmatched task segment. ΔU is the paired utility difference against the two-way policy at $\rho = 10$; it and Cov are undefined for methods without three-way verdicts. The p column gives raw paired-test values against SOMA-Probe; intervals and Holm-adjusted values are in Table 6. Bold marks the best non-oracle value in each arrow's direction. Matched probe steps are not ranked. The lowest abort rate, B3's 0.018, appears in the full table (Table 5); no displayed abort value is bolded. SOMA-Probe is shaded.

Method	**E2E** ↑	**Task-only** ↑	T_{probe}	**Total**	**Abort** ↓	Cov ↑	ΔU ↑	p
B1 RMA	0.331	0.372	61.4	243.4	0.081	n/a	n/a	0.001
B5 ASID	0.344	0.412	61.4	247.0	0.149	n/a	n/a	0.003
B8 Oracle signature	0.561	0.566	0.0	171.2	0.003	n/a	n/a	<0.001
B9 No adaptation	0.196	0.199	0.0	174.6	0.031	n/a	n/a	<0.001
B10 SafeOpt-MC	0.336	0.398	88.7	271.7	0.052	0.427	0.038	<0.001
SOMA-Probe (ours)	**0.446**	**0.502**	61.4	276.3	0.068	**0.610**	**0.126**	ref.

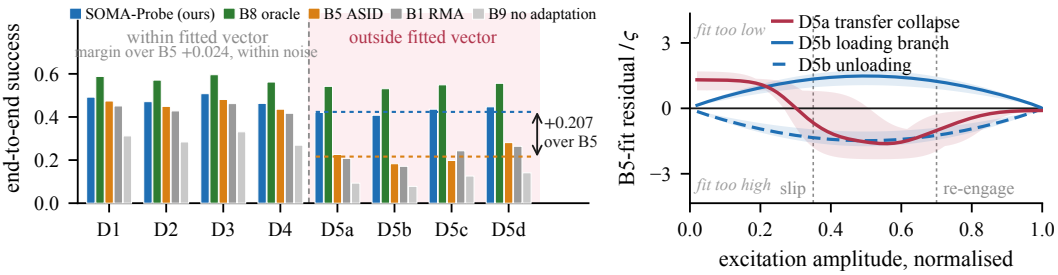


Figure 2: End-to-end success by perturbation tier and residuals of ASID's fitted model. Left: all 60 perturbations, matching the population used for the tier decomposition and interaction test. Right: residuals change sign across the slip load on D5a and between loading and unloading on D5b. Lines are subclass means; bands show the empirical spread over six perturbations per subclass. Subclass comparisons are descriptive under the pre-registration; significance is tested for D5 against D1 through D4. The sign pattern persists across the observed spread, while residual magnitudes vary.

Task success improves at matched probe budgets. At a mean probe budget of 61.4 steps, below $T_{\max} = 96$, SOMA-Probe improves end-to-end success over RMA by 0.115 and ASID by 0.102; both exceed the 0.091 minimum detectable effect (Table 1). The task-only versus end-to-end gap is 0.056 for SOMA-Probe and 0.068 for ASID. Their abort rates are 0.068 and 0.149, respectively. SOMA-Probe uses the most total steps because of conservative routes. The ground-truth-signature oracle reaches 0.561.

The advantage is concentrated outside ASID's fitted parameter vector. The margin is 0.024 on D1–D4, below the 0.091 minimum detectable effect, and 0.207 on D5, above the pre-specified 0.15 threshold (Figure 2). The interaction test compares these margins (G3). This analysis uses all 60 perturbations, including eighteen D5 clusters; the primary endpoint uses the blind 36. Perturbations are the independent units, so retaining all D5 clusters supports interaction estimation.

Residuals reveal omitted load and history dependence. ASID's constant-parameter fit leaves positive residuals below the slip load and negative residuals above it on D5a; residuals reverse between loading and unloading on D5b. The registered sign pattern occurs in 87% of D5a and D5b instances. The capability target remains defined through physical response when the fitted parameter model omits these dependencies.

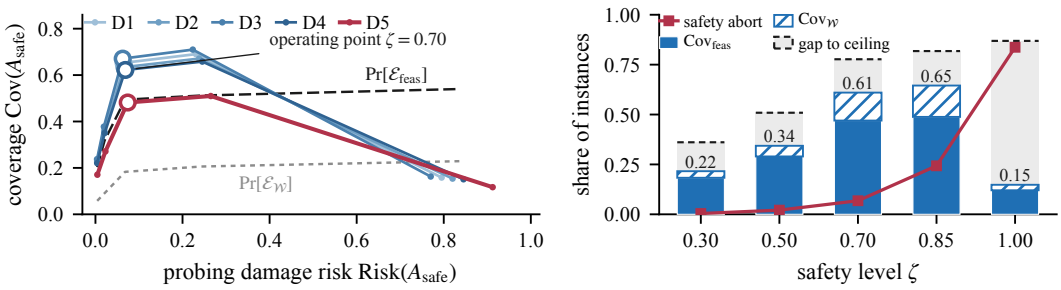


Figure 3: Coverage across safety levels. Left: Cov against probe damage risk, with one curve per tier and open markers at $\zeta = 0.70$. Compare tiers at matched safety levels. The two dashed curves show reported reference rates $p^{\text{ref}}_{\text{feas}}$ and $p^{\text{ref}}_{\mathcal{W}}$. Right: coverage decomposed into feasible certificates and structural witnesses, with the abort rate overlaid. The legend "gap to ceiling" denotes $R - \text{Cov}$, where $R = p^{\text{ref}}_{\text{feas}} + p^{\text{ref}}_{\mathcal{W}} + 2\alpha_{\text{cert}}$ is a descriptive reference; its identification with the bound in Proposition B.5 has not been established.

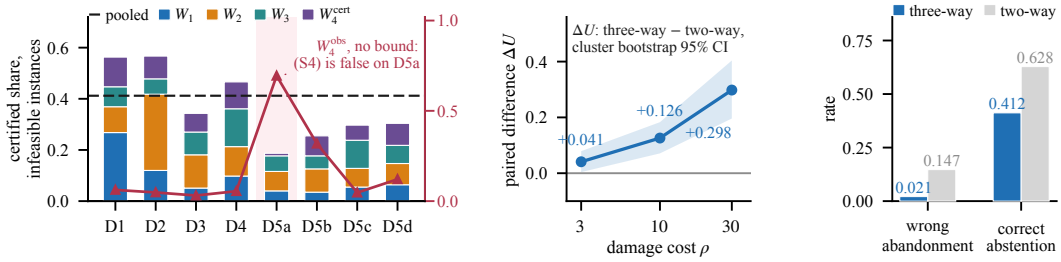


Figure 4: Witness coverage and the utility of three-way decisions. Left: each witness's contribution to the tier's conditional infeasibility coverage $\text{Cov}^{\text{inf}}_{\mathcal{W}}$; the dashed line is the pooled value. The shaded D5a column contains instances that all violate (S4). W_4^{obs} (right axis) triggers frequently there, while W_4^{cert} contributes 0.010 to coverage. Middle and right: the paired utility difference against a two-way policy merging UNCERTIFIED with CERTIFIED-INFEASIBLE, and the corresponding correct-abstention and wrong-abandonment rates.

4.3 CERTIFICATION COVERAGE AND PROBING RISK

We vary the safety envelope to examine the trade-off between available evidence and probe risk.

The most permissive envelope incurs frequent safety aborts. Coverage rises from 0.22 at $\zeta = 0.30$ to 0.65 at $\zeta = 0.85$. At $\zeta = 1.00$, aborts reach 0.84 and coverage falls to 0.15 (Figure 3); aborted episodes remain in the coverage denominator as UNCERTIFIED. Over $\zeta \in [0.50, 0.85]$, the descriptive reference R exceeds coverage by 0.166–0.173. Applying Proposition B.5 numerically requires verifying that the reported rates equal or upper-bound its two event probabilities; aggregate rates alone do not establish this correspondence.

Feasible instances are certified more often at the operating point. The same algorithm and budget yield $\text{Cov}^{\text{fea}}_{\text{feas}} = 0.87$ and $\text{Cov}^{\text{inf}}_{\mathcal{W}} = 0.41$, conditioned respectively on $\theta^\star \geq \kappa\hat{h}^\star$ and $\theta^\star < \hat{h}^\star$. The difference, 0.46, exceeds the pre-specified asymmetry threshold of 0.25 (G1). These rates are consistent with Proposition 3.1; instance margins and witness coverage also affect their difference.

4.4 STRUCTURAL WITNESSES AND THREE-WAY DECISIONS

We examine how witnesses supply infeasibility evidence and how preserving unresolved decisions affects behaviour.

Witness validity depends on the stated shape assumptions. Witnesses contribute 0.14 to total coverage of 0.61. On D5a, every instance violates (S4); the separately reported W_4^{cert} mis-certification

Table 2: Ablations on all 60 perturbations, grouped by whether they lie within the fitted parameter vector. This population differs from the blind split in Table 1. The parameter-target ablation (AB3) uses the same probe data as the signature target. Direct measurement beyond the envelope (AB6) is a diagnostic and is excluded from the ranking. The risk-weight row (AB7) reports $\lambda = 0$; Table 11 gives the full sweep. Bold marks the best value among methods restricted to the safety envelope; SOMA-Probe is shaded.

		D1–D4 (within fitted vector)			D5 (outside fitted vector)		
	Configuration	E2E↑	Cov↑	MCR↓	E2E↑	Cov↑	MCR↓
AB0	Full method	**0.483**	0.634	**0.010**	**0.423**	0.552	0.014
AB1	Probe direction: random / uniform	0.441	0.567	0.012	0.371	0.468	0.017
AB2	Stopping: fixed $\{24, 48, 96\}$ / precision	0.478	0.618	0.011	0.416	0.531	0.016
AB3	Target: parameters / latent / signature	0.475	0.621	0.011	0.243	0.318	0.048
AB4	Witnesses ablated singly (worst case)	0.482	0.589	**0.010**	0.421	0.487	**0.013**
AB5	Multi-axis: top-2 / top-3	0.479	0.628	0.012	0.419	0.544	0.016
AB6	Direct measurement past the envelope	0.541	0.812	0.008	0.498	0.744	0.011
AB7	Risk weight $\lambda \in \{0, 0.5, 2, 8, 32\}$ (λ=0)	0.402	**0.658**	0.012	0.338	**0.571**	0.018
AB8	Conformal grouping: marginal / Mondrian	0.480	0.629	0.019	0.421	0.547	0.041
AB9	L misspecification scan ($L = 4\hat{L}_{\mathrm{emp}}$)	0.481	0.631	0.011	0.422	0.549	0.014

rate is 0.104 (Table 4). Pooled MCR excludes W_4^{cert} on D5a and the optional regularity branch. The D5a result illustrates the assumption-failure term in Proposition 3.2.

Keeping UNCERTIFIED distinct reduces wrong abandonment. Merging UNCERTIFIED with CERTIFIED-INFEASIBLE increases correct abstention from 0.412 to 0.628 and wrong abandonment from 0.021 to 0.147. Under the pre-registered utility matrix, the paired advantage of the three-way policy grows from $+0.041$ at $\rho = 3$ to $+0.298$ at $\rho = 30$ (Figure 4).

4.5 ABLATIONS OF THE CERTIFICATION COMPONENTS

The estimation target matters most on D5. With probe data fixed, replacing the capability signature with a parameter target (AB3) changes end-to-end success from 0.483 to 0.475 on D1–D4 and from 0.423 to 0.243 on D5. This comparison connects the D5 advantage to the target inferred from the observations.

Family-specific calibration lowers the reported error rate. Replacing it with marginal calibration (AB8) raises D5 MCR from 0.014 to 0.041, while success changes from 0.423 to 0.421. The Lipschitz scan (AB9) uses $L \in \{0.25, 0.5, 1, 2, 4\} \times \hat{L}_{\mathrm{emp}}$. The main branch does not consume L; at $4\hat{L}_{\mathrm{emp}}$, reported success is 0.481 on D1–D4 and 0.422 on D5. Table 2 reports the remaining ablations.

5 CONCLUSION

Safe probing yields asymmetric evidence about robot capability. A safe-grid response above c_{feas} can support a feasibility certificate. Under the common scalar-observation law, an infeasible response in the general class remains indistinguishable from a feasible extension outside the envelope when the demand is below the trivial bound. The full-procedure coverage bound is specific to the prescribed certification rules.

SOMA-Probe combines an anytime-valid lower bound with four structural witnesses and returns UNCERTIFIED when neither certificate is supported. Its infeasibility certificates depend on explicit shape assumptions; D5a tests their failure by violating unimodality.

Our evaluation covers simulated tasks with one binding capability axis. The hand-specified witnesses limit which infeasible instances can be certified. Jointly reporting coverage and end-to-end success, with probe damage charged to the latter, measures both the information gained and its cost.

AI USE STATEMENT

AI assistants supported manuscript organization, reference summarization, inspection of supplied code and experiment records, mathematical analysis and proof revision, bilingual drafting and editing, and LaTeX preparation. The authors are responsible for the final technical content and references.

REFERENCES

Gustavo Belforte, Basilio Bona, and Vito Cerone. Parameter estimation algorithms for a set-membership description of uncertainty. *Automatica*, 26(5):887–898, 1990. doi: 10.1016/0005-1098(90)90005-3.

Felix Berkenkamp, Andreas Krause, and Angela P. Schoellig. Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics. *Machine Learning*, 112(10):3713–3747, 2023. doi: 10.1007/s10994-021-06019-1.

Nico Bohlinger and Jan Peters. Active embodiment identification with reinforcement learning for legged robots. *arXiv preprint arXiv:2605.08020*, 2026. URL https://arxiv.org/abs/2605.08020.

R. Bouc. Forced vibrations of mechanical systems with hysteresis. In *Proceedings of the Fourth Conference on Nonlinear Oscillations*, Prague, 1968. Academia. Conference held in Prague, 5–9 September 1967; proceedings published in 1968.

Shouyuan Chen, Tian Lin, Irwin King, Michael R. Lyu, and Wei Chen. Combinatorial pure exploration of multi-armed bandits. In *Advances in Neural Information Processing Systems*, volume 27, pp. 379–387. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/d3ea0f3316d2da934d79b8b344eafee4-Paper.pdf.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 844–853. PMLR, 2017. URL https://proceedings.mlr.press/v70/chowdhury17a.html.

Antoine Cully, Jeff Clune, Danesh Tarapore, and Jean-Baptiste Mouret. Robots that can adapt like animals. *Nature*, 521(7553):503–507, 2015. doi: 10.1038/nature14422.

Christian Fiedler, Johanna Menn, Lukas Kreisköther, and Sebastian Trimpe. On safety in safe Bayesian optimization. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=tgFHZMsl1N.

Michel Gevers. Identification for control: From the early achievements to the revival of experiment design. *European Journal of Control*, 11(4-5):335–352, 2005. doi: 10.3166/ejc.11.335-352.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. doi: 10.1214/18-PS321.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

Jonas Hübotter, Bhavya Sukhija, Lenart Treven, Yarden As, and Andreas Krause. Transductive active learning with application to safe Bayesian optimization. In *ICML 2024 Workshop: Aligning Reinforcement Learning Experimentalists and Theorists*, 2024a. URL https://openreview.net/forum?id=aaC86cHArL.

Jonas Hübotter, Bhavya Sukhija, Lenart Treven, Yarden As, and Andreas Krause. Transductive active learning: Theory and applications. In *Advances in Neural Information Processing Systems*, volume 37, pp. 124686–124755, 2024b. doi: 10.52202/079017-3961.

Minsoo Jo, Taeju Kwon, Junha Chun, Youngjoon Jeong, and Taesup Kim. Uncovering vulnerability of vision-language-action models under joint-level physical faults. *arXiv preprint arXiv:2606.10501*, 2026. URL https://arxiv.org/abs/2606.10501.

Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. RMA: Rapid motor adaptation for legged robots. In *Proceedings of Robotics: Science and Systems*, Virtual, 2021. doi: 10.15607/RSS.2021.XVII.011.

Andrea Locatelli, Maurilio Gutzeit, and Alexandra Carpentier. An optimal algorithm for the thresholding bandit problem. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1690–1698, New York, New York, USA, 2016. PMLR. URL https://proceedings.mlr.press/v48/locatelli16.html.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer Series in Statistics. Springer, New York, NY, 2003. ISBN 978-0-387-00454-9. doi: 10.1007/b97478.

Marius Memmel, Andrew Wagenmaker, Chuning Zhu, Dieter Fox, and Abhishek Gupta. ASID: Active exploration for system identification in robotic manipulation. In *International Conference on Learning Representations*, volume 2024, pp. 39979–39997, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/ae8afec5dabb07091d8f85128250f2ff-Abstract-Conference.html.

Mario Milanese and Carlo Novara. Set membership identification of nonlinear systems. *Automatica*, 40(6):957–975, 2004. doi: 10.1016/j.automatica.2004.02.002.

Tom M. Mitchell. Generalization as search. *Artificial Intelligence*, 18(2):203–226, 1982. doi: 10.1016/0004-3702(82)90040-6.

Joshua K. Pickard, Vincent Padois, Milan Hladík, and David Daney. Efficient set-based approaches for the reliable computation of robot capabilities. *arXiv preprint arXiv:2104.00279*, 2021. URL https://arxiv.org/abs/2104.00279.

Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5331–5340. PMLR, 2019. URL https://proceedings.mlr.press/v97/rakelly19a.html.

Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 661–682. PMLR, 2023. URL https://proceedings.mlr.press/v229/ren23a.html.

Nikhil Sobanbabu, Guanqi He, Tairan He, Yuxiang Yang, and Guanya Shi. Sampling-based system identification with active exploration for legged Sim2Real learning. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 578–598. PMLR, 2025. URL https://proceedings.mlr.press/v305/sobanbabu25a.html.

Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias W. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 1015–1022. Omnipress, 2010. URL https://icml.cc/Conferences/2010/papers/422.pdf.

Yanan Sui, Alkis Gotovos, Joel Burdick, and Andreas Krause. Safe exploration for optimization with Gaussian processes. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 997–1005, Lille, France, 2015. PMLR. URL https://proceedings.mlr.press/v37/sui15.html.

Elie Tamer. Partial identification in econometrics. *Annual Review of Economics*, 2(1):167–195, 2010. doi: 10.1146/annurev.economics.050708.143401.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 23–30, 2017. doi: 10.1109/IROS.2017.8202133.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109.

Matteo Turchetta, Felix Berkenkamp, and Andreas Krause. Safe exploration in finite Markov decision processes with Gaussian processes. In *Advances in Neural Information Processing Systems*, volume 29, pp. 4312–4320. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf.

Antonio Vicino and Giovanni Zappa. Sequential approximation of feasible parameter sets for identification with set membership uncertainty. *IEEE Transactions on Automatic Control*, 41(6):774–785, 1996. doi: 10.1109/9.506230.

Andrew Wagenmaker, Guanya Shi, and Kevin G. Jamieson. Optimal exploration for model-based RL in nonlinear systems. In *Advances in Neural Information Processing Systems*, volume 36, pp. 15406–15455. Curran Associates, Inc., 2023. doi: 10.52202/075280-0678.

Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16 (2):117–186, 1945. doi: 10.1214/aoms/1177731118.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009.

Yi-Kwei Wen. Method for random vibration of hysteretic systems. *Journal of the Engineering Mechanics Division*, 102(2):249–263, 1976. doi: 10.1061/JMCEA3.0002106.

Doguhan Yeke, Elif Su Temirel, Ananth Shreekumar, Brandon Lee, Dongyan Xu, and Z. Berkay Celik. The yes-man syndrome: Benchmarking abstention in embodied robotic agents. *arXiv preprint arXiv:2605.20544*, 2026. URL https://arxiv.org/abs/2605.20544.

Kevin Zakka, Yuval Tassa, and MuJoCo Menagerie Contributors. MuJoCo Menagerie: A collection of high-quality simulation models for MuJoCo, 2022. URL http://github.com/google-deepmind/mujoco_menagerie.

A NOTATION

e, s, τ	embodiment, scene, task	$\mathcal{F}$	function class, §2
$u, u^\star$	capability axis; the binding one	$B, \widetilde{B}$	nominal ceiling; $\widetilde{B} = \min(B, a_{\max})$
$h_{e,s}(u)$	capability signature, §2	$\theta(f)$	$\sup_{[0,a_{\max}]} f$, the target functional
$\theta^\star$	$h_{e,s}(u^\star)$, the quantity certified	$[L_T, U_T]$	identification interval, §2
$\hat{h}^\star$	demand, from scene geometry	$\hat{\mu}_j, r_j$	band centre and radius at α_j
κ	conservatism margin, 1.15	ς	noise scale (unknown, may vary with α)
$c_{\text{feas}}, c_{\text{inf}}$	$\kappa\hat{h}^\star$; $\hat{h}^\star$	ρ_{eB}	mixture-boundary parameter
α	probe amplitude	α_{cert}	total significance budget, 0.05
$\mathcal{A}$	amplitude grid, $J = 24$	$\alpha^{\text{CS}}, \alpha^{(3)}$	0.04 band; 0.01 split, W_3
$a_{\max}$	nominal amplitude ceiling	$\alpha^{(R)}$	0.05, optional branch, off-budget
$A_{\text{safe}}(\zeta)$	safety envelope, Eq. (1)	$\theta^\star_{\text{safe}}$	$\max_{j:\alpha_j \le A_{\text{safe}}} f(\alpha_j)$
ζ	safety level, main 0.70	$T_{\max}, n_0$	test budget 96; calibration 8
f	response function on $[0, a_{\max}]$	N_{ch}	certification channels per task, 1
$W_k, \bar{\theta}_k$	witness and its implied bound	Π, ϵ_Π	task progress; its tolerance
(S_k)	shape assumption used by W_k	Δ_+	certifiable gap, $\theta^\star_{\text{safe}} - c_{\text{feas}}$
Cov, Risk, MCR	coverage, probe risk, mis-certification	π_{inf}	$\mathbb{P}_{\mathcal{P}}[\theta^\star < \hat{h}^\star]$, infeasible mass
$\mathcal{P}$	instance distribution, §2	H_ϵ	$\sum_j (\Delta_j + \epsilon)^{-2}$, sample complexity
pen, λ	selection penalty and its weight, 2.0	Δ_j	$\lvert f(\alpha_j) - c_{\text{feas}}\rvert$, per-arm gap

B PROOFS

Throughout, $A_{\text{safe}} < a_{\max}$, all probes satisfy $\alpha_t \le A_{\text{safe}}$, and $\mathcal{F}$, θ, $\widetilde{B}$, L_T, U_T are as in §2. The safe grid $\mathcal{A} \cap [0, A_{\text{safe}}]$ is assumed non-empty whenever its maximum is used. The consistent set the main text names is, in full: given $\mathcal{D}_T = \{(\alpha_t, y_t)\}_{t \le T}$, the set $\mathcal{F}(\mathcal{D}_T)$ collects those $f \in \mathcal{F}$ with $f(\alpha_t) = y_t$ at every probed point in the noiseless case, and with $\underline{\mu}_j \le f(\alpha_j) \le \bar{\mu}_j$ at every visited grid point in the noisy one. The second form is what makes $f^\star \in \mathcal{F}(\mathcal{D}_T)$ hold for all T simultaneously with probability at least $1-\alpha^{\text{CS}}$: it is the band event $\mathcal{C}$ of the proof of Proposition 3.2, read as a statement about which functions remain admissible.

B.1 LEMMA B.1: SMOOTH INDISTINGUISHABILITY

Lemma B.1 (Smooth indistinguishability). *Let $\mathcal{F}$ be as in §2, let $A_{\text{safe}} < a_{\max}$, and let $\varepsilon \in (0, \widetilde{B})$ with $h := \widetilde{B} - \varepsilon$. For every $f_0 \in \mathcal{F}$ there is an $f_1 \in \mathcal{F}$ with* (1) $f_1|_{[0,A_{\text{safe}}]} = f_0|_{[0,A_{\text{safe}}]}$; (2) $\theta(f_1) = \max\{\theta(f_0), h\} \ge \widetilde{B} - \varepsilon$; *and* (3) *if f_0 is C^∞, or monotone non-decreasing, or Lipschitz, then so is f_1.*

Proof. Fix $f_0 \in \mathcal{F}$ and $\varepsilon \in (0, \widetilde{B})$, and set $h := \widetilde{B} - \varepsilon$. If $\theta(f_0) \ge h$, take $f_1 := f_0$ and all three claims are immediate. Assume therefore $\theta(f_0) < h$.

Since $h < \widetilde{B} \le a_{\max}$ and $A_{\text{safe}} < a_{\max}$, the interval $(\max\{A_{\text{safe}}, h\}, a_{\max})$ is non-empty; choose $\alpha^\dagger$ in it and $\alpha^\ddagger \in (\alpha^\dagger, a_{\max}]$. Let $\psi : [0, a_{\max}] \to [0, 1]$ be non-decreasing with $\psi \equiv 0$ on $[0, \alpha^\dagger]$, $\psi \equiv 1$ on $[\alpha^\ddagger, a_{\max}]$, and $\psi \in C^\infty$ (a standard mollifier bump; for the Lipschitz and monotone clauses a piecewise-linear ψ also suffices). Define the *interpolation*

$$f_1 := (1 - \psi)\, f_0 + \psi\, h. \tag{3}$$

Admissibility, $f_1 \in \mathcal{F}$. Measurability is clear. For the range, $f_1(\alpha)$ is a convex combination of $f_0(\alpha) \in [0, B]$ and $h \in (0, \widetilde{B}) \subseteq [0, B]$, hence lies in $[0, B]$. For the constraint $f_1 \le \alpha$ we distinguish two regions. On $[0, \alpha^\dagger]$ we have $\psi = 0$, so $f_1 = f_0 \le \alpha$. On $(\alpha^\dagger, a_{\max}]$, both arguments of the convex combination are at most α: indeed $f_0(\alpha) \le \alpha$ by assumption, and $h < \alpha^\dagger < \alpha$ by the choice $\alpha^\dagger > h$. Convex combinations preserve the bound, so $f_1(\alpha) \le \alpha$. This is exactly where an *additive* bump $f_0 + \psi(h - \theta(f_0))$ would fail: it could push the value above α near $\alpha^\dagger$.

Claim 1. $A_{\text{safe}} < \alpha^\dagger$, so $\psi \equiv 0$ on $[0, A_{\text{safe}}]$ and $f_1 = f_0$ there.

Claim 2. $f_1(\alpha) = h$ for $\alpha \geq \alpha^\ddagger$, hence $\theta(f_1) \geq h = \widetilde{B} - \varepsilon$. Conversely every value of f_1 is a convex combination of a value of f_0 and h, so $\theta(f_1) \leq \max\{\theta(f_0), h\}$, and the two bounds give equality.

Claim 3. If f_0 is continuous then so is f_1, being a convex combination of continuous functions with continuous weights, and if $f_0 \in C^\infty$ then $f_1 \in C^\infty$ for the same reason. If f_0 is Lipschitz with constant ℓ_0, then $f_1' = (1-\psi)f_0' + \psi'(h - f_0)$ almost everywhere is bounded by $\ell_0 + \|\psi'\|_\infty(B + \widetilde{B})$, so f_1 is Lipschitz with a larger but finite constant, which is precisely why an *unknown* Lipschitz constant provides no protection. If f_0 is non-decreasing with $\theta(f_0) = \sup f_0 < h$, then for $\alpha < \beta$ in $(\alpha^\dagger, \alpha^\ddagger)$, $f_1(\beta) - f_1(\alpha) = \big(1 - \psi(\beta)\big)\big(f_0(\beta) - f_0(\alpha)\big) + \big(\psi(\beta) - \psi(\alpha)\big)\big(h - f_0(\alpha)\big) \geq 0$, both terms being non-negative, so f_1 is non-decreasing there; it agrees with the non-decreasing f_0 to the left and is constant to the right, and at $\alpha^\dagger$ the two pieces agree in value, so f_1 is non-decreasing throughout. □

B.2 Two auxiliary lemmas

Lemma B.2 (Lower endpoint)**.** *In the noiseless case $L_T = \max_{t \leq T} y_t$, and this is attained: the constant-completion f_{low} that equals $f^\star$ on the probed points and 0 elsewhere lies in $\mathcal{F}$ and is data-consistent.*

Proof. Any data-consistent f satisfies $f(\alpha_t) = y_t$ for probed α_t, hence $\theta(f) \geq \max_t y_t$, giving $L_T \geq \max_t y_t$. For the reverse, f_{low} is Borel, lies in $[0, B]$, and satisfies $f_{\mathrm{low}}(\alpha) \leq \alpha$ (it equals $f^\star(\alpha_t) \leq \alpha_t$ at probed points and 0 elsewhere), so $f_{\mathrm{low}} \in \mathcal{F}$ and $\theta(f_{\mathrm{low}}) = \max_t y_t$. □

Lemma B.3 (Upper endpoint and witness upper bounds)**.** *Without shape assumptions, $U_T = \widetilde{B}$ whenever the data-consistent class is non-empty. For the shaped case, fix a realised history and a non-empty class $\mathcal{C}_\mathcal{S}$ of complete instances compatible with that history, including its response, calibration and auxiliary measurements, and satisfying the shape assumptions and side conditions used below. Suppose each retained witness bound $\bar{\theta}_k$ bounds $\theta(g)$ for* every *$g \in \mathcal{C}_\mathcal{S}$. Then*

$$U_{\mathrm{id}}^{\mathcal{S}} := \sup_{g \in \mathcal{C}_\mathcal{S}} \theta(g) \;\leq\; \overline{U}^{\mathcal{S}} := \min\Big\{\widetilde{B}, \min_{k \in \mathcal{K}} \bar{\theta}_k\Big\}. \tag{4}$$

Here $\theta(g)$ denotes the capability of the response component of g; $\min \emptyset = +\infty$. A bound is retained only when its confidence coverage, shape assumption and trigger side conditions hold throughout $\mathcal{C}_\mathcal{S}$. If the safe-grid response is known exactly, the grid in (S1) is non-empty, and (S1) holds for all candidates, one may also retain $\max\{\bar{\varrho}^{\mathrm{safe}} a_{\max}, a_{\min}\}$, where $\bar{\varrho}^{\mathrm{safe}} = \max_{j : a_{\min} \leq \alpha_j \leq A_{\mathrm{safe}}} f(\alpha_j)/\alpha_j$ is then common to all candidates. With noisy data, its replacement must use a valid upper confidence bound, as in W_1. These are upper bounds, not an equality characterisation of the identification endpoint or an assertion of attainment.

Proof. If $\widetilde{B} = 0$, every admissible response is zero and the first claim is immediate. Otherwise choose a response from the non-empty data-consistent class. The first claim is Lemma B.1 with $\varepsilon \downarrow 0$: for every ε there is a data-consistent f_1 with $\theta(f_1) \geq \widetilde{B} - \varepsilon$, so the supremum defining U_T is $\widetilde{B}$; it cannot exceed $\widetilde{B}$ because $\widetilde{B} = \sup_\mathcal{F} \theta$. For the second, every retained bound holds pointwise for every candidate, so it continues to hold after taking the supremum over $\mathcal{C}_\mathcal{S}$. Lemmas B.7–B.10 justify a retained witness bound when their confidence events and side conditions hold for each candidate. The optional exact-grid (S1) term follows by bounding responses above $a_{\min}$ by $\bar{\varrho}^{\mathrm{safe}}\alpha$ and responses below $a_{\min}$ by α. No construction attaining the minimum is required or claimed. □

B.3 Proposition 3.1: certification asymmetry

Proof. (a), the two endpoints. Lemma B.2 gives L_T; Lemma B.3 gives $U_T = \widetilde{B}$, which involves no data. Monotonicity of L_T is immediate since the maximum is over a growing set. Revealing $f^\star$ on all of $[0, A_{\mathrm{safe}}]$ changes the argument of Lemma B.1 only by making the "agrees on $[0, A_{\mathrm{safe}}]$" clause binding on an interval, which clause 1 of the lemma already provides. The lower endpoint is the largest value the data force, which is $\theta^\star_{\mathrm{safe}}$ when the safe grid is exhausted and $\sup_{\alpha \leq A_{\mathrm{safe}}} f^\star(\alpha)$ when the whole safe interval is revealed, in both cases by the completion of Lemma B.2. The grid version is the operative one, since Algorithm 1 probes only $\mathcal{A}$.

Under noise the upper endpoint is unchanged, because Lemma B.1 is a statement about $\mathcal{F}$, not about the observation process. For the lower endpoint, $\hat{\mu}_j(t) - r_j(t)$ is a time-uniform lower confidence bound for $f(\alpha_j)$ at level $\alpha^{\mathrm{CS}}/(2JN_{\mathrm{ch}})$ by Howard et al. (2021); a union bound over the $2JN_{\mathrm{ch}}$ edge terms of the two-sided band makes $\max_j(\hat{\mu}_j - r_j)^+$ a simultaneous lower bound for $\max_j f(\alpha_j) \le \theta^\star$ with probability at least $1 - \alpha^{\mathrm{CS}}$, uniformly in t. For equality with the lower endpoint of the non-empty band-consistent class, set $f_{\mathrm{low}}(\alpha_j) = \max\{0, \underline{\mu}_j\}$ at visited grid points and set it to zero elsewhere. Non-emptiness ensures these values satisfy the upper band edges and the bounds $f \le B$ and $f(\alpha) \le \alpha$. This finite-point completion is admissible and has supremum $\max_j(\underline{\mu}_j)^+$. Every consistent response has a supremum at least this large. With no visited points, both lower endpoints are defined as zero.

(b.1) The hypothesis $\theta^\star_{\mathrm{safe}} > c_{\mathrm{feas}}$ is exactly $\Delta_+ > 0$ for the gap of Proposition B.11, which then gives a finite n after which $L_n \ge c_{\mathrm{feas}}$ with the stated probability. Stating the hypothesis on the grid maximum rather than on $\sup_{\alpha \le A_{\mathrm{safe}}} f^\star(\alpha)$ is necessary: a peak strictly between two grid points is unreachable by a probe, so it cannot make any event certifiable in finitely many steps.

(b.2) Work in the scalar-observation experiment stated in the proposition, with a common measurable history space and a common conditional observation kernel $Q(dy \mid h, \alpha, f(\alpha))$. The initial information has the same law for responses agreeing on $[0, A_{\mathrm{safe}}]$, and the algorithm's internal random seed is drawn from the same response-independent law. All information available to the algorithm is included in this history; no additional channel may reveal the response outside the safe interval. Fix an algorithm $\mathcal{A}$ confined to $[0, A_{\mathrm{safe}}]$ whose probability of a false CERTIFIED-INFEASIBLE is at most α_{cert}, the supremum being over the feasible instances $\{f \in \mathcal{F} : \theta(f) \ge c_{\mathrm{inf}}\}$ on which CERTIFIED-INFEASIBLE is the wrong verdict, and fix any $f_0 \in \mathcal{F}$. Requiring the bound only there is what keeps the hypothesis a validity condition; asked everywhere, it would instead assume $\mathcal{A}$ powerless to begin with. Apply Lemma B.1 with $\varepsilon < \widetilde{B} - c_{\mathrm{inf}}$, obtaining f_1 that agrees with f_0 on $[0, A_{\mathrm{safe}}]$ and has $\theta(f_1) \ge \widetilde{B} - \varepsilon > c_{\mathrm{inf}}$, so f_1 is feasible in the sense that CERTIFIED-INFEASIBLE is wrong on it. Conditional on the same history and random seed, the action rule is the same under f_0, f_1; their responses agree at that action, so the common kernel gives the same next observation law. Induction gives equality of every finite augmented-history law, including the random seed and actions. An output by any adaptive finite stopping time is measurable with respect to such histories. For unbounded stopping, the event of eventually issuing a verdict is the increasing union of its finite-time events; non-termination counts as UNCERTIFIED. Hence $\mathbb{P}_{f_0}[\mathcal{A} \text{ outputs CERTIFIED-INFEASIBLE}] = \mathbb{P}_{f_1}[\mathcal{A} \text{ outputs CERTIFIED-INFEASIBLE}] \le \alpha_{\mathrm{cert}}$, the inequality by the error guarantee applied at f_1, which lies in the feasible set over which the supremum is taken. Since f_0 was arbitrary, $\mathcal{A}$ outputs CERTIFIED-INFEASIBLE with probability at most α_{cert} on every instance, so no such algorithm has non-trivial power. Note the necessity of $\varepsilon < \widetilde{B} - c_{\mathrm{inf}}$: if $c_{\mathrm{inf}} = \widetilde{B}$ this construction does not cross the threshold; if $c_{\mathrm{inf}} > \widetilde{B}$, infeasibility follows directly from the nominal bound. Conditional mean-zero noise by itself is insufficient for the equality of history laws used here.

(b.3) By $f \le \alpha$, every safe-grid response satisfies $f(\alpha_j) \le \alpha_j \le A_{\mathrm{safe}}$. Noiselessly this gives $L_T \le A_{\mathrm{safe}} < c_{\mathrm{feas}}$. Under noise, on the simultaneous band-coverage event, $\hat{\mu}_j(t) - r_j(t) \le f(\alpha_j)$ at every visited point, so $L_t \le A_{\mathrm{safe}} < c_{\mathrm{feas}}$ for all t as well. Thus the CERTIFIED-FEASIBLE branch cannot fire on that event. By (b.2), the CERTIFIED-INFEASIBLE branch has no uniform guarantee over the unrestricted class; in the absence of a triggered witness, the algorithm returns UNCERTIFIED. □

Corollary B.4 (Regularity does not help unless its constant is known)**.** *The upper-endpoint conclusion of Proposition 3.1(a) and the testing conclusion (b.2), under its observation model, hold on the subclasses of continuous functions, of monotone non-decreasing functions, and of Lipschitz functions with unspecified constant, whenever the corresponding consistent class is non-empty. On the unimodal class, the same conclusions hold for a safe restriction that is non-decreasing on* $[0, A_{\mathrm{safe}}]$.

Proof of Corollary B.4. Lemma B.1 and its proof preserve continuity, monotonicity, or a finite unspecified Lipschitz constant, so its upper-endpoint construction and the history-law argument for (b.2) apply within each subclass. For a unimodal f_0 non-decreasing on $[0, A_{\mathrm{safe}}]$, choose any $h < \widetilde{B}$. If $\theta(f_0) \ge h$, retain f_0. Otherwise first replace its tail by the constant $f_0(A_{\mathrm{safe}})$, leaving its safe restriction untouched. This gives an admissible globally non-decreasing function with supremum

below h. Applying the interpolation (3) to this function yields an admissible non-decreasing, hence unimodal, extension reaching h. If instead f_0 declines inside $[0, A_{\text{safe}}]$, unimodality forbids a later rise and the construction fails, which is the content of W_4. □

B.4 Proposition B.5: coverage for the prescribed certificate family

Fix a certification protocol: its amplitude grid, envelope, thresholds, test budget, calibration procedure, auxiliary measurements, safety checks and CF/W_1–W_4 decision gates. Let $\mathfrak{A}$ be the measurable selection and stopping policies obeying this protocol. An instance x specifies its response, demand, physical side conditions and the joint observation law for all channels. All policies and instances use a common measurable augmented history space that includes actions, calibration, auxiliary and safety feedback, and internal randomisation. The instance distribution and the policy-induced history laws are measurable, so the probabilities below exist.

For each x, let $\mathcal{H}_x$ contain every history that can be generated under x by an allowed policy, with the fixed budget and safety rules. Define $\mathcal{H}_x^W$ as its histories on which all confidence statements used by the CI gate cover their corresponding truths at every step. Write $d_W(h) = 1$ when the prescribed decision rule issues CI on history h, including its priority and dual-trigger checks. Define

$$\begin{aligned} \mathcal{E}_{\text{feas}} &:= \{x : \theta^\star_{\text{safe}}(x) \geq c_{\text{feas}}(x)\}, \\ \mathcal{E}_{\mathcal{W}} &:= \{x : \exists h \in \mathcal{H}_x^W \text{ with } d_W(h) = 1\}. \end{aligned} \tag{5}$$

Assume these instance events are measurable. A verified measurable superset of the second event may replace it throughout; this weakens the bound without changing its validity. The definition concerns the fixed gate's opportunity to issue CI on a confidence-valid history. It makes no claim that a closed formula is the exact identification endpoint, or that a supremum is attained. It also does not identify these events with previously supplied aggregate statistics; using the bound numerically requires separately establishing the event probabilities for this protocol.

Proposition B.5 (Coverage bound for the prescribed certificate family)**.** *For each $\mathcal{A} \in \mathfrak{A}$ and instance x, suppose simultaneous confidence events G_F and G_W satisfy $\mathbb{P}_x^{\mathcal{A}}(G_F^c) \leq \alpha_{\text{cert}}$ and $\mathbb{P}_x^{\mathcal{A}}(G_W^c) \leq \alpha_{\text{cert}}$. On G_F, every safe-grid lower bound read by CF is at most its true response; no off-grid calibration reading enters that lower-bound statistic. On G_W, the realised history belongs to $\mathcal{H}_x^W$. Then*

$$\text{Cov}(A_{\text{safe}}) \leq \min\Big\{1,\, \mathbb{P}_{\mathcal{P}}[\mathcal{E}_{\text{feas}}] + \mathbb{P}_{\mathcal{P}}[\mathcal{E}_{\mathcal{W}}] + 2\alpha_{\text{cert}}\Big\}. \tag{6}$$

The bound holds across selection and stopping policies in $\mathfrak{A}$. It does not assert a limit for algorithms using other certification rules.

Proof. Fix x and $\mathcal{A}$. If CF is issued on G_F, then some safe-grid lower bound reaches $c_{\text{feas}}(x)$. Its true response is at least that lower bound, so $x \in \mathcal{E}_{\text{feas}}$. Therefore

$$\mathbb{P}_x^{\mathcal{A}}[\textsc{Certified-Feasible}] \leq \mathbf{1}[x \in \mathcal{E}_{\text{feas}}] + \alpha_{\text{cert}}.$$

If CI is issued on G_W, the realised history itself is an element of $\mathcal{H}_x^W$ with $d_W(h) = 1$. Thus $x \in \mathcal{E}_{\mathcal{W}}$, and

$$\mathbb{P}_x^{\mathcal{A}}[\textsc{Certified-Infeasible}] \leq \mathbf{1}[x \in \mathcal{E}_{\mathcal{W}}] + \alpha_{\text{cert}}.$$

Integrate these two inequalities over $\mathcal{P}$ and add them. The output verdicts are mutually exclusive, with safety abort and budget exhaustion without a certificate counted as UNC, hence $\text{Cov} = \mathbb{P}[\textsc{Certified-Feasible}] + \mathbb{P}[\textsc{Certified-Infeasible}]$. Finally $\text{Cov} \leq 1$ gives the displayed bound. No independence of the confidence events is needed. The instance events need not be disjoint, and the proof makes no assertion that the bound is attainable by a selection policy. □

The gate-opportunity inclusion above remains a coverage statement even on an instance violating a shape assumption: such an instance can belong to $\mathcal{E}_{\mathcal{W}}$. Whether its CI verdict is correct is a separate question, addressed by the assumption-failure term in Proposition 3.2. When the stated confidence events cover the fixed protocol, their per-instance failure probabilities follow from its simultaneous band and auxiliary-test budgets. Changing that protocol requires checking those guarantees again.

Remark B.6 (The (R) variant). For a declared valid Lipschitz constant L, include the corresponding regularity gate in the fixed protocol and define its opportunity event by Eq. (5) with that enlarged CI decision rule. The same inclusion proof applies, using a failure budget that covers every confidence statement read by the enlarged gate. If the main CI gate has failure budget α_{cert} and the added gate has budget $\alpha^{(R)}$, their joint confidence event has failure probability at most $\alpha_{\mathrm{cert}} + \alpha^{(R)}$ by a union bound. Together with the CF budget α_{cert}, the enlarged-protocol bound has additive term $2\alpha_{\mathrm{cert}} + \alpha^{(R)}$. In-envelope data can lower-bound slopes but cannot establish the upper Lipschitz bound required for extrapolation. The main certificate excludes this branch, which has its own $\alpha^{(R)}$ budget outside α_{cert}; Ablation AB9 scans L over a factor of sixteen instead of fitting it.

B.5 Witness validity

Each witness converts its trigger into an upper bound $\bar{\theta}_k$ on $\theta^\star$. The following lemmas establish the bounds used by Proposition 3.2 under the corresponding shape assumptions. They use the upper edge of the two-sided band and the shape assumption to extend a bound from grid observations to the continuum. Throughout, $\mathcal{C}$ is the band event in the proof of Proposition 3.2, $\mathcal{C}_3$ is the contact-position event, and trigger conditions are given in Appendix D. Conditional on the required band events, the remaining arguments are deterministic.

Lemma B.7 (W_1 validity)**.** *On $\mathcal{C}$, if W_1 triggers and* (S1) *holds for $f^\star$, then $\theta^\star \le \bar{\theta}_1 = \max\{\bar{\varrho}_n a_{\max},\, a_{\min}\}$.*

Proof. On $\mathcal{C}$, $f^\star(\alpha_j) \le \bar{\mu}_j$ and hence $f^\star(\alpha_j)/\alpha_j \le \bar{\mu}_j^+/\alpha_j$ for every $j \in V$ with $\alpha_j \ge a_{\min}$. Since W_1 is evaluated only when V exhausts the admissible grid, the grid maximum in (S1) is over a subset of those j, so $\max_{j:\, a_{\min} \le \alpha_j \le A_{\mathrm{safe}}} f^\star(\alpha_j)/\alpha_j \le \bar{\varrho}_n$. Then (S1) gives $f^\star(\alpha) \le \bar{\varrho}_n \alpha \le \bar{\varrho}_n a_{\max}$ for $\alpha \in [a_{\min}, a_{\max}]$, while (A6) gives $f^\star(\alpha) \le \alpha < a_{\min}$ for $\alpha \in [0, a_{\min})$. Taking the supremum over $[0, a_{\max}]$ and using (A2) gives the claim. □

Lemma B.8 (W_2 validity)**.** *On $\mathcal{C}$, if W_2 triggers on the window $[\alpha_p, \alpha_q]$ and* (S2) *holds for $f^\star$, then $\theta^\star \le \bar{\theta}_2$ as defined in* (9)*.*

Proof. The trigger certifies the hypothesis of (S2) on $\mathcal{C}$: band oscillation at most η_2 over the window means $\max_{p \le j \le q} \bar{\mu}_j - \min_{p \le j \le q} \underline{\mu}_j \le \eta_2$, and since $\underline{\mu}_j \le f^\star(\alpha_j) \le \bar{\mu}_j$ on $\mathcal{C}$, the oscillation of $f^\star$ over the window's grid points is at most η_2 as well; the width and torque-limit conditions are checked directly. Applying (S2) and then $f^\star(\alpha_j) \le \bar{\mu}_j$ to each of its two grid maxima gives $\theta^\star \le \max\{\bar{M}_q,\, \bar{M}_{pq} + \eta_2\} = \bar{\theta}_2$. Both terms are needed. The second alone bounds only the plateau and what lies to its right, so a response that rises to a peak before the window and then saturates would escape it; the first term is there to make the bound hold on $[0, \alpha_q]$. □

Lemma B.9 (W_3 validity)**.** *On $\mathcal{C}_3$, if W_3 triggers on a P1 displacement channel and* (S3) *holds for $f^\star$, then $\theta^\star \le \bar{\theta}_3 = \bar{z}_K^+$.*

Proof. On $\mathcal{C}_3$ the contact-position band covers, so $z^\dagger \le \bar{z}_K^+$. The repeatability test bounds the dispersion of the K independent approaches, which licenses treating them as observations of one position $z^\dagger$. (S3) then gives $\theta^\star \le z^\dagger \le \bar{z}_K^+$. The channel restriction turns the first inequality into a comparison of like with like: on a P1 channel $f^\star$ and $z^\dagger$ are both displacements along the capability axis, so the block position is literally a ceiling on the reachable displacement. On a load channel it would not be, so the quota sets W_3 to NOT-EVALUATED there. □

Lemma B.10 (W_4 validity)**.** *On $\mathcal{C}$, if W_4 triggers at the pair (j_1, j_2) and* (S4) *holds for $f^\star$, then $\theta^\star \le \bar{\theta}_4 = \max_{j \in V:\, \alpha_j \le \alpha_{j_2}} \bar{\mu}_j^+$.*

Proof. Trigger condition (i) reads $\underline{\mu}_{j_1} - \bar{\mu}_{j_2} > \eta_4 \ge 0$. On $\mathcal{C}$ this gives $f^\star(\alpha_{j_1}) \ge \underline{\mu}_{j_1} > \bar{\mu}_{j_2} \ge f^\star(\alpha_{j_2})$, so the observed decline establishes the antecedent of (S4) on the band event. This step uses both edges of the band. (S4) then puts the peak on the grid to the left of α_{j_2}, and $f^\star(\alpha_j) \le \bar{\mu}_j$ on $\mathcal{C}$ turns that grid maximum into $\bar{\theta}_4$. Trigger condition (iii), which requires the decline to reach the envelope boundary without a detected later rise, is not used here: it cannot verify (S4), since absence

of a detected rise is not absence of a rise, and it is included as a falsification check that makes (S4) harder to violate silently. □

B.6 PROPOSITION 3.2: ANYTIME VALIDITY

Proof. Let $\mathcal{F}_t$ be the filtration generated by the probe history. The two branches read opposite edges of the same band, so the proof establishes the band events once, at the top, and every later step reads them.

The band event. Fix a $(j, \text{channel})$ pair. The empirical-Bernstein mixture supermartingale of Howard et al. (2021), whose sub-Gaussian and sub-gamma boundary family is developed in Howard et al. (2020), combined with Ville's maximal inequality gives, at crossing level $\alpha^{\mathrm{CS}}/(2JN_{\mathrm{ch}})$ on each edge,

$$\mathbb{P}\big[\exists t : \hat{\mu}_j(t) - r_j(t) > f(\alpha_j)\big] \;\le\; \frac{\alpha^{\mathrm{CS}}}{2JN_{\mathrm{ch}}}, \qquad \mathbb{P}\big[\exists t : \hat{\mu}_j(t) + r_j(t) < f(\alpha_j)\big] \;\le\; \frac{\alpha^{\mathrm{CS}}}{2JN_{\mathrm{ch}}}. \tag{7}$$

This is the reason r_j in Eq. (2) is instantiated at $\alpha^{\mathrm{CS}}/(2JN_{\mathrm{ch}})$ and not at $\alpha^{\mathrm{CS}}/(JN_{\mathrm{ch}})$: the lower edge alone would leave every $\bar{\mu}_j$ uncontrolled, and the witnesses read $\bar{\mu}_j$. Define the band event

$$\mathcal{C} := \big\{\forall t \ge 1,\ \forall j,\ \forall \text{ channel} :\ \hat{\mu}_j(t) - r_j(t) \;\le\; f(\alpha_j) \;\le\; \hat{\mu}_j(t) + r_j(t)\big\},$$

and let $\mathcal{C}_3$ be the analogous event for the contact-position band read by W_3, whose K approach statistics are instantiated at $\alpha^{(3)}/(4K)$ per edge, jointly with the χ^2 repeatability bound of Appendix D, which Lemma B.9 also uses and which is therefore run at $\alpha^{(3)}/2$. A union bound over the $2JN_{\mathrm{ch}}$ edge-pair terms of (7) gives $\mathbb{P}[\mathcal{C}^c] \le \alpha^{\mathrm{CS}}$, and the same argument over the two W_3 components gives $\mathbb{P}[\mathcal{C}_3^c] \le \alpha^{(3)}$, so

$$\mathbb{P}\big[\mathcal{C}^c \cup \mathcal{C}_3^c\big] \;\le\; \alpha^{\mathrm{CS}} + \alpha^{(3)} \;=\; 0.04 + 0.01 \;=\; \alpha_{\mathrm{cert}}. \tag{8}$$

Equation (7) is uniform in t, so $\mathcal{C}$ and $\mathcal{C}_3$ are defined for all t at once and hold at any stopping time. A fixed-time interval would not survive that, which is why the construction is a confidence sequence.

Both branches fail only outside that event. On $\mathcal{C}$, the lower edge gives $L_t \le \max_j f(\alpha_j) \le \theta^\star$ for every t, so CERTIFIED-FEASIBLE is never issued falsely. On $\mathcal{C} \cap \mathcal{C}_3$ and under the shape assumptions in force, Lemmas B.7–B.10 give $\theta^\star \le \bar{\theta}_k$ for every triggered W_k, so a triggered witness with $\bar{\theta}_k < c_{\mathrm{inf}}$ implies $\theta^\star < c_{\mathrm{inf}}$ and CERTIFIED-INFEASIBLE is never issued falsely either. Contrapositively,

$$\{\text{a false certification is ever issued}\} \;\subseteq\; \big(\mathcal{C}^c \cup \mathcal{C}_3^c\big) \;\cup\; \big\{\exists k : W_k \text{ triggerable and } (S_k) \text{ fails}\big\},$$

and (8) bounds the first term by α_{cert}, which is the proposition. The witnesses need no multiplicity correction of their own: they are deterministic measurable functionals of a band whose J-point multiplicity (7) has already paid for, and they introduce no new randomness. If some (S_k) fails on $f^\star$ the corresponding lemma does not apply, $\bar{\theta}_k$ need not bound $\theta^\star$, and that error is what the indicator term names and the significance budget does not cover.

The GP posterior enters only through α_t, which is $\mathcal{F}_{t-1}$-measurable; supermartingale validity is insensitive to how the next design point is chosen, so the bound holds for an arbitrary action rule. Marginalising over the sub-population of $\mathcal{P}$ on which the (S_k) in force hold preserves it; on the rest, the indicator term governs, and Table 4 reports the realised rate on D5a separately for that reason.

Finally, a simultaneous trigger of both branches implies $\bar{\theta}_k < c_{\mathrm{inf}} \le c_{\mathrm{feas}} \le L_t \le \bar{\theta}_k$, a contradiction; such an event therefore witnesses either a band-coverage failure or a violated (S_k). Every simultaneous trigger is recorded as a diagnostic event. Its rate combines possible failures of band coverage and of the shape assumptions; it does not by itself estimate the frequency of shape-assumption violations. □

B.7 PROPOSITION B.11: SAMPLE COMPLEXITY

Proposition B.11 (Sample complexity, feasible side)**.** *Let $\Delta_+ = \theta^\star_{\mathrm{safe}} - c_{\mathrm{feas}} > 0$. Probing the best safe arm $n = O\big(\varsigma^2\Delta_+^{-2}\log(2JN_{\mathrm{ch}}/\alpha^{\mathrm{CS}}) + \varsigma^2\Delta_+^{-2}\log\log(\varsigma^2/\Delta_+^2)\big)$ times suffices for $\mathbb{P}[L_n \ge c_{\mathrm{feas}}] \ge 1 - \alpha^{\mathrm{CS}}$. When the best safe arm is unknown, classifying the grid at c_{feas} to precision ϵ reduces to a thresholding bandit, and $T = O(H_\epsilon \log(1/\alpha^{\mathrm{CS}}))$ probes suffice with $H_\epsilon = \sum_j(\Delta_j + \epsilon)^{-2}$ and $\Delta_j = |f(\alpha_j) - c_{\mathrm{feas}}|$ (Chen et al., 2014); the same H_ϵ governs the fixed-budget error exponent of Locatelli et al. (2016).*

Proof. Fix the best safe arm $j^\star$ and probe it n times. The certificate fires once $\hat{\mu}_{j^\star} - r_{j^\star} \geq c_{\text{feas}}$, for which it suffices that $r_{j^\star}(n) \leq \Delta_+/2$ together with a deviation of $\hat{\mu}$ no worse than $\Delta_+/2$. The empirical-Bernstein boundary satisfies $r_j(n) = O\big(\sqrt{(\varsigma^2/n)(\log(2JN_{\text{ch}}/\alpha^{\text{CS}}) + \log\log n)}\big)$; setting this to $\Delta_+/2$ and solving for n gives the two stated terms, the second arising from the iterated-logarithm factor that ensures time-uniformity. The factor of two from the two-sided band (2) enters as an additive $\log 2$ inside the first logarithm and therefore leaves the stated order untouched. Both terms are sufficient conditions read off an upper bound on $r_{j^\star}$, so the statement is O and not Θ; we make no matching lower-bound claim.

When $j^\star$ is unknown, the problem is a thresholding bandit at c_{feas}, with gaps $\Delta_j = |f(\alpha_j) - c_{\text{feas}}|$. An arm at the threshold has $\Delta_j = 0$, so the unregularised sum $\sum_j \Delta_j^{-2}$ diverges. The precision parameter ϵ permits either classification for arms within ϵ of c_{feas}, giving $H_\epsilon = \sum_j (\Delta_j + \epsilon)^{-2}$ as in Locatelli et al. (2016). The fixed-confidence bound $T = O(H_\epsilon \log(1/\alpha^{\text{CS}}))$ follows from the combinatorial-bandit result of Chen et al. (2014) specialised to thresholding. The fixed-budget counterpart of Locatelli et al. (2016) has error probability decaying as $\exp(-T/H_\epsilon)$ up to logarithmic factors. □

C Assumptions and their audits

	Statement	Strength	Direction of violation, and audit
(A1)	Probing is quasi-static: $\|\dot{q}\|_\infty \leq 0.05$ rad/s, $\|\ddot{q}\|_\infty \leq 0.5$ rad/s^2	medium	Inertial terms inflate y and hence L_T: dangerous. Enforced as a hard online constraint (violation ⇒ SAFETY-ABORT); violation rate reported; violating samples excluded from band updates
(A2)	$\theta^\star = \sup_{[0,a_{\max}]} f$	definitional	Requires the primitive library to be exhaustive for $u^\star$; if not, $\theta^\star$ is under-estimated (fail-safe). Because execution is not quasi-static, the cell "truly infeasible & executed successfully" can be non-zero; it is an audit counter with a pre-registered ceiling of 2%, and its value measures how far (A1) is relaxed at execution
(A3)	$\mathbb{E}[\varepsilon_t \mid \mathcal{H}_{t-1}] = 0$, observations bounded; ς neither known nor constant in α	weak	Confidence sequence fails and all guarantees void. Variance-adaptive bands remove the calibration requirement; audited by injecting $\varsigma(\alpha)$ and checking the Clopper–Pearson upper bound on $\mathbb{P}[\theta^\star < L_T]$, plus QQ plots and excess kurtosis of calibration residuals
(A4)	$0 \leq f \leq B$, B read coarsely from the nominal model	weak	No substantive consequence; B enters only through $\widetilde{B}$. Sensitivity over $B \times \{0.5, 1, 2\}$
(A5)	f Borel measurable and nothing more	baseline	Not a condition to satisfy; it is the premise of Proposition 3.1. Its relaxation is the (R) branch
(A6)	$f(\alpha) \leq \alpha$	weak	Used in Proposition 3.1(b.3) and the W_1 bound below $a_{\min}$. Checked directly in simulation after the triple-equality audit; violations prompt a check of the reading map

	Statement	Strength	Direction of violation, and audit
(A7)	$\varrho_{\text{sens}} = \partial y/\partial\theta > 0$	medium	Identically 1 on the certification channel, as a corollary of the triple-equality convention, not a measured quantity. Note $\partial y/\partial\alpha \approx 0$ is a different derivative and is not a failure but witness W_1; W_1 additionally requires the P3 channel to show the actuator exerted force, since an unresponsive reading may only mean the primitive was misdirected
(A8)	$u^\star$ is stable during probing (bottleneck-static tasks)	strong; scope limit	Probing the wrong axis inflates L_T: dangerous. Scope written explicitly; the within-episode fraction of episodes with $u^\star$ drift $\leq 10^\circ$ is measured, with a $\geq 90\%$ gate for a task to remain in the main set
(A9)	Isolation does not alter reachability	medium	Two directions: a virtual wall removing reachable configurations under-estimates L_T (fail-safe); an anchor altering contact dynamics inflates force readings (dangerous). Audited by a Kolmogorov–Smirnov comparison of L_T against unisolated control seeds; $\epsilon_{\text{wall}} = 2\,\text{cm}$ of slack; a rigidified weld, not added mass; anchor residual displacement measured against ϵ_Π
(A10)	$\hat{h}^\star$ exactly computable from scene geometry	medium	Bias shifts both thresholds together, and the small direction is dangerous. Holds in simulation; on hardware it needs perception and is a stated limitation. $\pm 10\%$ injected-bias sensitivity
(A11)	The GP posterior is a reasonable approximation	weak	Costs budget only; the certification branch does not read the posterior. Online leave-one-out residual z-scores with a non-parametric fallback
(A12)	$\mathcal{W} = \{W_1, \ldots, W_4\}$ is complete	not assumed	A missed witness lowers $\text{Cov}_{\mathcal{W}}$ and cannot cause a false certificate. NOT-TRIGGERED and NOT-EVALUATED are recorded separately so that a budget shortfall is never read as evidence of feasibility

The four shape assumptions below supplement (A1)–(A12). Each witness uses one continuum-to-grid inequality to bound the supremum over an interval from observations at J grid points, as formalised in Lemmas B.7–B.10.

	Statement	Strength	Direction of violation, and audit
(S1)	No delayed breakout: $\sup_{\alpha\in[a_{\min},a_{\max}]} f(\alpha)/\alpha \le \max_{j:\, a_{\min}\le\alpha_j\le A_{\text{safe}}} f(\alpha_j)/\alpha_j$	medium	Forbids a jump after some threshold; counterexamples are static-friction breakaway and transmission deadband $f(\alpha) = (\alpha - \delta)_+$, for which the violation factor $\frac{1-\delta/a_{\max}}{1-\delta/A_{\text{safe}}}$ is explicit and, for $\delta \le 0.1A_{\text{safe}}$, bounded by 1.111, absorbed by $\kappa = 1.15$. P3's minimum amplitude is set above twice the nominal backlash. It must be stated as a difference quotient, since (A4) and (A6) already force $f(0) = 0$ and a point-value form would collapse the admissible class to $\{0\}$. The floor $a_{\min}$ is the same one that keeps $\bar{\theta}_1$ finite in Lemma B.7; below it, (A6) already gives $f(\alpha) \le \alpha < a_{\min}$
(S2)	Saturation is irreversible. Write M_q for $\max_{j:\alpha_j\le\alpha_q} f(\alpha_j)$ and M_{pq} for $\max_{p\le j\le q} f(\alpha_j)$. On a window $[\alpha_p, \alpha_q]$ of width $\ge w_{\min}$ over which f oscillates by at most η_2 at the torque limit, $\sup_{(0,a_{\max}]} f \le \max\{M_q,\, M_{pq} + \eta_2\}$	medium	Violated by re-engagement after slip; the mechanistic side condition requires the actuator to be at its torque limit throughout the plateau. Written, like (S1), as a continuum-to-grid inequality: the second term is the rightward saturation claim, the first says the pre-plateau peak is seen on the grid. Without that first term a peak before the plateau escapes the bound
(S3)	Static scene: a repeatable geometric block at position $z^\dagger$ caps the same-axis displacement, $\sup_{\alpha\in(0,a_{\max}]} f(\alpha) \le z^\dagger$	medium	Violated if the obstacle can move or be circumvented; requires the block position to repeat across independent trials, and uses its own significance budget $\alpha^{(3)}$. Stated only for P1 displacement channels, where $z^\dagger$ and f carry the same units
(S4)	Unimodality, in grid form: if $f(\alpha_{j_1}) > f(\alpha_{j_2})$ for some $\alpha_{j_1} < \alpha_{j_2}$ then $\sup_{\alpha\in(0,a_{\max}]} f(\alpha) \le \max_{j:\,\alpha_j\le\alpha_{j_2}} f(\alpha_j)$	medium	Also continuum-to-grid: one peak, no second rise, and the peak is attained on the grid left of the decline. The certifying form requires the decline to extend to the envelope boundary; the observational form on the auxiliary transfer channel reports state-dependent degradation but implies no bound and enters no verdict

The main branch does not require an accurately calibrated regularity constant. Variance adaptivity uses ς only to set the order of ρ_{eB}. The certification reading and $\hat{h}^\star$ share the same unit, origin, and frame, so the reading-to-capability map is the identity and $\varrho_{\text{sens}} \equiv 1$. The bound $\widetilde{B}$ and envelope A_{safe} require a coarse reading or pre-registration. The optional (R) branch additionally requires a Lipschitz bound.

D METHOD DETAILS

D.1 CONFIDENCE ALLOCATION AND PROBE BUDGET

The significance budget is split once. Here $\alpha^{\text{CS}} = 0.04$ covers the whole band, and with it L_t, W_1, W_2 and W_4, all deterministic functionals of that band, while $\alpha^{(3)} = 0.01$ covers the two independent statistics W_3 reads, its contact-position band and its repeatability test, at $\alpha^{(3)}/2$ each. When the shape assumptions hold, false certification in either branch requires a failure of one of the two band events, whose union bound is $\alpha_{\text{cert}} = \alpha^{\text{CS}} + \alpha^{(3)}$; the optional regularity branch carries its own $\alpha^{(R)} = 0.05$ off-budget.

On the feasible side, the probe requirement scales as $1/\Delta_+^2$ for $\Delta_+ = \theta^\star_{\text{safe}} - c_{\text{feas}}$, with an iterated-logarithm factor for time-uniform validity. When the best safe amplitude is unknown, thresholding-bandit results apply (Chen et al., 2014; Locatelli et al., 2016); Proposition B.11 states the resulting bounds. We set $T_{\max} = 96$ from the quota table, which requires six repeats at every admissible grid amplitude before witness evaluation. Checking this budget against the complexity expression

Algorithm 1 SOMA-Probe (one episode)

Require: scene s, axis $u^\star$, demand $\hat{h}^\star$, margin κ, level ζ

$c_{\text{feas}} \leftarrow \kappa\hat{h}^\star$; $c_{\text{inf}} \leftarrow \hat{h}^\star$; $\mathcal{A} \leftarrow \textsc{Grid}(0, a_{\max}, J{=}24)$; $\textsc{safe} \leftarrow \{\alpha \in \mathcal{A} : \alpha \le A_{\text{safe}}(\zeta)\}$
$\textsc{snap}_0 \leftarrow \textsc{GetState}$ ▷ before isolation is installed
install isolation (weld anchor, virtual wall); run $n_0 = 8$ calibration steps at $0.1A_{\text{safe}}$ ▷ outside the test budget
freeze ρ_{eB} in both bands; $L \leftarrow 0$; $\mathcal{W} \leftarrow \emptyset$
for $t = 1$ **to** $T_{\max} = 96$ **do**
  if $L \ge c_{\text{feas}}$ **then return** CERTIFIED-FEASIBLE
  end if
  if $\exists W_k \in \mathcal{W}$ with $\bar{\theta}_k < c_{\text{inf}}$ **then return** CERTIFIED-INFEASIBLE
  end if
  $\alpha_t \leftarrow \arg\max_{\alpha \in \textsc{safe}}[\text{IG}(\alpha) - \lambda\,\text{pen}(\alpha)]$ ▷ $\lambda = 2.0$
  $y_t, \textsc{ok} \leftarrow \textsc{ExecuteInSitu}(\alpha_t)$; **if** $\neg\textsc{ok}$ **return** SAFETY-ABORT
  update the certification band; $L \leftarrow \max_j(\hat{\mu}_j - r_j)^+$; $\mathcal{W} \leftarrow \textsc{DetectWitnesses}$; update the GP posterior
end for
repeat both checks once more: **return** CERTIFIED-FEASIBLE if $L \ge c_{\text{feas}}$, CERTIFIED-INFEASIBLE if a witness fires, else UNCERTIFIED with $[L, \widetilde{B}]$
finally: release isolation, assert $\Pi \equiv 0$ and anchor residual $< \epsilon_\Pi$, restore $\textsc{snap}_0$

gives the scope boundary $\Delta_+ = 1.0\varsigma$. Proposition 3.1(b.2) rules out a corresponding finite-sample guarantee for infeasibility over the unrestricted class in its stated scalar-observation experiment.

D.2 Probe selection and state isolation

Amplitudes maximise $\text{IG}(\alpha) - \lambda\,\text{pen}(\alpha)$ over the masked grid, with $\text{IG}(\alpha) = \frac{1}{2}\ln(1 + \sigma^2(\alpha)/\varsigma^2)$ and $\sigma^2(\alpha)$ the posterior predictive variance at α of a Matérn-$3/2$ GP fitted to the probe returns so far. The penalty $\text{pen}(\alpha) \in [0, 1]$ is a deterministic logistic function of the offline-scanned margin to the nearest safety criterion, a search heuristic distinct from the population damage rate $\text{Risk}(A_{\text{safe}})$ the experiments report. The posterior never touches the certificate: it decides where to probe, while L_t and the witnesses come from the band alone, so Proposition 3.2 holds under an arbitrary, even adversarial, action rule. The IG formula is the classical point-wise information gain of a Gaussian observation (Srinivas et al., 2010); from Hübotter et al. (2024b) we take the framing, their formalisation of the case where the sampling domain and the prediction target differ.

Probing happens in the task scene, since $\theta^\star$ is scene-indexed, and it must not accomplish the task, or comparing probing against executing becomes circular. A welded massless mocap body absorbs probe forces, a thin virtual wall outside the goal set keeps a reach probe out of the completion region, and each primitive restores the integration state. Task progress and the weld's residual displacement are checked separately, so a leaking anchor surfaces as a failed check and never as slack.

Probe primitive library. P1 (end-effector displacement along d) is the certification channel for reach and clearance tasks, reading the absolute projection $\langle p_{\text{ee}} - p_{\text{base}}, d\rangle$ for T1 and the entry depth relative to the gap plane for T3. P2 (end-effector force) is the certification channel for load tasks; P2′ adds a complete load–unload cycle, which is the only excitation on which a hysteresis witness can be distinguished. P3 (joint reciprocation) and P4′ (stepped transfer-ratio sweep) are auxiliary mechanism channels: they supply deterministic side checks such as "the actuator did exert force" and "the actuator is against its torque limit", enter no $\bar{\theta}_k$ and no L_t, and consume no significance budget. For P4′ the primary observation is the end-effector force projection divided by the equivalent tip force obtained by pushing the commanded torque through the full Jacobian and contact solver. The simulator's own constraint-force readout is kept only as a cross-check: under quasi-static rigid anchoring it returns the injected transfer function back algebraically, so a ratio built on it would be true by construction and never measured.

Amplitude ceilings and the safety envelope. The five levels are indexed by the single degree of freedom ζ, which maps to thresholds on joint force, tip-over moment, self-collision distance, robot–

Table 3: The four structural witnesses. $\bar{\mu}_j := \hat{\mu}_j(t) + r_j(t)$ and $\underline{\mu}_j := \hat{\mu}_j(t) - r_j(t)$ are the band envelopes at α_j; $\eta_1 = \eta_4 = 3\varsigma$, $\eta_2 = 0.10\hat{h}^\star$, $w_{\min} = 0.15 a_{\max}$. A witness certifies CERTIFIED-INFEASIBLE only when $\bar{\theta}_k < c_{\inf}$. Full trigger conditions are in Appendix D.

	Observed shape	**Implied bound $\bar{\theta}_k$**	**Shape assumption spent**
W_1	Pinned at the noise floor, $\max_j \bar{\mu}_j < \eta_1$, actuator effort proven	$\max\{\bar{\varrho}_n a_{\max}, a_{\min}\}$, $\bar{\varrho}_n = \max_{\alpha_j \geq a_{\min}} \bar{\mu}_j^+ / \alpha_j$	(S1) *No delayed breakout*: on $[a_{\min}, a_{\max}]$ the ratio $f(\alpha)/\alpha$ is maximised on the safe grid
W_2	Plateau of width $\geq w_{\min}$, oscillation $\leq \eta_2$, actuator at its torque limit	$\max\{\bar{M}_q,\ \bar{M}_{pq} + \eta_2\}$, positive-part band maxima up to α_q and over the window	(S2) *Saturation is irreversible*: past the plateau $f \leq v + \eta_2$, before it the peak is on the grid
W_3	Persistent contact at a repeatable position that does not advance as the push force is ramped; P1 displacement channels only	$\bar{z}_K^+$, upper band on the block position	(S3) *Static scene*: a repeatable block at $z^\dagger$ caps same-axis displacement
W_4	Reversal, $\underline{\mu}_{j_1} - \bar{\mu}_{j_2} > \eta_4$ for $\alpha_{j_1} < \alpha_{j_2}$, reaching the boundary	$\max_{j:\alpha_j \leq \alpha_{j_2}} \bar{\mu}_j^+$	(S4) *Unimodality*: a grid decline puts the peak on the grid left of it

scene contact force, and quasi-static kinematics. The contact-force ceiling is fixed by the same threshold table. Scanning the nominal, unperturbed model offline gives $A_{\text{safe}}(\zeta)$ per primitive; the pre-registered requirement $A_{\text{safe}}(\zeta) < a_{\max}$ holds strictly at all five levels.

Primitive	**Unit**	$a_{\max}$	$A_{\text{safe}}(1.00)$ / $A_{\text{safe}}(0.70)$	**Ratio** $A_{\text{safe}}(0.70)/a_{\max}$
P1	m	0.855	0.72 / 0.60	0.70
P2	N	100	85 / 40	0.40
P3	rad	0.20	0.15 / 0.08	0.40
P4′	–	$\tau_j^{\max}$	$0.90\tau_j^{\max}$ / $0.60\tau_j^{\max}$	0.60

Full witness trigger conditions. Write $\bar{\mu}_j := \hat{\mu}_j(t) + r_j(t)$ and $\underline{\mu}_j := \hat{\mu}_j(t) - r_j(t)$ for the two edges of the band (2), and $\bar{\mu}_j^+ := \max\{\bar{\mu}_j, 0\}$, which is the edge clipped to the range f is known to take by (A4). Let V index the visited admissible grid points and $S \subseteq V$ those meeting the repeat quota $n_{\text{rep}}^{\min} = 6$. A witness returns NOT-EVALUATED unless its amplitude-role quota is met, so a budget shortfall is never recorded as NOT-TRIGGERED. That quota includes coverage: W_1, W_2 and W_4 are evaluated only once every admissible grid point has been visited at least once, so $V = \mathcal{A} \cap [0, A_{\text{safe}}]$ whenever an implied bound is read. The grid maxima in (S1), (S2) and (S4) are therefore over a set the algorithm has actually observed, as Lemmas B.7–B.10 require.

W_1, zero response. Triggers when $\max_{j \in S} \bar{\mu}_j < \eta_1$ and the auxiliary P3 channel confirms that the actuator exerted effort above its own floor. The implied bound is $\bar{\theta}_1 = \max\{\bar{\varrho}_n a_{\max},\ a_{\min}\}$ with $\bar{\varrho}_n = \max_{j \in V:\, \alpha_j \geq a_{\min}} \bar{\mu}_j^+ / \alpha_j$. The amplitude floor $a_{\min}$ is set above twice the nominal backlash and below $c_{\inf}$. Since r_j does not shrink with α_j, very small probe amplitudes can make $\bar{\mu}_j^+/\alpha_j$, and hence $\bar{\theta}_1$, too large to certify infeasibility. For amplitudes below $a_{\min}$, (A6) bounds f by $a_{\min}$, giving the second term in the maximum. The trigger is evaluated on S, while $\bar{\varrho}_n$ uses every visited point above $a_{\min}$. If all of these points also satisfy $\bar{\mu}_j < \eta_1$, then $\bar{\varrho}_n < \eta_1 / a_{\min} = 3\varsigma / a_{\min}$. In this case, $3\varsigma a_{\max}/a_{\min} < \hat{h}^\star$ together with $a_{\min} < \hat{h}^\star$ is sufficient for $\bar{\theta}_1 < c_{\inf} = \hat{h}^\star$. These two sufficient inequalities imply $(\hat{h}^\star)^2 > 3\varsigma a_{\max}$; they do not give a necessary condition for certification. The decision uses the computed $\bar{\varrho}_n$ and the threshold $\bar{\theta}_1 < c_{\inf}$, including when the sufficient inequalities do not hold. The maximum defining $\bar{\varrho}_n$ runs over V, never S, so a half-sampled point with a high reading blocks certification instead of being excluded from it; the same convention governs W_2 and W_4.

W_2, saturation plateau. Triggers on a window $[\alpha_p, \alpha_q]$ with $p, q \in S$, $\alpha_q - \alpha_p \geq w_{\min}$ and band oscillation $\max_{p \leq j \leq q} \bar{\mu}_j - \min_{p \leq j \leq q} \underline{\mu}_j \leq \eta_2$, provided the actuator saturation duty over the window's steps reaches its pre-registered fraction. The implied bound is

$$\bar{\theta}_2 \;=\; \max\bigl\{\bar{M}_q,\;\; \bar{M}_{pq} + \eta_2\bigr\}, \qquad \bar{M}_q := \max_{j \in V:\ \alpha_j \leq \alpha_q} \bar{\mu}_j^{+}, \qquad \bar{M}_{pq} := \max_{p \leq j \leq q} \bar{\mu}_j^{+}, \tag{9}$$

and among qualifying windows the one minimising $\bar{\theta}_2$ is kept. The positive parts enforce $\theta^\star \geq f(0) = 0$; a negative band edge cannot upper-bound this nonnegative supremum. W_1 uses the same truncation. The first term includes any peak preceding the plateau. For example, a response that peaks at 1.5 before settling at 1.0 has $\theta^\star = 1.5$, although the second term alone is about 1.1. With $c_{\text{feas}} = 1.4$, the feasible branch can remain silent, so the upper bound must also include the pre-plateau grid.

W_3, geometric blockage. Evaluated only on P1 displacement channels, where the contact position z and the certification reading are the same physical quantity in the same units; the pre-registered quota gives W_3 no steps on the load axis, so W_3 is NOT-EVALUATED on T2 and T4 by design and never compares a displacement against a force. It requires at least the pre-registered number of independent approaches and triggers when contact is held (normal force above F_{contact} for at least 0.3 s), the approach positions are repeatable in the sense that a χ^2 upper confidence bound on their variance at level $\alpha^{(3)}/2$ falls below ϵ^2_{geo}, and no position sampled during the subsequent force ramp exceeds $\bar{\theta}_3 = \bar{z}_K^{+}$, the upper edge of the contact-position band. The certificate is explicitly conditional on the tier's force ceiling, which the evidence pack records alongside the verdict, since a heavier push might topple the obstacle.

W_4, monotonic reversal. Evaluation requires the following three conditions. (i) There exist $j_1 < j_2$ in S with $\underline{\mu}_{j_1} - \bar{\mu}_{j_2} > \eta_4$; the leftmost such j_2 is taken. (ii) The implied bound is $\bar{\theta}_4 = \max_{j \in V,\ \alpha_j \leq \alpha_{j_2}} \bar{\mu}_j^{+}$. (iii) The decline reaches the envelope boundary: at least the pre-registered number of admissible points lie beyond α_{j_2}, and among them no later point rises significantly above an earlier one, that is, there is no pair $j_2 < k < l$ with $\underline{\mu}_l > \bar{\mu}_k$. Reading the tail through the band instead of through raw upper edges matters, because three readings on a noise floor would fail a raw ordering test about half the time on noise alone. Quantifying (i) over all grid pairs at once frees the data-driven peak location from any selective-inference correction: an existential over pairs is a deterministic functional of the band, whose multiplicity the J-point union bound has already paid for. Condition (iii) is a partial defence against (S4) failing, and D5a is the subclass on which it is deliberately made to fail; this is why the W_4^{cert} mis-certification rate is reported on its own rather than pooled. The observational form W_4^{obs} reads a drop on the P4′ transmission-ratio staircase, lives on a dimensionless axis on which the demand side never defined a threshold, and therefore carries $\bar{\theta} = \perp$: it can never imply infeasibility, and no policy decision is involved.

The action-selection penalty. pen uses the same offline scan. The action-selection posterior models the scalar response and supplies no predictive distributions for the safety criteria: joint torques, tip-over moments, clearances, contact forces and joint velocities. Writing $\hat{g}_c(\alpha)$ for the peak of criterion c over the whole primitive trajectory at amplitude α on the nominal model, and $g_c^{\text{thr}}(\zeta)$ for its threshold,

$$\text{pen}(\alpha) = \max_c \varsigma_{\text{lgt}}\left(\frac{c_{\text{infl}}\,\hat{g}_c(\alpha) - g_c^{\text{thr}}(\zeta)}{w_c}\right), \qquad w_c = 0.10\, g_c^{\text{thr}}(\zeta), \tag{10}$$

with ς_{lgt} the logistic function and the sign reversed for the criteria on which larger is safer. The logistic maps five quantities in five different units onto a common dimensionless scale, which leaves λ measured in nats. The inflation factor $c_{\text{infl}} = 1.3$ is used to allow for higher criterion peaks in degraded bodies at the same amplitude. We assess its adequacy with a separate diagnostic using a GP for each criterion. Because pen affects action selection, errors in this penalty affect budget use while certificate validity is governed by the certification branch.

Hyperparameters. $J = 24$ grid points; $n_0 = 8$ calibration steps, outside the test budget and entering no hypothesis test, whose only purpose is to initialise the variance process and fix the order of ρ_{eB} independently of the test data; $T_{\max} = 96$; $n_{\text{rep}}^{\min} = 6$ repeats per grid point, which is derived and not chosen: it is the smallest n at which inverting the band radius $r_j(n)$ against

the trigger threshold $\eta_1 = 3\varsigma$ leaves W_1 able to fire at all under the mixture boundary in force, so a smaller quota would only produce NOT-EVALUATED where NOT-TRIGGERED was meant; $\kappa = 1.15$, swept over $\{1.0, 1.15, 1.25, 1.5\}$; $\lambda = 2.0$ nats, swept over $\{0, 0.5, 2, 8, 32\}$; $\alpha_{\text{cert}} = 0.05 = \alpha^{\text{CS}}(0.04) + \alpha^{(3)}(0.01)$, with $\alpha^{(R)} = 0.05$ off-budget; $\eta_1 = \eta_4 = 3\varsigma$, $\eta_2 = 0.10\hat{h}^\star$, $w_{\min} = 0.15 a_{\max}$; $\epsilon_{\text{wall}} = 2\,\text{cm}$; a Matérn-$3/2$ kernel for the action-selection posterior. The budget quota allocates $T_{\max}$ across grid coverage, repeats and the union of the per-witness minimum requirements, and the three agree exactly: one visit to every admissible amplitude, $n_{\text{rep}}^{\min}$ repeats at each amplitude a witness reads, and nothing over. No witness asks for more than that already buys, which is what makes NOT-EVALUATED a statement about the quota, never about the budget running out.

E EXPERIMENT DETAILS

Tables 4–12 report certification rates, complete primary results, method access, ablations and compute. The protocol below specifies the statistical model, utility matrix, task design and perturbations. Figure 5 reports scene-specific witness rates; Figure 6 examines the assumed Lipschitz constant.

Table 4: Certification statistics across the safety ladder. The last two columns, labelled by the event symbols, retain the reported reference rates $p^{\text{ref}}_{\text{feas}}$ and $p^{\text{ref}}_{\mathcal{W}}$. Their sum plus $2\alpha_{\text{cert}}$ defines R in Figure 3; their correspondence to the events in Proposition B.5 remains unverified. Rates are unconditional except $\text{Cov}^{\text{inf}}_{\mathcal{W}}$ (marked †), which conditions on $\theta^\star < \hat{h}^\star$; the infeasible mass is $\pi_{\text{inf}} = 0.34$. Rows correspond to safety levels. MCR includes a one-sided Clopper–Pearson 95% upper bound and excludes W_4^{cert} on D5a, where (S4) fails on every instance. That excluded rate is 0.104, with upper bound 0.187. The optional regularity branch is also reported separately. Aborted episodes count as UNCERTIFIED in the coverage denominator, accounting for the decline at $\zeta = 1.00$.

ζ	Cov_{feas}	$\text{Cov}_{\mathcal{W}}$	$\text{Cov}^{\text{inf}\,\dagger}_{\mathcal{W}}$	UNCERTIFIED	SAFETY-ABORT	MCR (CP up.)	$\mathbb{P}[\mathcal{E}_{\text{feas}}]$	$\mathbb{P}[\mathcal{E}_{\mathcal{W}}]$
0.30	0.183	0.033	0.097	0.784	0.004	0.006 (0.019)	0.203	0.058
0.50	0.290	0.053	0.157	0.657	0.021	0.008 (0.022)	0.318	0.091
0.70 (main)	0.470	0.140	0.412	0.390	0.068	0.011 (0.026)	0.494	0.183
0.85	0.488	0.157	0.462	0.355	0.243	0.013 (0.029)	0.512	0.206
1.00	0.120	0.028	0.081	0.852	0.837	0.004 (0.017)	0.540	0.229

Table 5: Primary endpoints for all methods, extending Table 1. Δ is task-only minus end-to-end success. "Cons." is the share of successful conservative fallbacks, which receive 0 under the end-to-end metric. "Wrong" is the share of episodes abandoned or refused despite a feasible ground-truth verdict. Both columns are undefined for methods without three-way verdicts. Paired differences, intervals and p values are in Table 6.

Method	**E2E↑**	**Task-only**	Δ	T_{probe}	**Total**	**Abort↓**	**Cons.**	Cov↑	ΔU↑	**Wrong**
B1 RMA	0.331	0.372	0.041	61.4	243.4	0.081	n/a	n/a	n/a	n/a
B2 Residual calibration	0.318	0.334	0.016	61.2	240.9	0.043	n/a	n/a	n/a	n/a
B3 PEARL	0.271	0.278	0.007	61.1	238.2	**0.018**	n/a	n/a	n/a	n/a
B4 Task-cost exploration	0.325	0.381	0.056	61.3	245.5	0.126	n/a	n/a	n/a	n/a
B5 ASID	0.344	0.412	0.068	61.4	247.0	0.149	n/a	n/a	n/a	n/a
B6 MAP-Elites / M-BOA	0.307	0.359	0.052	61.0	242.7	0.171	n/a	n/a	n/a	n/a
B7 Domain-randomised policy	0.311	0.316	0.005	0.0	178.4	0.022	n/a	n/a	n/a	n/a
B8 Oracle signature	0.561	0.566	0.005	0.0	171.2	0.003	n/a	n/a	n/a	n/a
B9 No adaptation	0.196	0.199	0.003	0.0	174.6	0.031	n/a	n/a	n/a	n/a
B10 SafeOpt-MC	0.336	0.398	0.062	88.7	271.7	0.052	0.093	0.427	0.038	0.058
SOMA-Probe (ours)	**0.446**	**0.502**	0.056	61.4	276.3	0.068	0.184	**0.610**	**0.126**	0.021

Table 6: Uncertainty on the primary endpoint. Paired differences are each method's end-to-end success minus SOMA-Probe's on the same (perturbation, task, episode) triples. The 95% BCa intervals use 2,000 bootstrap resamples clustered by perturbation. The final column adjusts the pre-registered family {B1, B5, ΔU at $\rho = 10$} by Holm's method. For the eight comparators outside this family, the table gives raw p values for descriptive comparison and no adjusted significance claim. The utility contrast has raw and Holm-adjusted $p = 0.0089$.

Method	E2E	Paired diff.	95% cluster CI	p	Holm
B1 RMA	0.331	-0.115	$[-0.161, -0.070]$	0.0010	0.0030
B2 Residual calibration	0.318	-0.128	$[-0.174, -0.083]$	0.0003	–
B3 PEARL	0.271	-0.175	$[-0.223, -0.128]$	0.0001	–
B4 Task-cost exploration	0.325	-0.121	$[-0.168, -0.075]$	0.0004	–
B5 ASID	0.344	-0.102	$[-0.149, -0.056]$	0.0026	0.0052
B6 MAP-Elites / M-BOA	0.307	-0.139	$[-0.187, -0.092]$	0.0002	–
B7 Domain-randomised policy	0.311	-0.135	$[-0.181, -0.089]$	0.0002	–
B8 Oracle signature	0.561	$+0.115$	$[+0.069, +0.162]$	0.0006	–
B9 No adaptation	0.196	-0.250	$[-0.298, -0.202]$	0.0001	–
B10 SafeOpt-MC	0.336	-0.110	$[-0.157, -0.064]$	0.0005	–

Table 7: Method inputs, training and outputs. "Budget" gives the source of the interaction allowance: *producer* sets the per-episode allowance in pass one, *pass 1* reads that allowance, and *own* uses a separate budget. The coverage and utility columns in Table 1 require three-way verdicts.

	Probe library	Budget	Output arity	Privileged information	Training
AB0	shared	producer	3 (CF/UNC/CI)	none	none
B1	free interaction	pass 1	1 (execute)	sim params in stage 1	PPO + regression
B2	free interaction	pass 1	1 (execute)	none	GP fit online
B3	free interaction	pass 1	1 (execute)	none	variational, shared trunk
B4	shared	pass 1	1 (execute)	offline task Hessian	none
B5	shared	pass 1	1 (execute)	parameter names	none
B6	shared	pass 1	1 (execute)	none	offline archive
B7	none	pass 1	1 (execute)	none	PPO under wide DR
B8	none	own	2 (execute/refuse)	$\theta^\star$ table	none
B9	none	pass 1	1 (execute)	none	PPO, nominal
B10	shared	own	3 (CF/UNC/CI)	assumed L	none

Table 8: Removing one witness at a time (AB4). Each row partitions infeasible instances into the three recorded witness states and sums to one. The last column subtracts the witness's triggered share from $\mathrm{Cov}^{\mathrm{inf}}_{\mathcal{W}} = 0.412$, giving a lower bound on coverage after removal. Witnesses can overlap, and a triggered bound can remain above c_{inf}, so the triggered share can exceed the stacked certification contribution in Figure 4. Conclusions exclude any evaluated (witness, task) cell with more than 20% NOT-EVALUATED instances. The two load tasks receive no W_3 steps by design and are excluded from that witness's evaluation. The evaluated reach and clearance cells satisfy the threshold; the pooled W_3 NOT-EVALUATED share is therefore close to one half.

Witness removed	Triggered	Not-triggered	Not-evaluated	$\mathrm{Cov}^{\mathrm{inf}}_{\mathcal{W}}$ left
W_1	0.148	0.771	0.081	0.264
W_2	0.163	0.749	0.088	0.249
W_3	0.098	0.395	0.507	0.314
W_4	0.089	0.822	0.089	0.323

Table 9: Multi-axis probing at a fixed budget (AB5), by task. On T3, clearance angle and insertion depth compete as the binding axis at the widest slot; probing the top two axes increases end-to-end success by 0.026. On the other three tasks, $u^\star$ is unambiguous and distributing the budget across more axes reduces success.

Task	Top-1 axis	Top-2 axes	Top-3 axes
T1	**0.508**	0.491	0.472
T2	**0.463**	0.449	0.431
T3	0.421	**0.447**	0.428
T4	**0.447**	0.436	0.418

Table 10: Validity of the optional regularity branch (G5). The branch extrapolates beyond A_{safe} under local concavity and is enabled only on tiers where that condition holds for at least 60% of perturbations. It remains closed on D5a, D5b and D5d. Forcing it open on D5a and D5b produces relative extrapolation errors of 0.63 and 0.48, compared with at most 0.17 on D1–D4.

Tier	(R) hold rate	Rel. extrapolation error	(U2) channel
D1	0.784	0.09	OPEN
D2	0.731	0.12	OPEN
D3	0.802	0.08	OPEN
D4	0.646	0.17	OPEN
D5a	0.187	0.63	CLOSED
D5b	0.241	0.48	CLOSED
D5c	0.612	0.21	OPEN
D5d	0.573	0.29	CLOSED

Table 11: Risk-weight sweep (AB7). Removing the risk penalty increases coverage and aborts, while reducing end-to-end success. An aborted episode receives zero under the end-to-end metric. The operating value of λ is selected using that metric.

λ	**E2E** $\uparrow$	Cov $\uparrow$	MCR $\downarrow$	Risk $\downarrow$
0	0.371	0.622	0.014	0.196
0.5	0.418	0.618	0.013	0.131
2 (main)	**0.446**	0.610	0.011	0.068
8	0.427	0.564	0.010	0.041
32	0.389	0.487	0.009	0.023

Table 12: Compute by activity. Per-episode estimates are 0.10 CPU core-hours for SOMA-Probe, 0.55 for ASID, 0.35 for MAP-Elites, 0.18 for SafeOpt-MC, 0.095 for a privileged ground-truth scan, 0.042 for a fallback-only episode and 0.075 for the remaining methods. The method-specific estimates include finite-difference identification, archive use and posterior updates where applicable.

Line	GPU-h	CPU core-h
Pass 1, AB0 over five ζ levels	0	1 400
Pass 2, nine baselines at $\zeta = 0.70$	0	2 400
B10 SafeOpt-MC, own budget and L sweep	0	2 500
Policy pre-training (PPO, RMA stage two, PEARL)	169	16 850
B5 identification and B6 archive	2	15 250
Privileged ground-truth scans, development and blind	0	600
Ablations AB1–AB9	46	22 200
Conservative-route episodes	0	200
Pre-checks, calibration and audits	12	2 000
Total	**229**	**63 400**

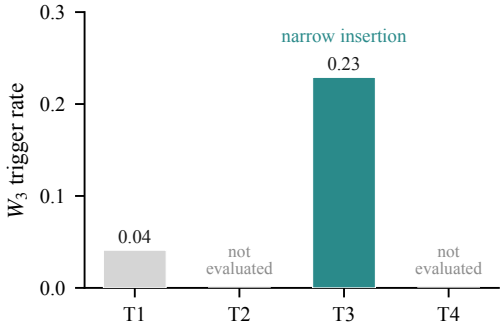


Figure 5: Task-specific trigger rates for W_3. The load tasks T2 and T4 receive no W_3 steps and are recorded as NOT-EVALUATED. Bars use all instances of each task as their denominator; Table 8 conditions on infeasible instances.

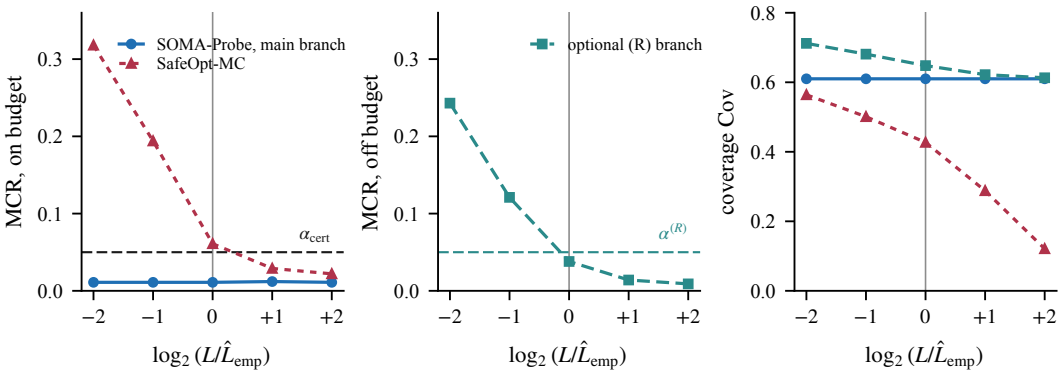


Figure 6: Sensitivity to the assumed Lipschitz constant. Each panel scans L from $0.25\hat{L}_{\text{emp}}$ to $4\hat{L}_{\text{emp}}$, where the empirical reference is measured beyond the envelope in AB6. Left: main-branch and SafeOpt-MC mis-certification rates relative to α_{cert}. The main branch does not use L; SafeOpt-MC exceeds the level when L is underestimated. Middle: the optional (R) branch is assessed against its separate $\alpha^{(R)}$ budget and is not pooled with the main branch. Right: an overestimated L also reduces SafeOpt-MC coverage by making its bounds more conservative. In-envelope observations provide a lower bound on L, whereas extrapolation requires an upper bound.

Statistical model. Let $d_{ipte} \in \{-1, 0, +1\}$ be the paired success difference between our method and a comparator on perturbation p, task t and episode e. The analysis model is $d \sim \beta_0 + \beta_1 \,\mathrm{par}_p + b_p + b_t + \varepsilon$, where par_p indicates that perturbation p lies within the fitted parameter vector, with $b_p \sim \mathcal{N}(0, \sigma_p^2)$ and $b_t \sim \mathcal{N}(0, \sigma_t^2)$ independent. The interaction test compares the model containing β_1 with the intercept-only model to assess whether the advantage differs between D1–D4 and D5. BCa intervals use 2000 bootstrap replicates that resample perturbations, preserving the dependence among episodes sharing a body. Three contrasts, against RMA, against ASID and the ΔU test at $\rho = 10$, form the Holm family at 0.05. The pre-registered protocol assigns other inferential comparisons to Benjamini–Hochberg control at $q = 0.10$. Table 6 displays only raw values for comparisons outside the Holm family; these are treated descriptively. With 3600 paired differences, the planned minimum detectable effect is 0.091 at 80% power. Perturbations are the independent units for this calculation.

The asymmetry gap for G1 is $\mathrm{Cov}^{\mathrm{fea}}_{\mathrm{feas}} - \mathrm{Cov}^{\mathrm{inf}}_{\mathcal{W}} = 0.458$, with bootstrap interval $[0.391, 0.522]$. The D5 paired difference over ASID for G3 is 0.207, with interval $[0.152, 0.263]$; the interaction likelihood-ratio test gives $p = 0.0021$ in its own family. Both meet their pre-specified criteria. Benchmark retention requires ground-truth-verdict agreement of at least 0.80 for the privileged oracle and B9 success of at most 0.30. Across retained D5 subclasses, oracle agreement is at least 0.861 and B9 success is at most 0.141. Levels failing either condition are excluded before the method comparison.

The utility matrix. Six outcomes are scored, with task value normalised to 1 and damage cost ρ swept over $\{3, 10, 30\}$. The feasible and infeasible rows assign the same value to each outcome. The two-way comparator uses the same matrix and merges UNCERTIFIED into CERTIFIED-INFEASIBLE.

Outcome	Utility	Outcome	Utility
Execute, task succeeds	$+1.0$	Conservative route succeeds	$+0.6$
Execute, safety criterion tripped	$-\rho$	Conservative route fails	-0.4
Execute, task fails, no trip	-0.3	Refuse	-0.2

The utility matrix values a successful conservative route at $+0.6$ and refusal at -0.2, a difference of 0.8, so it rewards completion of a reduced task over declining it. A failed conservative route receives -0.4. These values were fixed in the pre-registration. G2 measures completion of the original task and scores a successful conservative route as 0; G7 assigns that route $+0.6$. The two endpoints therefore evaluate different benefits of the three-way policy.

Table 13: Pre-registered falsification criteria. G6 is descriptive and excluded from all multiple-comparison families. Its reported value is $R - \mathrm{Cov}$ from Figure 3; interpreting this value as a gap to Proposition B.5 requires verifying the reference rates against the proposition's events.

	Claim	Passes only if
G1	The $(\mathrm{Cov}, \mathrm{Risk})$ frontier is non-trivial and asymmetric	Cov spans ≥ 0.25 over the middle three ζ; $\mathrm{Cov}^{\mathrm{inf}}_{\mathcal{W}} \geq 0.15$; $\mathrm{Cov}^{\mathrm{fea}}_{\mathrm{feas}} - \mathrm{Cov}^{\mathrm{inf}}_{\mathcal{W}} \geq 0.25$ with bootstrap lower bound above 0
G2	At matched probe budget, higher probe-risk-inclusive success than B1 and B5	paired difference > 0, Holm $p < 0.05$
G3	The advantage is concentrated on D5	interaction $p < 0.05$ *and* D5 paired difference ≥ 0.15
G4	Witnesses certify infeasible cases at the specified error level	$\mathrm{Cov}^{\mathrm{inf}}_{\mathcal{W}} \geq 0.15$ and MCR upper bound ≤ 0.05
G5	The optional regularity branch meets its validity threshold	enabled per tier only where the assumption holds on $\geq 60\%$ of perturbations
G6	Gap to the Proposition B.5 ceiling	descriptive only; no threshold
G7	The three-way policy has higher utility than the two-way policy	$\Delta U > 0$ at $\rho = 10$ with bootstrap lower bound above 0, Holm $p < 0.05$

Perturbations. Sixty perturbations across five tiers: D1 joint damage (10), D2 torque decay (10), D3 payload and centre-of-mass shift (10), D4 morphology (12), D5 outside the identified parame-

Table 14: The four tasks. Success is scored on the executed task, not the probe. Fallback R1 changes to an energy-saving configuration and does not read L_T; R2 derates the demand once and R3 splits it into $n_{\text{batch}} = \lceil \kappa_{\text{cons}} \hat{h}^\star / L_T \rceil \leq 4$ batches, both of which do.

	Task	**Axis $u^\star$**	**Demand $\hat{h}^\star$**	**Bottleneck (11 levels)**	**Fallbacks**
T1	High reach	reach along d	shelf height projection	$z_{\text{shelf}} \in [0.45, 0.95]$ m	R1
T2	Heavy transport	lift force	mg	$m \in [0.5, 6.0]$ kg	R1→R2→R3
T3	Narrow insertion	insertion depth	required depth	$d_{\text{ins}} \in [20, 70]$ mm	R1
T4	Push heavy object	push force	$\mu_{\text{gnd}}\, mg$	$\mu_{\text{gnd}} \in [0.2, 1.0]$	R1→R2→R3

Table 15: Estimation-target ablation by tier (AB3), using the same probe data. Rows compare signature, parameter and latent targets; columns give end-to-end success by perturbation tier. The kinematic-calibration variant is evaluated on D5c, where kinematic parameters can represent the bent link.

Estimation target	D1	D2	D3	D4	D5a	D5b	D5c	D5d
Capability signature (ours)	0.492	0.471	0.508	0.463	0.421	0.408	0.436	0.447
Explicit physical parameters	0.486	0.464	0.499	0.455	0.238	0.196	0.271	0.318
+ kinematic calibration	–	–	–	–	–	–	0.418	–
Unstructured latent	0.471	0.448	0.483	0.436	0.211	0.174	0.243	0.287

ter vector (18: six D5a load-dependent transfer, six D5b Bouc–Wen hysteresis (Bouc, 1968; Wen, 1976), three D5c permanent 3–8 mm link bend, three D5d anisotropic friction). Each carries two labels fixed before inspecting results: membership in the fitted parameter space and whether an equivalent rigid body can represent the perturbation. D5c lies outside the registered inertial parameter vector but can be represented through kinematic calibration. Table 15 reports that variant separately. The pooled D5 comparison includes all eighteen perturbations, including the three in D5c. The 24/36 development/blind split is enforced by sealing the blind realisations and scanning logs to assert no blind identifier was read before the unsealing date.

Baseline catalogue. We compare ten baselines. B1, RMA (Kumar et al., 2021), and B3, PEARL (Rakelly et al., 2019), adapt passively; B2 calibrates from residuals following Jo et al. (2026). The active baselines are B5, ASID (Memmel et al., 2024), which identifies a parameter vector with 60 to 80 dimensions; B4, task-cost-weighted exploration (Wagenmaker et al., 2023); and B6, MAP-Elites (Cully et al., 2015), which searches behaviours. A kinematic-calibration variant of B5 tests D5c. B10, SafeOpt-MC (Berkenkamp et al., 2023), shares our $J = 24$ amplitude grid and uses its confidence bounds to produce three-way verdicts. B7 is a domain-randomised policy, B8 receives the ground-truth signature, and B9 uses the nominal model without adaptation. Table 7 specifies their information access and outputs.

Baseline configuration. RMA receives a latent-dimension sweep over $\{1, 2, 4, 8, 16\}$, with the best result reported per tier, and an augmented variant with free exploratory interaction matched to SOMA-Probe's probe allowance. ASID's pre-registered parameter vector is $\{m, c_{xyz}, I_{\text{diag}}, \tau^{\max}, \mu, k_{\text{joint}}\}$, with 60 to 80 dimensions, identified using MuJoCo's system-identification utilities. Its kinematic-calibration variant evaluates the geometric change in D5c. The pre-registered hypothesis orders the difficulty of fitting the D5 subclasses as D5b > D5a > D5d > D5c. Our implementation of task-cost exploration approximates the model-task Hessian by finite differences at the decision boundary. SafeOpt-MC uses the shared amplitude grid and produces three-way verdicts from its own confidence bounds.

Compute. Our per-episode cost is estimated at 0.10 CPU core-hours, based on 8 calibration steps plus at most 96 probing iterations, totalling at most 104 steps at 0.45 to 1.3 s per probing step, at most 60 s of task execution, 40 s for rollback, logging and model compilation, and a safety factor of 1.5. Per-episode costs are estimated separately for each method: ASID uses 0.55 core-hours for finite-difference identification in 60 to 80 dimensions, SafeOpt-MC uses 0.18 for $O(t^3)$ posterior

updates, and methods without either operation use 0.075. The complete study uses 229 GPU-hours and $63{,}400$ CPU core-hours, including $15{,}000$ CPU core-hours for PEARL training. Table 12 gives the breakdown.

F AUDITS

We check confidence-sequence calibration, anchor isolation, budget matching and witness-state logging. Under injected amplitude-dependent noise $\varsigma(\alpha)$, the one-sided Clopper–Pearson upper bound on the empirical false-feasibility rate $\mathbb{P}[\theta^\star < L_T]$ remains below α_{cert} across the tested range. Each episode checks anchor residual displacement against ϵ_Π. The second pass must read the allowance written for the same episode in the first pass; otherwise the episode does not run. Every witness records one of its three states, and the quota table determines NOT-EVALUATED before the detector outcome is observed.

We also recompute dependent constants and check their consistency: η_1, the minimum repeat count and boundary choice; the quota sums, detector minima and $T_{\max}$; $a_{\max} > A_{\mathrm{safe}}(\zeta)$ at all five levels; and per-episode steps, costs and study totals. Certification-channel readings and $\hat{h}^\star$ must share the same $(\mathrm{unit}, \mathrm{origin}, \mathrm{frame})$ triple, with $y(0) = 0$ at the registered origin. The origin and frame checks also detect comparisons of relative displacement with absolute distance, which would invalidate the four ratio-valued criteria.

Execution relaxes the quasi-static assumption (A1), so the recorded frequency of "truly infeasible and executed successfully" can be non-zero. We assess that frequency against the pre-registered 2% ceiling.

Thresholds, budgets, significance splits, the analysis model and the falsification criteria of Table 13 were frozen before any blind perturbation was realised. Development used the 24 development perturbations only; the 36 blind realisations were unsealed on a pre-declared date, after the analysis was fixed, and a scan of the run logs confirms no blind identifier was read beforehand. Every episode writes a record carrying the ground-truth signature, the three-way decision, the per-witness states, the dual-trigger flag, the safety-abort stage and the observable-registry hash.

G EXTENDED RELATED WORK

SOMA-Probe combines active excitation for embodiment information (Memmel et al., 2024; Sobanbabu et al., 2025), information-based experimental design (Hübotter et al., 2024b), and sequential testing (Locatelli et al., 2016; Wald, 1945; Howard et al., 2021). These tools determine how to gather observations and when to stop. Our analysis determines which capability verdicts those observations can support under a safety envelope. It connects the limits of safe exploration (Turchetta et al., 2016) to abstention at a declared error rate (Ren et al., 2023) through the distinction between certified feasibility, certified infeasibility, and an unresolved decision.

G.1 ACTIVE IDENTIFICATION

ASID (Memmel et al., 2024), SPI-Active (Sobanbabu et al., 2025) and Bohlinger & Peters (2026) use deliberate excitation for body identification. Their parameter estimates support generalisation across tasks when the body lies within the assumed model family. For perturbations outside that family, a rigid-body fit can still be returned, but its parameters cannot represent the perturbation mechanism. SOMA-Probe defines its target directly as task capability, which remains well-defined for these perturbations. The hard amplitude truncation considered here introduces a further restriction on certification, characterised by Proposition 3.1. Following Wagenmaker et al. (2023), we weight identification by its downstream use. Their method selects directions in parameter space; ours selects probe amplitudes for a capability decision.

G.2 PASSIVE AND REACTIVE ADAPTATION

RMA's low-dimensional sufficient statistic (Kumar et al., 2021) and PEARL's probabilistic context (Rakelly et al., 2019) infer a body summary from interaction. PEARL represents task uncertainty

explicitly and uses its posterior to guide exploration and conditioning. The meta-training distribution determines the range of bodies on which this inference is trained. A chosen model parameterization may omit mechanisms such as a bent link or a hysteretic gripper (Tobin et al., 2017). J-PARC (Jo et al., 2026) infers a latent joint-fault regime from degraded execution and then corrects it. It also distinguishes performance loss from infeasibility: increased friction can make a task harder while leaving it feasible. These methods adapt behaviour from interaction; SOMA-Probe supplies a capability verdict before task execution.

G.3 Safe Bayesian optimisation

SafeOpt (Sui et al., 2015) and SafeOpt-MC (Berkenkamp et al., 2023) expand a safe set under a GP prior and use known RKHS-norm or Lipschitz bounds to reason about unvisited inputs. SafeOpt-MC's decoupled constraints closely match our envelope (1). SafeOpt guarantees convergence to an optimum reachable within the safe set. Proposition 3.1 characterises how a safe observation domain restricts capability certification; the resulting coverage is measured by $\mathrm{Cov}(A_{\mathrm{safe}})$.

Fiedler et al. (2024) document that available frequentist bounds (Srinivas et al., 2010; Chowdhury & Gopalan, 2017) can prevent exploration or require an RKHS-norm bound unavailable from realistic prior knowledge; they propose an assumed Lipschitz constant as an alternative. Slip responses can make such regularity bounds difficult to specify. A discontinuity at the slip load violates every finite Lipschitz bound. Our main branch uses the stated observation assumptions and, for infeasibility certificates, the order property associated with the triggered witness. Without such a witness, U_T retains its trivial value, and the method returns Uncertified if the lower bound also fails to certify feasibility. The guarantee $\mathrm{MCR} \leq \alpha_{\mathrm{cert}}$ holds under these assumptions. The optional regularity branch requires a Lipschitz bound and has a separate significance budget.

G.4 Set-membership identification and guaranteed capability sets

Set-membership identification returns a guaranteed set of admissible models (Belforte et al., 1990; Vicino & Zappa, 1996; Milanese & Novara, 2004). Milanese & Novara (2004) use gradient bounds without prescribing a functional form. A global gradient bound supplies Lipschitz regularity and requires prior information about the response. Pickard et al. (2021) compute inner approximations of robot capability sets by interval analysis using a known model with interval-valued parameters. These methods support capability guarantees when the required model and bounds are available. SOMA-Probe defines a capability target for perturbations outside the chosen parameterization and studies which verdicts safe observations can certify.

G.5 Sequential testing under a threshold

Our stopping rule and its sample complexity apply established sequential-testing results: Chen et al. (2014) for the fixed-confidence bound, Locatelli et al. (2016) for the fixed-budget error exponent at a threshold, and Wald (1945) for sequential testing. The confidence sequences of Howard et al. (2021), with the variance-adaptive refinement of Waudby-Smith & Ramdas (2024), allow stopping at an arbitrary time without inflating α_{cert}. Proposition B.11 applies these results to capability certification. Anytime validity controls error under adaptive stopping. The additional identification question arises because observations are restricted to amplitudes below A_{safe}, while the target supremum ranges beyond that envelope. The decision-focused precision criterion follows the identification-for-control perspective (Gevers, 2005).

G.6 Abstention and sampling domains that exclude the target

KnowNo (Ren et al., 2023) combines abstention with a conformal guarantee at a user-specified error rate. We adopt its use of error-controlled abstention and study physical capability uncertainty. KnowNo calibrates semantic uncertainty over instruction-scene pairs; in our experiment, a worn and a healthy gearbox produce the same prediction set. It outputs execute or ask; our three verdicts separate Certified-Feasible, Certified-Infeasible, and Uncertified.

Semantic refusal and physical capability certification require different evidence. On $6{,}069$ instructions that should be refused, Yeke et al. (2026) report refusal rates of 39.0% for their best model and

16.5% for an embodied planner under direct prompting; revised prompts raise refusal above 88%. These results concern instruction-level decisions. Our question requires physical observations of the robot's current capabilities, such as the load a worn gearbox can sustain.

Transductive active learning (Hübotter et al., 2024b;a) distinguishes a sampling domain S from a target domain $A \not\subseteq S$. Here these domains are $[0, A_{\text{safe}}]$ and $[0, a_{\max}]$, and its action-selection rule can be used for probe selection. GP or RKHS regularity couples observations across the domain gap. Under our baseline assumptions, the upper endpoint for θ remains at the trivial bound $\widetilde{B}$, even after observing the entire safe domain. Transductive active learning targets point values on A; Proposition 3.1(a) concerns their supremum over the full domain. Turchetta et al. (2016) establish an algorithm-independent ceiling on the safely reachable state set, using a known Lipschitz constant. Proposition B.5 bounds coverage for policies using the prescribed feasible-certificate rule and witness family. It uses the instances on which these rules can issue a certificate under valid confidence bounds. For a fixed instance distribution, budget and certificate family, it gives a coverage bound as a function of A_{safe}, without requiring an exact identification endpoint. The interval $[L_T, U_T]$ is the image of a version space (Mitchell, 1982) under θ, placing this work in the partial identification tradition (Manski, 2003; Tamer, 2010).