

Agent Behavioral Contracts II:

Certifying Compositional Reliability Without Assuming Independence

Varun Pratap Bhardwaj
Qualixar / Independent Researcher, India
varun.pratap.bhardwaj@gmail.com
ORCID: 0009-0002-8726-4289

Garima Singh
Independent Researcher, India
garima1213@gmail.com

Arun Pratap Bhardwaj
Independent Researcher, India
arun.pratap.bhardwaj@gmail.com

Abstract

Compositional reliability bounds for multi-agent systems multiply component reliabilities, a step licensed by a conditional-independence assumption that is routinely stated and rarely tested. We test it. Two instances of one model, composed in a two-agent handoff, co-fail on 90.0% of the missions on which either fails (log OR = 6.66, 95% CI [6.38, 7.00]; $\phi = 0.916$). The evidence is a preregistered confirmatory evaluation of 18,000 missions with deterministic scoring and no model in the judging loop, inside a larger campaign whose other topologies we report as secondary. Substituting a different model reduces the association significantly in the confirmatory motif and in both secondary topologies (six of six contrasts); substituting a different *vendor*, with the model already different, does not — a registered hypothesis that fails to replicate and that we report as a null. An unmanipulated same-model pair, present in every arm, returns fifteen of fifteen null contrasts: the result a design free of graph-wide confounding would produce, and one that would have been informative had it come out otherwise.

The error is signed and runs against the operator: positive dependence inflates joint failure above the independence product, so redundancy is over-credited exactly when components share a model. The assumption-free alternative is often vacuous — the certified floor is zero whenever mean component reliability falls below $1 - 1/m$ — and fitting a dependence model is worse: we prove that a bootstrap bound on a fitted model’s functional loses coverage of the true reliability as $n \rightarrow \infty$, because the identification gap is $O(1)$ while the bootstrap haircut is $O(n^{-1/2})$. More data makes such a certificate worse, with no visible symptom.

We give a finite-sample certificate that assumes no dependence structure: a linear program over the joint, taken over a Bonferroni–Clopper–Pearson box around measured co-execution moments. It is sound, sharp for the information supplied, and monotone in the moment family under a stated Bonferroni allocation. On four-stage data, enriching from ten moment functionals to fourteen narrows the identified interval by 85.7% and lifts the certified floor from 0.2455 to 0.4116. A companion anytime-valid certificate holds its empirical type-I error at 0.0471 or below across every admissible betting fraction, recovering the SPRT exactly at the optimal bet. An ablation at a conceded design effect 1.31 times the largest measured moves the floor by at most 2.69 percentage points.

We also show that the dependence statistics in common use — Jaccard, ϕ , Kendall’s τ_a — are bounded by the marginals and can significantly *reverse* an apparent ordering of conditions when the compared agents fail at different rates, a reversal we observe and then replicate on two further inference backends. The contracts, the mission generators, the scoring code, the analysis scripts, and the preregistration are released; every reported statistic is regenerated by those scripts rather than transcribed.

1 Introduction

1.1 The independence gap

A multi-agent pipeline is certified the way a series system is certified: bound each component’s reliability, multiply, report the product. The v1 Agent Behavioral Contract framework (Bhardwaj, 2026) does this, and so does every compositional reliability argument we are aware of for agent systems. The step is licensed by a conditional-independence condition — C5 in Definition 3.10 — which asserts that a downstream agent’s contract satisfaction is independent of an upstream agent’s internal execution given a compliant handoff.

C5 is stated in v1 and never tested. It is also, on reflection, implausible exactly where multi-agent systems are most often deployed: a reviewer agent checking a writer agent is very often the same model with a different prompt. Two instances of one model do not have independent blind spots. They share them.

We measure how much this matters. The preregistered confirmatory evaluation is a two-agent handoff over 18,000 missions with deterministic scoring and no model in the judging loop; the registration places the other topologies of the campaign in secondary work, and we report them as such. Two instances of `mistral-small-24b` co-fail on 90.0% of the missions on which either fails, with $\log \text{OR} = 6.66$ (95% CI [6.38, 7.00]). Replacing the second agent with a different model reduces the association significantly, in the confirmatory motif and in both secondary topologies. Replacing the *vendor*, with the model already different, does not.

The consequence runs against the operator. By Proposition 4.1 the compositional gap is signed: under positive dependence, joint failure exceeds the independence product, so redundancy is over-credited precisely when the redundant components share a model. A dashboard that multiplies reliabilities reports a reassuring number whose governing assumption the data reject, and nothing in the pipeline signals it.

1.2 Why the obvious repairs fail

Dropping the assumption gives the Fréchet–Hoeffding bounds, which are sharp and frequently vacuous: by Corollary 4.2 the certified floor is exactly zero whenever mean component reliability falls below $1 - 1/m$, which for four components at $p = 0.75$ is the entire regime of interest.

Fitting a dependence model — a Gaussian one-factor copula, the standard choice in reliability practice — is worse than either. Theorem 4.2 proves that a bootstrap lower bound on the fitted model’s functional loses coverage of the *true* reliability as $n \rightarrow \infty$: the identification gap is $O(1)$ while the bootstrap haircut shrinks like $n^{-1/2}$, so past some finite sample size the interval sits entirely above the truth and never returns. Collecting more data does not repair such a certificate; it narrows the interval around the wrong target, and nothing in the interval signals it.

1.3 Approach

We constrain the joint law with measured co-execution moments and optimise over everything consistent with them. The resulting certificate is a linear program over the 2^m cells of the joint, taken over a Bonferroni–Clopper–Pearson box around the empirical moments. It is sound with no dependence assumption (Theorem 5.2), sharp for the information supplied (Theorem 6.1), and tightens monotonically as the moment family grows (Propositions 6.1 and 6.2). On real four-stage data, enriching from ten moment functionals to fourteen narrows the identified interval by 85.7% and lifts the certified floor from 0.2455 to 0.4116.

Because deployed systems are monitored continuously rather than at a sample size fixed in advance, we add a certificate valid under arbitrary stopping (Theorem 7.1). Its null constrains only a conditional mean, so it requires no independence assumption at all — it is immune to the failure the rest of the paper documents.

1.4 Contributions

C1 (co-lead)

A preregistered evaluation that *manipulates* model sharing at three levels, confirmatory over 18,000 two-agent-handoff missions and replicated in two secondary topologies within a 30,820-mission campaign, converting an observation about particular systems into a contrast attributable to substitution. Correlated failure among same-backbone agents is itself established by concurrent work (Section 2.3); the controlled design, and the finding that the ordering holds at the model level and vanishes at the vendor level, are ours.

C2 (lead)

A finite-sample, copula-agnostic reliability certificate for composed agent pipelines (Theorem 5.2), with a Bonferroni allocation that makes the floor monotone in the moment family by construction.

C3 (supporting)

Theorem 4.2: coverage of a model-based reliability floor tends to zero as the sample grows. The phenomenon is misspecification bias; what we add is the explicit witness of Example 4.2, indistinguishable from the fitted family on the moments the model uses.

C4 (co-lead)

Anytime-valid certification of graph reliability, with the SPRT recovered exactly at the optimal bet (Proposition 7.1).

C5 (supporting)

A demonstration that marginal-sensitive dependence statistics can significantly *reverse* an apparent condition ordering, replicated on independent backends (Sections 10.2.3 and 10.6).

C6 (supporting)

An internal negative control returning fifteen of fifteen nulls, and a full artifact release including the preregistration.

Section 12 states what we do *not* claim, including one registered hypothesis that fails to replicate.

1.5 Organisation

Section 2 places the work against contract-based specification, guardrails, the recent correlated-failure literature, and dependence-free bounding. Section 3 states the ABC framework in full, including formulas withheld from v1 under a patent claim now withdrawn. Section 4 establishes what breaks without C5. Sections 5 to 7 construct the certificate. Sections 8 and 9 describe runtime enforcement and the implementation. Section 10 reports E1–E6. Section 11 covers limitations and threats to validity, and Sections 12 to 14 close.

2 Background and Related Work

2.1 Contracts for software and for agents

Design by Contract (Meyer, 1992; Hoare, 1969) specifies preconditions, postconditions, and invariants that a component must honour, and makes composition tractable by letting each component’s postcondition discharge the next one’s precondition. The formal-methods lineage that followed (Clarke et al., 1999; Lamport, 2002; Leino, 2010; Barnett et al., 2004) verifies such specifications statically, static analysis bounds behaviour by abstraction (Cousot and Cousot, 1977), and dynamic-invariant work (Ernst et al., 2007) infers them from traces.

None of it transfers directly to agents driven by natural-language instructions, because the specification surface is prompts rather than types and the execution is stochastic. The v1

Table 1: Capability comparison against representative approaches. ✓ = supported, ◦ = partial, — = not addressed. **RE** runtime enforcement · **FS** formal specification · **MA** multi-agent scope · **DM** dependence measured · **CB** composed reliability bound · **CA** copula-agnostic · **AV** anytime-valid. The last three are the contribution of this paper; **DM** is where concurrent work has recently arrived.

	RE	FS	MA	DM	CB	CA	AV
Training-time alignment (Bai et al., 2022; Ouyang et al., 2022)	—	—	—	—	—	—	—
Output guardrails (Rebedea et al., 2023)	✓	◦	—	—	—	—	—
Static verification (Clarke et al., 1999; Leino, 2010)	—	✓	◦	—	◦	—	—
Agent observability	—	—	✓	—	—	—	—
Multi-agent debate (Du et al., 2024; Liang et al., 2024)	—	—	✓	—	—	—	—
Byzantine-tolerance view (Zheng et al., 2025; Berdoz et al., 2026)	—	◦	✓	◦	—	—	—
Failure attribution (Rafi et al., 2026; Qiao et al., 2026)	—	—	✓	◦	—	—	—
Correlated-failure detection (McDonnell et al., 2026)	—	—	✓	✓	—	—	—
Multi-agent calibration (Huang et al., 2026)	—	—	✓	✓	—	—	—
ABC v1 (Bhardwaj, 2026)	✓	✓	◦	—	◦	—	—
This work	✓	✓	✓	✓	✓	✓	✓

framework (Bhardwaj, 2026) closes that gap by making preconditions, hard and soft invariants, governance policies, and recovery mechanisms first-class and runtime-enforceable, and by replacing deterministic satisfaction with the probabilistic (p, δ, k) -notion of Definition 3.6. We take that framework as given and measure what happens when several contracted agents are composed.

2.2 Steering, filtering, and guarding

Three families of technique constrain agent behavior, and none of them certifies a composed system.

Training-time alignment — Constitutional AI (Bai et al., 2022), RLHF (Ouyang et al., 2022) — shapes general tendencies but cannot encode a deployment-specific invariant, and offers no runtime guarantee. Output guardrails (Rebedea et al., 2023) filter or redirect responses matching prohibited patterns at inference time; the current generation adds programmable policy languages and validator libraries, and operates per-response. Observability platforms trace and score agent runs after the fact.

The common limitation is scope: each acts on a single turn or a single agent. None specifies invariants over a multi-agent pipeline, and none produces a statement of the form “this composition satisfies its contract with probability at least $\hat{L}$.” Table 1 makes the comparison explicit.

2.3 Correlated failure in multi-agent systems

That agents sharing a base model may fail together is no longer a conjecture, and we state plainly what is already established so that our own contribution is not overstated.

McDonnell et al. (2026) study a three-agent triage architecture and quantify correlated failure directly, reporting a joint error rate inflated $3.53\times$ over independence (BCa 95% CI [3.50, 3.59]) with $\phi = 0.612$, and finding 57.2% of errors occurring under agent agreement. Their agents are classical learners — a random forest, a k -nearest-neighbour model, and a calibrated meta-model — over intrusion-detection and clinical-readmission data, and model sharing is not a manipulated variable; the design varies learner *family* by construction rather than varying it experimentally. Huang et al. (2026) recalibrate multi-agent confidence against a counterfactual no-communication baseline and observe, qualitatively, that within-family blocks of agents sharing

a backbone exhibit shared blind spots. The Bengio et al. (2026) report lists correlated failure among same-base-model agents as a known concern, as does the survey of Hammond et al. (2025); work on governed capability evolution (Qin et al., 2026) addresses the related problem of keeping a guarantee valid as components are upgraded. Work on failure attribution (Rafi et al., 2026; Qiao et al., 2026) traces which agent caused a trajectory to fail, and Byzantine-tolerance analyses (Zheng et al., 2025; Berdoz et al., 2026) study whether agents can reach agreement at all under faults.

Three things separate the present work from that literature. First, model sharing here is a *manipulated* experimental variable — three sharing levels, preregistered, with a confirmatory core of 18,000 missions on the two-agent handoff motif and two further topologies reported as secondary replication — rather than a property of a fixed architecture, which is what makes the arm contrasts of Section 10.2 causal claims about substitution rather than observations about a particular ensemble. Second, none of this work produces a *bound*: measuring that failures correlate does not tell a practitioner what reliability they may certify, and Sections 5 and 6 supply exactly that. Third, the statistic matters. McDonnell et al. (2026) lead with ϕ , which Definition 3.17 classifies as marginal-sensitive; Section 10.2.3 shows that such statistics can reverse the apparent ordering of two conditions purely through a difference in marginal failure rates. That result is a caveat on our own headline numbers and, we think, a useful one for reading theirs.

2.4 Bounds without independence

Bounding a joint probability from marginals is classical. The Fréchet–Hoeffding inequalities (Fréchet, 1951; Hoeffding, 1940) give the sharp sandwich of Theorem 4.1, and Boole’s problem (Boole, 1854; Hailperin, 1965) recasts the search for the extremal joint as a linear program — the device Section 6 instantiates over co-execution moments. Copula theory (Sklar, 1959; Nelsen, 2006) parameterises dependence structures, and the Gaussian one-factor model of Remark 5.2 is the standard choice in reliability practice (Barlow and Proschan, 1975). Theorem 4.2 is our reason for refusing to certify on it.

The moment-problem view (Bertsimas and Popescu, 2005) bounds expectations subject to moment constraints, and Bonferroni-type inequalities (Bonferroni, 1936) supply the multiplicity correction Proposition 6.2 uses. Our contribution is not the LP, which is standard, but its use as a *finite-sample certificate* for agent pipelines: the combination of an exact Clopper–Pearson box (Clopper and Pearson, 1934) with the extremal LP, and the resulting guarantee of Theorem 5.2.

2.5 Sequential and anytime-valid inference

Wald’s SPRT (Wald, 1945) certifies with an expected sample size far below the fixed- n requirement, but at a stopping rule fixed in advance. Game-theoretic probability (Shafer and Vovk, 2019; Ville, 1939) and the recent literature on e-values and testing by betting (Shafer, 2021; Ramdas et al., 2023; Waudby-Smith and Ramdas, 2024; Grünwald et al., 2024), together with time-uniform confidence sequences (Robbins, 1970; Howard et al., 2021) replace that with certificates valid under arbitrary stopping. Section 7 applies this to graph reliability, where the payoff is specific: the null constrains only a conditional mean, so the certificate requires no independence assumption at all — the exact assumption this paper shows to be false for the composition bound. Proposition 7.1 records that the SPRT is recovered exactly at the optimally tuned bet, so anytime validity is obtained without loss against a known alternative.

2.6 Agent evaluation

Benchmarks for agent capability (Jimenez et al., 2024; Liu et al., 2024; Zhou et al., 2024) measure task success, and agent architectures (Yao et al., 2023; Shinn et al., 2023; Wang et al.,

2023; Park et al., 2023; Hong et al., 2024; Qian et al., 2024; Wu et al., 2024; Chase, 2022) optimise it. None reports whether component failures are dependent, which is the quantity a composed guarantee needs; task-success rates are marginals, and Section 10.2 shows marginals do not determine the composed outcome. Section 10 treats cross-agent failure dependence as the measured quantity under a preregistered manipulation of model sharing. We know of no earlier evaluation combining the four elements that design rests on: a registered hypothesis set, a manipulated sharing condition, three graph topologies, and deterministic contract scoring with no model in the judging loop. Correlated failure itself is not new — Section 2.3 reports it. What is new here is measuring it as an intervention rather than observing it, with a scoring rule that cannot itself induce correlation across arms.

3 Preliminaries: the ABC framework

This section states the Agent Behavioral Contract framework in full. The framework is due to the v1 paper (Bhardwaj, 2026), where several of these formulas were withheld under a patent claim that has since been withdrawn; every one is stated here without redaction, and Section B gives the complete catalogue with the corresponding v1 equation numbers. Readers familiar with v1 can skip to Section 4, which is where this work departs from it.

3.1 Contracts and compliance

Definition 3.1 (Behavioral contract; v1 Def. 3.1). A *behavioral contract* is a tuple

$$\mathcal{C} = (\mathcal{P}, \mathcal{I}_{\text{hard}}, \mathcal{I}_{\text{soft}}, \mathcal{G}_{\text{hard}}, \mathcal{G}_{\text{soft}}, \mathcal{R}),$$

where $\mathcal{P} = \{p_1, \dots, p_m\}$ is a finite set of *preconditions*, predicates over the initial state s_0 ; $\mathcal{I}_{\text{hard}}$ and $\mathcal{I}_{\text{soft}}$ are *hard* and *soft invariants*; $\mathcal{G}_{\text{hard}}$ and $\mathcal{G}_{\text{soft}}$ are hard and soft *governance policies* constraining operational authority such as tool use and spending; and $\mathcal{R} : (\mathcal{I}_{\text{soft}} \cup \mathcal{G}_{\text{soft}}) \times S \rightarrow A^*$ is a partial *recovery map* from a violated soft constraint and a state to a corrective action sequence. We write $\mathcal{C} = (\mathcal{P}, \mathcal{I}, \mathcal{G}, \mathcal{R})$ with $\mathcal{I} = \mathcal{I}_{\text{hard}} \cup \mathcal{I}_{\text{soft}}$ where the partition is not in play.

The hard/soft partition is the load-bearing distinction. A single hard violation is a contract breach. A soft violation is tolerated if recovery occurs inside a bounded window, which is what makes contracts usable against systems that are stochastic by construction.

Definition 3.2 (Constraint evaluation scores; v1 Def. 3.6, eqs. 1–2). For a state–action pair (s_t, a_t) ,

$$C_{\text{hard}}(t) = \frac{|\{c \in \mathcal{I}_{\text{hard}} \cup \mathcal{G}_{\text{hard}} : c(s_t, a_t) = \text{true}\}|}{|\mathcal{I}_{\text{hard}} \cup \mathcal{G}_{\text{hard}}|}, \quad (1)$$

$$C_{\text{soft}}(t) = \frac{|\{c \in \mathcal{I}_{\text{soft}} \cup \mathcal{G}_{\text{soft}} : c(s_t, a_t) = \text{true}\}|}{|\mathcal{I}_{\text{soft}} \cup \mathcal{G}_{\text{soft}}|}. \quad (2)$$

Both lie in $[0, 1]$. The evaluator is stateless: any turn is evaluable independently of the turns before it.

Definition 3.3 (Turn compliance). A turn is *hard-compliant* iff it incurs zero hard violations, i.e. $C_{\text{hard}}(t) = 1$, and *soft-compliant* iff zero soft violations remain outstanding after recovery. The binary hard verdict $h_t = \mathbf{1}[C_{\text{hard}}(t) = 1]$ is the quantity every dependence estimate in Section 10 is computed on.

3.2 Drift

Definition 3.4 (Composite drift score; v1 Def. 3.12, eqs. 7–9). The *drift* of an agent at time t is the convex combination

$$D(t) = w_c D_{\text{compliance}}(t) + w_d D_{\text{distributional}}(t), \quad w_c + w_d = 1, \quad (3)$$

with

$$D_{\text{compliance}}(t) = 1 - \bar{C}(t) = \frac{\sum_i w_i (1 - \sigma_i(t))}{\sum_i w_i}, \quad D_{\text{distributional}}(t) = \text{JSD}(P_{\text{obs}}(t) \parallel P_{\text{ref}}). \quad (4)$$

Defaults are $w_c = 0.6$, $w_d = 0.4$. Both components lie in $[0, 1]$, so $D(t) \in [0, 1]$.

Definition 3.5 (Distributional drift term; v1 eq. 9). $P_{\text{obs}}(t)$ is the empirical action distribution over a sliding window of recent actions and P_{ref} is a calibrated baseline from a compliant reference session. With $M = \frac{1}{2}(P + Q)$,

$$\text{JSD}(P \parallel Q) = \frac{1}{2} D_{\text{KL}}(P \parallel M) + \frac{1}{2} D_{\text{KL}}(Q \parallel M). \quad (5)$$

JSD (Lin, 1991) is symmetric and bounded in $[0, 1]$, and $\sqrt{\text{JSD}}$ is a metric (Endres and Schindelin, 2003), which is why it is preferred here to raw KL divergence.

Definition 3.6 ((p, δ, k) -satisfaction; v1 Def. 3.7, eqs. 3–4). An agent A (p, δ, k) -satisfies contract $\mathcal{C}$ over session length T if both hold:

$$\begin{aligned} \mathbb{P}[C_{\text{hard}}(t) = 1 \ \forall t \in \{0, \dots, T\} \mid P(s_0)] &\geq p, \quad (6) \\ \mathbb{P}[\forall t: C_{\text{soft}}(t) < 1 - \delta \implies \exists t' \in \{t, \dots, \min(t + k, T)\}: C_{\text{soft}}(t') \geq 1 - \delta \mid P(s_0)] &\geq p. \quad (7) \end{aligned}$$

Here $p \in [0, 1]$ is a probability threshold, $\delta \in [0, 1]$ a tolerable deviation, and $k \in \mathbb{N}$ a recovery window.

Remark 3.1. Equation (7) is a probabilistic guarantee about *recoverable* compliance. It is not the deterministic condition $\max_t |C_{\text{soft}}(t) - 1| \leq \delta$, which appears in some descriptions of this framework and states something strictly stronger and different. Definition 3.6 follows the v1 paper.

Definition 3.7 (Reliability index; v1 Def. 3.20, eq. 13).

$$\Theta = \alpha_1 \bar{C}(t) + \alpha_2 (1 - \bar{D}(t)) + \alpha_3 \frac{1}{1 + E} + \alpha_4 S, \quad \sum_{i=1}^4 \alpha_i = 1, \quad (8)$$

with defaults $\alpha_1 = 0.35$ (compliance), $\alpha_2 = 0.25$ (stability), $\alpha_3 = 0.20$ (event frequency, E the count of violation events), and $\alpha_4 = 0.20$ (stress resilience). S is the *stress resilience index* $\mathbb{E}[C(t) \mid \text{stressed}] / \mathbb{E}[C(t) \mid \text{baseline}]$ (v1 eq. 12), which is a distinct quantity from the recovery success rate; descriptions that equate the two are in error.

3.3 Drift dynamics

Definition 3.8 (Ornstein–Uhlenbeck drift dynamics; v1 Def. 4.1, eq. 14). Drift evolves as

$$dD = (\alpha - \gamma D(t)) dt + \sigma dW(t), \quad (9)$$

with $\alpha > 0$ the natural drift rate, $\gamma > 0$ the contract-recovery mean-reversion rate, $\sigma > 0$ the diffusion coefficient, and W a standard Wiener process. Treating $D(t) \geq 0$ is a modelling simplification (v1 Remark 4.2); the Gaussian stationary law of Theorem 3.1 places negligible but non-zero mass below zero.

Theorem 3.1 (Stationary drift law and bound; v1 Thm. 4.3). *Let D solve Definition 3.8 with $\alpha, \gamma, \sigma > 0$, and let the initial condition satisfy $\mathbb{E}[D(0)^2] < \infty$ with $D(0)$ independent of the driving Wiener process. Then:*

- (i) *the stationary law is $\pi_D = \mathcal{N}(\alpha/\gamma, \sigma^2/(2\gamma))$;*
- (ii) *$\mathbb{E}_\pi[D] = \alpha/\gamma$, so $\mathbb{E}_\pi[D] < 1$ iff $\gamma > \alpha$;*
- (iii) *$\text{Var}_\pi(D) = \sigma^2/(2\gamma)$;*
- (iv) *for $\eta > 0$, $\mathbb{P}_\pi(D > \alpha/\gamma + \eta) \leq \exp(-\gamma\eta^2/\sigma^2)$;*
- (v) *writing $e(t) = D(t) - \alpha/\gamma$, $\mathbb{E}[e(t)^2] = \mathbb{E}[e(0)^2]e^{-2\gamma t} + \frac{\sigma^2}{2\gamma}(1 - e^{-2\gamma t})$, so convergence to stationarity is exponential at rate 2γ . (For deterministic $D(0)$ this reads $e(0)^2e^{-2\gamma t} + \dots$)*

Proof. See Section A.1. □

Part (ii) is the design criterion: a contract whose recovery rate γ exceeds the natural drift rate α holds expected drift below 1 indefinitely. Part (iv) converts that mean statement into the tail bound a practitioner needs, and part (v) says the transient decays at rate 2γ , so the recovery rate sets both the stationary level and the time to reach it.

3.4 Composition

Definition 3.9 (Composability conditions; v1 Def. 4.7). Agents A and B with contracts $\mathcal{C}_A, \mathcal{C}_B$ and a handoff invariant $\mathcal{I}_h$ are *composable* if:

C1 Interface compatibility

$$\text{Type}(\text{Post}_A) \subseteq \text{Type}(\mathcal{P}_B).$$

C2 Assumption discharge

$$\text{Post}_A \wedge \mathcal{I}_h \implies \mathcal{P}_B.$$

C3 Governance consistency

$$\text{Allowed}(\mathcal{G}_A) \cap \text{Prohibited}(\mathcal{G}_B) = \emptyset.$$

C4 Recovery independence

any post-recovery state of A still satisfies $\mathcal{P}_B$.

Definition 3.10 (Conditional independence of contract satisfaction; v1 Thm. 4.11 preamble). Let E_A, E_B, E_h denote the events that A satisfies its contract, that B satisfies its contract, and that the handoff invariant holds. Condition **C5** states

$$\mathbb{P}(E_B \mid E_A \cap E_h) = \mathbb{P}(E_B \mid E_h). \quad (10)$$

C1–C4 are required for the deterministic composition theorem; C5 is required *additionally* for the probabilistic one. Treating the two sets as a single list “C1–C5” conflates the deterministic and probabilistic results.

Equation (10) is the assumption this paper is about. Section 4 shows what fails when it does not hold, and Section 10.2 measures the extent to which it does not hold in practice.

Definition 3.11 (Naive compositional bound; v1 Thm. 4.11, eqs. 28–29). Under C1–C5, for the two-agent composition $A \oplus B$,

$$p_{A \oplus B} \geq p_A \cdot p_B \cdot p_h, \quad \delta_{A \oplus B} \leq \delta_A + \delta_B + \delta_h, \quad (11)$$

extending to an N -agent chain (v1 Cor. 4.13, eqs. 30–31) as $p_{\text{chain}} \geq \prod_{i=1}^N p_i \prod_{i=1}^{N-1} p_{h_i}$ and $\delta_{\text{chain}} \leq \sum_i \delta_i + \sum_i \delta_{h_i}$.

3.5 Certification

Definition 3.12 (Sequential probability ratio certification; v1 §7). For $H_0 : p \leq p_0$ against $H_1 : p \geq p_1$, the SPRT accumulates the Bernoulli log-likelihood ratio and stops at the first boundary crossing. Its expected sample size satisfies

$$\mathbb{E}[N^*] \approx \frac{\log(1/\alpha_{\text{err}})}{D_{\text{KL}}(\hat{p} \| p_0)}, \quad (12)$$

scaling as $O(\log(1/\alpha)/D_{\text{KL}})$.

Definition 3.13 (Fixed-sample certification baseline). The fixed-sample requirement of Hoeffding (1963) for absolute accuracy ε at error α_{err} is $N_H = \log(2/\alpha_{\text{err}})/(2\varepsilon^2)$, scaling as $O(1/\varepsilon^2)$. At $p_0 = 0.85$, $p_1 = 0.95$, $\alpha = 0.05$ the SPRT needs roughly 58 sessions against Hoeffding’s 184; at $\varepsilon = 0.01$ the gap widens to about 300 against 18,445.

Remark 3.2. Both Definitions 3.12 and 3.13 certify at a stopping rule fixed in advance. Neither survives a team that watches a dashboard and stops when the number looks good. Section 7 replaces them with a certificate valid under arbitrary stopping, and Section 10.5 measures its type-I error.

3.6 Cost

Proposition 3.1 (Per-action enforcement cost; v1 Prop. 4.15). *Enforcement costs $O(k + |A|)$ per action, where k is the number of constraints and $|A|$ the action-vocabulary size: $O(k)$ for constraint evaluation, $O(|A|)$ for the JSD histogram update, and $O(1)$ for compliance aggregation. For $k < 100$ and $|A| < 50$ the measured overhead is under 10 ms per action (v1 Remark 4.16).*

3.7 Graphs, motifs, and co-failure

The remaining definitions are new to this paper and set up Section 4 onward.

Definition 3.14 (Agent graph and graph outcome). An *agent graph* G is a finite DAG whose nodes are agents, each carrying a contract, and whose edges are handoffs carrying handoff invariants. Node i produces a hard verdict $h_i \in \{0, 1\}$. A *composition rule* ρ maps the node verdicts to the graph outcome $Y_G = \rho(h_1, \dots, h_m) \in \{0, 1\}$. We write $p_i = \mathbb{P}(h_i = 1)$.

Definition 3.15 (Motifs). Two composition rules are used throughout: *series* $\rho_{\text{ser}} = \bigwedge_i h_i$ over a linear handoff chain, and *quorum* $\rho_{q/m} = \mathbf{1}[\sum_i h_i \geq q]$ over m workers feeding a deterministic q -of- m aggregator. The parallel motif is the $q = 1$ case: two independent branches feed a deterministic merge and $\rho_{\text{par}} = \mathbf{1}[\sum_i h_i \geq 1]$, so the graph succeeds when at least one branch passes and the merge is valid. Parallel is therefore a redundancy rule, not a conjunction; the distinction matters because Corollary 4.1 shows the two rules put positive dependence on *opposite* sides of the safety argument. Deterministic merge and aggregator nodes carry no contract and are excluded from dependence estimation.

Definition 3.16 (Co-failure table and sharing condition). For nodes i, j with failure indicators $F_i = 1 - h_i$, the *co-failure table* over n missions has cells $n_{11} = \sum \mathbf{1}[F_i F_j = 1]$, $n_{10} = \sum \mathbf{1}[F_i(1 - F_j) = 1]$, $n_{01} = \sum \mathbf{1}[(1 - F_i)F_j = 1]$, $n_{00} = \sum \mathbf{1}[(1 - F_i)(1 - F_j) = 1]$, with proportions $p_{ab} = n_{ab}/n$ and marginal failure rates $p_i = p_{11} + p_{10}$, $p_j = p_{11} + p_{01}$. The *sharing condition* of the pair is `same_model` when both nodes run the same model weights, `same_vendor` when they run different models from one vendor, and `different_vendor` otherwise.

Definition 3.17 (Dependence functionals). On a co-failure table we use the *overlap* $J = n_{11}/(n_{11} + n_{10} + n_{01})$, the binary Kendall $\tau_a = 2(p_{11}p_{00} - p_{10}p_{01})$, the phi coefficient $\phi =$

$(p_{11}p_{00} - p_{10}p_{01})/\sqrt{p_i(1-p_i)p_j(1-p_j)}$, the log odds ratio, and Yule's Q . We call J , τ_a , ϕ *marginal-sensitive* and the last two *marginal-free*: the latter are invariant under row and column scaling of the table, the former are not. Section 10.2.3 shows the distinction is decisive rather than cosmetic.

4 What breaks when independence fails

Definition 3.11 certifies a composed system by multiplying component reliabilities. That step is licensed by C5 (Definition 3.10) and by nothing else. This section establishes three things: that the multiplicative bound is not conservative when C5 fails, that the assumption-free alternative is too weak to be useful, and that a certificate built on a *fitted* dependence model can lose coverage of the truth entirely, and that collecting more data does not repair it.

4.1 The bound is not conservative

Proposition 4.1 (Signed compositional gap). *Let G be a series motif over components with hard verdicts $h_1, \dots, h_m$ and $Y_G = \bigwedge_i h_i$. Then*

$$\mathbb{P}(Y_G = 1) - \prod_{i=1}^m p_i = \text{Cov}\left(h_1, \prod_{i=2}^m h_i\right) + \sum_{k=2}^{m-1} \left(\prod_{i=1}^{k-1} p_i\right) \text{Cov}\left(h_k, \prod_{i=k+1}^m h_i\right). \quad (13)$$

*In particular for $m = 2$ the gap is exactly $\text{Cov}(h_1, h_2)$, so under positive dependence the independence product **understates** series reliability — and correspondingly overstates the probability that the series fails at all. The unsafe direction is a different estimand, stated separately in Corollary 4.1.*

Proof. See Section A.3. □

For a *series* system, then, the independence product is genuinely conservative and dependence only helps. The unsafe case is redundancy, and it is a distinct statement:

Corollary 4.1 (Redundant joint failure is understated). *Let $F_i = 1 - h_i$. For two components, $\mathbb{P}(F_1 = F_2 = 1) = (1 - p_1)(1 - p_2) + \text{Cov}(F_1, F_2)$ and $\text{Cov}(F_1, F_2) = \text{Cov}(h_1, h_2)$. Under positive dependence the probability that both redundant paths fail therefore exceeds the independence product, so a redundant design is over-credited exactly when its components are positively associated.*

Proof. See Section A.4. □

This is the direction that matters operationally: a quorum or a reviewer–writer pair is deployed precisely so that $\mathbb{P}(\text{all redundant paths fail})$ is small. Section 10.2 measures ϕ between 0.58 and 0.92; at those magnitudes the independent calculation is not close.

Example 4.1. Take a 2-of-3 quorum whose workers each satisfy their contract with $p = 0.61$, the `same_model` rate measured in Section 10.2. Under independence the quorum succeeds with probability $3p^2(1-p) + p^3 = 0.6623$ and therefore fails with probability 0.3377. An exchangeable joint consistent with the same marginals and the measured pairwise association puts quorum failure near 0.37, a relative increase of about 9.6% in the quantity the redundancy was bought to reduce. We stress that marginals and ϕ alone do not determine a 2-of-3 failure probability — the triple moment is free — so this is an illustration of magnitude under one consistent joint, not a bound. Section 6 is the machinery for the bound.

4.2 The assumption-free alternative is too weak

Dropping C5 entirely leaves the Fréchet–Hoeffding bounds, which hold for every joint law with the given marginals.

Theorem 4.1 (Fréchet–Hoeffding sandwich for all-success). *For any joint law of $(h_1, \dots, h_m)$ with marginals $p_1, \dots, p_m$,*

$$\max\left(0, \sum_{i=1}^m p_i - (m-1)\right) \leq \mathbb{P}\left(\bigwedge_i h_i = 1\right) \leq \min_i p_i, \quad (14)$$

and both bounds are attained, so neither can be improved without further information.

Proof. See Section A.5. □

Corollary 4.2 (Vacuity of the assumption-free floor). *The lower bound of Theorem 4.1 is 0 whenever $\sum_i p_i \leq m-1$, i.e. whenever the mean component reliability is at most $1-1/m$. For $m=4$ components each at $p=0.75$ the certified floor is exactly 0.*

Corollary 4.2 is the practical bind. C5 gives a number that is wrong in the unsafe direction; dropping C5 gives a number that is right and useless. The moment-set certificate of Section 6 occupies the space between them by constraining the joint with measured co-execution moments rather than with an independence assumption, and Section 10.3 measures what that buys: a floor of 0.4116 where the pairwise-only certificate gives 0.2455.

4.3 Fitting a dependence model does not rescue the situation

A natural response is to fit a parametric dependence model — a one-factor Gaussian copula, say — estimate its parameters, and report a confidence bound on the fitted model’s all-success functional. This fails, and it fails in a way that gets worse with more data.

Definition 4.1 (Model functional and identified set). Fix a moment family $\mathcal{J}$ and let $\mathcal{M}(\hat{\mu})$ be the set of joint laws on $\{0,1\}^m$ whose $\mathcal{J}$ -moments equal $\hat{\mu}$. The *identified set* for all-success is $[\underline{R}(\hat{\mu}), \overline{R}(\hat{\mu})]$ where $\underline{R} = \inf_{Q \in \mathcal{M}} Q(\bigwedge_i h_i = 1)$ and $\overline{R} = \sup$. A parametric family $\mathcal{F} \subset \mathcal{M}$ induces a *model functional* $R_{\mathcal{F}}(\hat{\mu})$, the all-success probability of the fitted member of $\mathcal{F}$.

Theorem 4.2 (Coverage collapse of a model-based floor). *Let Q^* be a true joint law with $\mathcal{J}$ -moments μ^* and true all-success probability $R^* = Q^*(\bigwedge_i h_i = 1)$. Suppose the parametric family $\mathcal{F}$ is misspecified at μ^* in the sense that*

$$\Delta := R_{\mathcal{F}}(\mu^*) - R^* > 0. \quad (15)$$

Let $\hat{L}_n$ be any bootstrap lower confidence bound for $R_{\mathcal{F}}(\mu^*)$ satisfying $\hat{L}_n = R_{\mathcal{F}}(\mu^*) - O_p(n^{-1/2})$. Then

$$\mathbb{P}(\hat{L}_n \leq R^*) \longrightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (16)$$

That is, coverage of the true all-success probability tends to zero. More data does not repair the certificate: the haircut shrinks while the target does not move. We claim the limit and not monotonicity in n , which tightness alone does not entail.

Proof. See Section A.6. □

The mechanism is a mismatch of orders. The identification gap Δ in Equation (15) is $O(1)$ — it is a property of the family, not of the sample — while the bootstrap haircut shrinks like $n^{-1/2}$. Past some finite n the entire interval sits above R^* and never returns. The bootstrap is not malfunctioning; it is covering the estimand it was asked about, which is the fitted model’s functional and not the truth.

Example 4.2 (An explicit witness). Take the equicorrelated Gaussian-copula joint at $p = 0.6$, $\lambda = 0.8$ (a factor *loading*, so the latent correlation is $\lambda^2 = 0.64$), $m = 3$. By Gauss–Legendre quadrature over the common factor its marginals are 0.600000, its pairwise co-success moments are 0.465237, and its all-success probability is 0.392144 — so the one-factor family is *correctly specified* for it and $\Delta = 0$. Every constant quoted for this law in this paper comes from that one deterministic routine, not from simulation. It is therefore **not** itself a witness for Theorem 4.2, a point worth stating because it is the natural law to reach for and it does not work.

The witness is instead the joint attaining the lower end of the LP-identified set. Over the pairwise moment family that set is $[0.330475, 0.465237]$, and by Theorem 6.1 the lower endpoint is attained. For this exchangeable case it has the closed form $2q - p$ in the pairwise moment q and marginal p , and the attaining law Q^* puts mass on five cells only:

$$\begin{aligned} Q^*(1, 1, 1) &= 0.330475, & Q^*(0, 0, 0) &= 0.265237, \\ Q^*(0, 1, 1) &= Q^*(1, 0, 1) = Q^*(1, 1, 0) = 0.134763, & \text{all other cells} &= 0. \end{aligned}$$

By construction Q^* has *exactly* the same marginals and *exactly* the same pairwise co-success moments as the Gaussian joint: the two are indistinguishable to any procedure that sees only those moments. Yet

$$\Delta = R_{\mathcal{F}}(\mu^*) - R^* = 0.392144 - 0.330475 = 0.061670 > 0,$$

so Theorem 4.2 applies to Q^* . Section 10.4 samples from Q^* and measures the collapse, and runs the Gaussian joint as a correctly-specified control in which no collapse should occur — and does not.

Corollary 4.3 (Why the copula-agnostic floor is the right target). *The LP floor $\underline{R}(\hat{\mu})$ of Definition 4.1 satisfies $\underline{R}(\mu^*) \leq R^*$ by construction for every $Q^* \in \mathcal{M}(\mu^*)$, with no parametric assumption. A confidence bound targeting $\underline{R}$ therefore cannot suffer the failure of Theorem 4.2; its conservatism is the price of that immunity, and Section 6 shows the price falls as the moment family grows.*

5 A tiered certificate

Section 4 leaves three unusable options: a multiplicative bound that is unsafe on the failure side, an assumption-free bound that is often exactly zero, and a fitted model whose coverage collapses. What survives depends on *what was actually measured*. A certificate should therefore not be a single number but a selection among tiers, each stating the evidence it requires and the assumptions it carries.

Definition 5.1 (Certification scope). A *certificate* is a triple $(\hat{L}, \mathcal{A}, \mathcal{S})$ where $\hat{L}$ is a lower confidence bound on all-success reliability at level $1 - \eta_{\text{conf}}$, $\mathcal{A}$ is the set of assumptions it rests on, and $\mathcal{S}$ is its scope: the mission distribution, the model versions, and the topology under which it was obtained. A certificate makes no claim outside $\mathcal{S}$; in particular it does not transfer across mission mixes, model upgrades, or topology changes.

Definition 5.2 (The three tiers). Let $\mathbf{h} \in \{0, 1\}^{m \times n}$ be a pass matrix over m stages and n missions.

Tier 0 — observed

Available only when the composition was executed end-to-end, i.e. every stage was scored on the same missions, so Y_G is directly observed. $\hat{L}_0$ is the exact Clopper–Pearson lower bound on $\mathbb{P}(Y_G = 1)$ from $\sum_r \mathbf{1}[Y_G^{(r)} = 1]$ successes in n trials. Assumptions: i.i.d. missions. No copula, no model.

Tier 1 — copula-agnostic

Available from per-stage and co-execution data alone, when the composed pipeline was *not* run end-to-end. $\hat{L}_1$ is the LP minimum of Section 6 over the Bonferroni–Clopper–Pearson moment box. Assumptions: i.i.d. missions. No copula, no model.

Tier 2 — model

The finite-sample Gaussian one-factor floor, via the Slepian (1962) corner. Assumptions: i.i.d. missions *and* correct specification of the one-factor family.

Theorem 5.1 (Tier-0 validity). *If missions are i.i.d. and Y_G is observed on each, then the Clopper–Pearson lower bound $\hat{L}_0$ satisfies $\mathbb{P}(\hat{L}_0 \leq \mathbb{P}(Y_G = 1)) \geq 1 - \eta_{\text{conf}}$ exactly, for every n and every true reliability.*

Proof. See Section A.7. □

Theorem 5.2 (Tier-1 validity). *Let $\mathcal{J}$ be a moment family of size J , let $\hat{\mu} = (\hat{\mu}_S)_{S \in \mathcal{J}}$ be the empirical moments, and let $B(\hat{\mu})$ be the box formed by two-sided Clopper–Pearson intervals for each μ_S at level $\eta_{\text{conf}}/(2J)$ per tail. Define*

$$\hat{L}_1 = \min \left\{ Q(\bigwedge_i h_i = 1) : Q \in \Delta(\{0, 1\}^m), (Q(\bigwedge_{i \in S} h_i = 1))_{S \in \mathcal{J}} \in B(\hat{\mu}) \right\}. \quad (17)$$

Then for i.i.d. missions and any true law Q^* , $\mathbb{P}(\hat{L}_1 \leq Q^*(\bigwedge_i h_i = 1)) \geq 1 - \eta_{\text{conf}}$.

Proof. See Section A.8. □

Theorem 5.2 is the reason Tier 1 is safe where Tier 2 is not: it is a minimum over *every* joint law consistent with the box, so the true law is one of the candidates whenever the box covers, and the union bound makes the box cover with the stated probability. Nothing is assumed about the dependence structure.

Proposition 5.1 (Fail-safe tier selection). *Let the certificate select Tier 0 only on an explicit assertion that the composition was executed end-to-end, and Tier 1 by default otherwise. Then no the two failure directions are not symmetric. Tier 1 requires only i.i.d. missions (Theorem 5.2), which Tier 0 requires as well, so the assumption sets are nested. Consequently:*

- (i) **Omission is safe.** *Failing to set the flag when it was warranted yields $\hat{L}_1$, which is valid — merely looser than the Tier-0 bound that was available.*
- (ii) **Commission is not.** *Setting the flag when it was not warranted yields $\hat{L}_0$, whose validity requires an assumption that does not hold; that guarantee is unsound.*

*The design content of the proposition is therefore that the unsound case requires a deliberate assertion, while the default — and every omission — lands in the safe case. It is **not** a claim that no setting of the flag can produce an unsound certificate; case (ii) is exactly such a setting, and no inspection of the pass matrix can detect it.*

Proof. See Section A.9. □

Remark 5.1 (No pointwise ordering). We deliberately do *not* claim $\hat{L}_1 \leq \hat{L}_0$ pointwise. Both bound the same true reliability from below on their covering events, but neither dominates the other as a random variable: an unlucky low all-success count depresses $\hat{L}_0$ while leaving the marginal and pairwise moments — and hence $\hat{L}_1$ — comparatively high. The fail-safety in Proposition 5.1 comes from the nesting of the assumption sets, not from an ordering of the two numbers.

Proposition 5.1 states a design constraint rather than a mathematical discovery, and it is stated because the alternative is a silent failure. Tier 0’s validity requires that the joint outcome was observed — a fact about how the data were produced, which no inspection of the pass matrix can confirm or refute. A default of Tier 0 would let a forgotten flag convert an extrapolation into an over-claim with no detectable symptom. Defaulting to Tier 1 makes the failure mode a needlessly weak certificate instead.

Remark 5.2 (Tier 2 is a diagnostic). Tier 2 is reported and never certified. By Theorem 4.2 its coverage of the true reliability tends to zero under a misspecification that marginal and pairwise data cannot detect, and Section 10.4 measures that collapse on the witness of Example 4.2. Reporting it is useful — a large gap between Tier 1 and Tier 2 signals that the moment family is leaving information on the table — but printing it as a guarantee would reintroduce exactly the failure this paper documents.

6 Moment-set certification

Tier 1 turns on a linear program. This section states it, proves it sharp, and characterises how it behaves as the moment family grows — which is the knob a practitioner actually turns, and the one Section 10.3 measures.

6.1 The program

Definition 6.1 (Cell representation). A joint law on $\{0, 1\}^m$ is a vector $x \in \mathbb{R}^{2^m}$ with $x_c \geq 0$ and $\sum_c x_c = 1$, where cell c ranges over binary patterns. For $S \subseteq \{1, \dots, m\}$ let $a_S \in \{0, 1\}^{2^m}$ be the indicator row $(a_S)_c = \mathbf{1}[c_i = 1 \ \forall i \in S]$, so $a_S^\top x$ is the probability that every stage in S succeeds. All-success is $a_{\{1..m\}}^\top x$.

Definition 6.2 (Moment-set LP). Given a moment family $\mathcal{J}$ and values $(\nu_S)_{S \in \mathcal{J}}$,

$$\underline{R} = \min_x a_{\{1..m\}}^\top x \quad \text{s.t.} \quad a_S^\top x = \nu_S \quad \forall S \in \mathcal{J}, \quad \mathbf{1}^\top x = 1, \quad x \geq 0, \quad (18)$$

and $\bar{R}$ the corresponding maximum. With $\mathcal{J} = \{\text{singletons}\} \cup \{\text{pairs}\}$ this is the pairwise program; adding triples gives the $J = 14$ program of Section 10.3 at $m = 4$. For a box constraint $\nu \in B$ rather than point values, replace each equality by the pair of inequalities defining B .

Theorem 6.1 (Soundness and sharpness). *Let Q^* be any joint law on $\{0, 1\}^m$ whose $\mathcal{J}$ -moments equal ν . Then $\underline{R} \leq Q^*(\bigwedge_i h_i = 1) \leq \bar{R}$, and both bounds are attained: there exist laws Q_-, Q_+ with the same $\mathcal{J}$ -moments achieving $\underline{R}$ and $\bar{R}$ respectively. Consequently $[\underline{R}, \bar{R}]$ is exactly the identified set of Definition 4.1, and no bound using only the moments in $\mathcal{J}$ can be tighter.*

Proof. See Section A.10. □

Sharpness matters for interpretation. A conservative-but-loose bound invites the response “the truth is surely much better than that”; Theorem 6.1 says that for the information supplied there is a consistent world in which the truth is exactly $\underline{R}$. Tightening requires more moments, not more optimism.

Proposition 6.1 (Monotonicity in the moment family). *If $\mathcal{J} \subseteq \mathcal{J}'$ and the moment values agree on $\mathcal{J}$, then $\underline{R}(\mathcal{J}) \leq \underline{R}(\mathcal{J}')$ and $\bar{R}(\mathcal{J}) \geq \bar{R}(\mathcal{J}')$. The identified set shrinks weakly as the family grows.*

Proof. See Section A.11. □

Proposition 6.1 holds at the level of *point* moments. It does not automatically survive the finite-sample box, and the reason is a Bonferroni budget rather than anything deep.

Proposition 6.2 (Bonferroni allocation and monotonicity). *Fix η_{conf} and let the box B be built from $J = |\mathcal{J}|$ Clopper–Pearson intervals.*

- (i) *Used-set allocation: spending $\eta_{\text{conf}}/(2J)$ per tail gives the tightest box for that family, but enlarging $\mathcal{J}$ widens every interval, so Proposition 6.1 can fail on the certified floor.*
- (ii) *Pre-allocated allocation: fixing a budget family $\mathcal{J}_{\text{max}} \supseteq \mathcal{J}$ and spending $\eta_{\text{conf}}/(2|\mathcal{J}_{\text{max}}|)$ per tail regardless of which moments are constrained makes every interval width independent of $|\mathcal{J}|$. Then enriching $\mathcal{J}$ only adds constraints, and the certified floor is monotone by construction.*

Both are valid at level $1 - \eta_{\text{conf}}$.

Proof. See Section A.12. □

The trade-off is measurable rather than theoretical. At $m = 4$, $n = 717$ in Section 10.3, pre-allocation costs 0.0098 of certified floor at $J = 10$ (0.2357 against 0.2455) and costs nothing at $J = 14$, where the budget family is the family in use. We report both, because reporting only the used-set number would present an empirically-monotone-so-far quantity as a guaranteed one.

Remark 6.1 (Cost). Equation (18) has 2^m variables and $|\mathcal{J}| + 1$ constraints, so it is exact and cheap for the small m that agent motifs actually use ($m \leq 8$ covers every topology in Section 10) and exponential beyond. For large m the practical route is to certify sub-compositions and compose the certificates, which costs tightness; we do not pursue it here because no motif in this evaluation needs it.

7 Anytime-valid certification

Every certificate so far is fixed- n : valid at a sample size chosen before the data arrive. Deployed agent systems are not monitored that way. A team watches a reliability dashboard, and when the number looks good enough they ship. That is optional stopping, and it inflates type-I error without bound for a fixed- n interval. This section gives a certificate that survives it, at a cost this section also states. The machinery is that of game-theoretic probability and testing by betting (Ville, 1939; Shafer and Vovk, 2019; Shafer, 2021; Ramdas et al., 2023; Waudby-Smith and Ramdas, 2024; Grünwald et al., 2024); the application to composed agent reliability is what is new here.

7.1 The e-process

Definition 7.1 (Betting e-process for graph reliability). Let $y_r \in \{0, 1\}$ be the graph outcome of mission r and let $\mathcal{F}_r$ be the σ -field generated by $y_1, \dots, y_r$. Test the *sequential* null

$$H_0 : \mathbb{E}[y_r \mid \mathcal{F}_{r-1}] \leq p_0 \quad \text{almost surely, for every } r \geq 1. \quad (19)$$

This is the null the betting construction requires, and it is strictly stronger than the marginal statement $\mathbb{P}(Y_G = 1) \leq p_0$. The distinction is not pedantic: a stream whose every mission has marginal success rate p_0 but whose outcomes are perfectly dependent satisfies the marginal statement, violates Equation (19), and drives the wealth process out of the supermartingale class entirely.¹ Let λ_r be a *predictable* betting fraction, i.e. λ_r is $\mathcal{F}_{r-1}$ -measurable, with $\lambda_r \in [0, 1/p_0]$. The wealth process is

$$E_R = \prod_{r \leq R} (1 + \lambda_r(y_r - p_0)), \quad E_0 = 1. \quad (20)$$

The certificate *fires* at the first R with $E_R \geq 1/\alpha$.

¹Take $p_0 = \frac{1}{2}$, $\lambda_r \equiv 1$, and $y_r \equiv Z$ for every r with $Z \sim \text{Bernoulli}(\frac{1}{2})$. Every marginal satisfies $\mathbb{P}(y_r = 1) = \frac{1}{2} \leq p_0$, yet $\mathbb{E}[E_2] = \frac{5}{4} > 1$ and $E_8 = (3/2)^8 > 20 = 1/\alpha$ on the event $\{Z = 1\}$, so the certificate fires with probability $\frac{1}{2}$ rather than $\alpha = 0.05$.

Lemma 7.1 (Supermartingale property). *Under H_0 , $(E_R)_{R \geq 0}$ is a non-negative supermartingale with $\mathbb{E}[E_R] \leq 1$ for all R .*

Proof. See Section A.13. □

Theorem 7.1 (Anytime validity; Ville, 1939). *Under H_0 , for any stopping time T — including data-dependent and unbounded ones —*

$$\mathbb{P}\left(\sup_{R \geq 1} E_R \geq 1/\alpha\right) \leq \alpha. \quad (21)$$

Consequently a team may inspect E_R after every mission, stop whenever they choose, and the probability of ever issuing a false certificate is at most α . The same device underlies time-uniform confidence sequences (Robbins, 1970; Howard et al., 2021).

Proof. See Section A.14. □

Two properties matter for this paper specifically. First, H_0 constrains only the scalar conditional mean of y_r , and constrains it one mission at a time, so **no independence assumption is required** — not across missions, not across components, not across shared-model shocks. The dependence may take any form whatever, provided it does not lift the conditional mean of the next outcome above p_0 ; and a correlated shock that *degrades* reliability is exactly a shock that does not. This is the relevant direction: the sequential certificate is immune to the exact failure mode Section 4 documents for the composition bound. Second, the guarantee is on the whole path, which is why certification must latch: once E_R has crossed, the event “ $\sup_R E_R \geq 1/\alpha$ ” has occurred and no later evidence undoes it. Section 10.5 verifies the shipped implementation latches, driving a certified process through 340 consecutive failures.

7.2 Relation to the SPRT

The fixed-stopping predecessor is Wald (1945)’s sequential test.

Proposition 7.1 (Exact SPRT recovery). *Fix an alternative $p_1 > p_0$ and set $\lambda^* = (p_1 - p_0)/(p_0(1 - p_0))$. Then for $y \in \{0, 1\}$,*

$$1 + \lambda^*(y - p_0) = \left(\frac{p_1}{p_0}\right)^y \left(\frac{1 - p_1}{1 - p_0}\right)^{1-y}, \quad (22)$$

so Equation (20) is exactly the Bernoulli likelihood ratio and E_R is the SPRT statistic. The e -process therefore contains Definition 3.12 as the special case of a constant, optimally-tuned bet.

Proof. See Section A.15. □

Proposition 7.1 places the cost of anytime validity precisely. Against a known alternative the SPRT is recovered with no loss; the price is paid only when λ is mis-tuned, and it is a loss of *power*, never of validity — Theorem 7.1 holds for every predictable λ . Section 10.5 measures both sides: the type-I rate stays under α across the whole admissible range of λ , while the mean time to certification varies from 14.6 missions at $\lambda = 1.2375$ to 403.9 at $\lambda = 0.125$. A badly chosen bet costs a factor of 28 in detection time and nothing in soundness.

Proposition 7.2 (Mixture over bets). *Let π be a prior on a finite set Λ of predictable betting strategies and let $E_R^{\text{mix}} = \sum_{\lambda \in \Lambda} \pi(\lambda) E_R^\lambda$. Then E^{mix} is a non-negative supermartingale under H_0 , so Theorem 7.1 applies unchanged, and*

$$\log E_R^{\text{mix}} \geq \max_{\lambda \in \Lambda} \log E_R^\lambda - \log \frac{1}{\pi(\lambda)}, \quad (23)$$

a pathwise log-regret guarantee against the best bet in hindsight.

Proof. See Section A.16. □

Proposition 7.2 is what makes the method usable when p_1 is unknown, which is the normal case: the mixture pays an additive $\log(1/\pi(\lambda))$ against the best fixed bet and retains anytime validity exactly.

8 Runtime enforcement

A certificate is a claim about a system that behaves as contracted. Producing that system requires intercepting agent actions before they take effect, which is an engineering problem with a small number of solutions and sharply different guarantees between them.

8.1 Three interception planes

An agent action can be intercepted at exactly three points, and where you intercept determines what you can enforce.

P1 — in-process.

A hook inside the orchestration framework, invoked around each tool call. Sees the parsed call, the agent identity, and the framework’s state; can deny by the framework’s own veto convention. Requires framework support and a per-framework adapter.

P2 — tool protocol.

An interposer on the tool-invocation protocol between the model host and the tool server. Sees every tool call and result as structured messages, and is framework-independent. Cannot see reasoning that never becomes a tool call.

P3 — transport.

A proxy on the HTTP path to the model provider. Sees requests and responses but must reconstruct semantic structure from wire format, and cannot attribute an action to an agent in a multi-agent process without additional context.

The planes are complementary rather than ranked. P1 gives the richest state and the weakest portability; P3 gives the reverse; P2 sits between and is the only plane that is both structured and framework-neutral.

8.2 Denial semantics

Definition 8.1 (Pre-action and post-action denial). A *pre-action* denial prevents the action from executing. A *post-action* denial permits execution and withholds the result. Only pre-action denial enforces a governance constraint on an irreversible operation; post-action denial is a redaction mechanism.

Definition 8.1 is the distinction that decides whether a deployment is enforced or merely monitored, and it is not always available. A plane that observes a tool result has already permitted the tool to run. Section 9 records, per surface, which of the two is achievable, because a framework that reports “enforcement enabled” while only offering post-action denial is making a claim its users will misread.

8.3 Failure policy

Definition 8.2 (Fail-closed and fail-open). When the enforcement path itself errors — a malformed contract, an evaluator exception, an unreachable policy store — a *fail-closed* policy denies the action and a *fail-open* policy permits it. A third state is required: *unevaluated*, in

which no verdict was reached and the action is neither credited as compliant nor recorded as violating.

The third state matters for the measurement in Section 10. Counting an unevaluated action as compliant inflates every reliability estimate; counting it as a violation deflates them. Both are silent. The implementation carries `evaluated` as an explicit field on every verdict so that unevaluated actions are excluded from the denominator rather than assigned to a side.

8.4 Enforcement and certification are different claims

Enforcement constrains what an agent does. Certification bounds how often the constrained system satisfies its contract. Neither implies the other: a perfectly enforced hard constraint still permits soft-constraint drift (Definition 3.6), and a high certified floor says nothing about whether any individual action was blocked. Section 10 measures the second; this section describes the machinery that makes the first possible, and the two are reported separately throughout.

9 Implementation

Everything in Section 3–Section 7 that is computable is implemented in `AGENTASSERT`, an open-source Python library (AGPL-3.0). We describe it at the level a reader needs to reproduce Section 10 or to check a claim against code; Section C gives the artifact details.

9.1 Structure

The library separates the three concerns this paper keeps distinct. Contract specification lives in `evaluator/`, `dsl/`, and `models.py` (Definitions 3.1 and 3.2); measurement in `metrics/` and `dependence/` (Definitions 3.4, 3.7 and 3.17); and certification in `certification/`, which carries one module per tier of Definition 5.2 — `observed_floor`, `lp_bound`, `slepian_floor` (diagnostic only), `eprocess`, `sppt` — assembled by `certificate` under Proposition 5.1. Runtime enforcement lives in `gateway/` and `enforce/`, with the interception surfaces of Section 8 as thin adapters over them.

9.2 One policy path, several surfaces

Enforcement across agent frameworks with different hook conventions would ordinarily mean one policy implementation per framework and one opportunity per framework to diverge. Instead a single bridge exposes a pre-action decision and a post-action outcome, and each shim translates those into its host’s veto convention — returning false from a before-call hook, declining to invoke the continuation, not awaiting the next middleware, or raising from a pre-hook. The shims match each host’s hook shapes structurally and import nothing from them, so the library depends on no agent framework.

Two consequences matter for this paper. Policy is evaluated in one place, so a contract enforced through a tool-protocol interposer and the same contract enforced in-process yield the same verdict on the same state — which is what makes cross-surface measurement comparable. And because Definition 8.1 distinguishes pre-action from post-action denial while not every host offers the former, the library publishes a per-surface capability matrix recording where enforcement is impossible. The negative entries are the point: a library whose documentation implies uniform coverage will be deployed against hosts where the guarantee does not hold.

9.3 State representation

Contracts are evaluated against a flat, dotted-key state (`output.pii_detected` rather than a nested object). The convention is not cosmetic. An earlier version returned nested dictionaries

from one adapter and flat keys from another; a constraint written against `output.pii_detected` then resolved correctly under enforcement and silently evaluated to false under measurement, so the same contract yielded two different compliance rates depending on which path produced the state. Flattening at every adapter boundary removes the divergence by construction.

9.4 Verification

The library carries 1697 tests at 93% statement coverage. Beyond conventional unit tests, three properties from this paper are pinned by tests because they are the ones an implementation silently gets wrong: that a certified e-process never un-certifies after wealth collapses (Section 10.5); that the default certificate tier is Tier 1 unless end-to-end execution is explicitly asserted (Proposition 5.1); and that a contract which could not be evaluated is recorded as unevaluated rather than counted as either outcome (Definition 8.2).

A paper-to-code parity test checks the formulas of Section 3 against their implementations, so a change to either that breaks correspondence fails the suite. That test exists because two defects found during preparation of this work — a mishandled constraint state across the proxy and hook surfaces, and an incorrect divergence term in a drift computation — produced confidently wrong numbers rather than errors.

10 Evaluation

We answer six questions with six experiments. E1 measures whether model sharing induces correlated contract failures across a pipeline (Section 10.2). E2 measures how much a moment-set certificate lifts the reliability floor over the Fréchet bound (Section 10.3). E3 measures how fast independence-assuming coverage collapses as dependence grows (Section 10.4). E4 measures the empirical type-I error of the anytime-valid certificate (Section 10.5). E5 measures contract behavior on frontier models under live API conditions (Section 10.6). E6 is an ablation over the clustering and serial-dependence assumptions the analysis rests on (Section 10.7). Section 10.8 re-presents the v1 single-agent evidence this work builds on, and Section 10.9 states what the six experiments jointly support and what they do not.

10.1 Setup, preregistration, and scoring

Preregistration. The confirmatory hypotheses, conditions, sample sizes, sampling parameters, primary estimator, stopping rule, and falsification criteria for E1 were committed to the repository before any confirmatory outcome was generated (Nosek et al., 2018), and the commit is git-timestamped. We reproduce the three registered hypotheses verbatim:

- **H1 (dependence exists under sharing).** In the `same_model` condition, co-failure dependence between the two pipeline agents is positive: Kendall’s $\tau_a > 0$ with a bootstrap CI excluding 0.
- **H2 (dependence decreases with less sharing).** τ_a is weakly monotone decreasing across `same_model` $\geq$ `same_vendor` $\geq$ `different_vendor`.
- **H3 (naive bound is anti-conservative under sharing).** Observed graph reliability $\mathbb{P}(Y_G = 1)$ *exceeds* the independence product $\prod_i p_i$ under positive dependence (the compositional gap is signed and non-zero), so a dependence-aware bound is required.

The registration further designates `series2` as the confirmatory motif and names the three OpenRouter arms at $n = 6000$ as the primary confirmatory arms, placing other motifs in “future/secondary work”. We hold to that scope: **the confirmatory evaluation is series2 over $3 \times 6000 = 18,000$ missions.** The `quorum2of3` and `parallel2` results, and the two breadth

arms of Section 10.6, are reported as *secondary and exploratory* throughout. Where they agree with the confirmatory arms we say so; they are not evidence at the registered level, and we have not filed a dated amendment broadening the registration.

The registration also fixed what would falsify the thesis: if the `same_model` τ_a CI covers 0, or τ_a fails to decrease from `same_model` to `different_vendor`, or the composition gap is approximately zero, the claim is unsupported and the null is reported. We report against that criterion in Section 10.2.2, including one registered comparison that does *not* survive.

Missions and scoring. Missions are drawn from six seeded generators spanning retail and financial agent tasks. Every mission is scored by *deterministic gold code*, never by an LLM judge: order arithmetic, refund-policy application, and promotional-cap enforcement on the retail side; transaction-limit checks, watchlist screening, and mandatory-disclaimer presence on the financial side. The mission-to-task assignment is a fixed SHA-256 hash of the mission identifier, which makes each arm’s task sequence deterministic and reproducible. It does *not* make the realised task mixtures identical across arms: identifiers embed the condition name (Section 11.3), so each arm is an independent deterministic draw from the same generator distribution rather than a paired design. At $n = 6000$ per confirmatory arm the resulting imbalance is absorbed by the bootstrap; for the under-run breadth arms it is a real limit.

Each agent invocation yields a hard verdict and a soft score. The hard verdict $h_i \in \{0, 1\}$ records whether the agent satisfied every hard clause of its contract; the soft score $\sigma_i \in [0, 1]$ records graded compliance. All dependence estimates in Section 10.2 are computed on the *hard* verdict, so a “failure” means a definite contract violation, not a low score. Graph success Y_G is the motif’s composition rule applied to the component hard verdicts.

Frozen sampling. Every call uses `temperature = 0.2`, `top_p = 1.0`, and `max_output_tokens = 160`, with client-side prompt truncation at 3200 characters. These values were fixed at registration and never varied.

Conditions. The manipulated variable is how much the two scored agents share. In `same_model` both run `mistral-small-24b`; in `same_vendor` the second agent is swapped to `ministral-8b` (same vendor, smaller model); in `different_vendor` it is swapped to `gemma-3-12b-it`. The first agent is held at `mistral-small-24b` throughout, so the contrast isolates the substitution.

Motifs. Three topologies are measured. `series2` is a two-agent handoff $A \rightarrow B$ where both must satisfy their contract for $Y_G = 1$. `parallel2` runs two branches into a deterministic merge. `quorum2of3` runs three workers into a deterministic 2-of-3 aggregator. The aggregator and merge nodes are deterministic code, not models, and are excluded from dependence estimation.

Scale. The campaign is 30,820 scored missions across 12 arms, recorded in 13 execution logs. The counting rule matters and we state it: the `quorum3of4` arm was executed twice, so its two logs share 212 mission identifiers, which we deduplicate with the later pass winning; that arm therefore contributes 717 missions rather than the 929 records on disk, and 717 is the n used throughout Section 10.3. The primary confirmatory arms are the three `series2` conditions at $n = 6000$ each.

Estimators. For two agents with failure indicators $F_A = 1 - h_A$ and $F_B = 1 - h_B$ we form the 2×2 co-failure table with cells $n_{11}, n_{10}, n_{01}, n_{00}$ and corresponding proportions $p_{11}, p_{10}, p_{01}, p_{00}$. We report five statistics, and the distinction between the first three and the last two carries the

analysis:

$$J = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}, \quad (24)$$

$$\tau_a = 2(p_{11}p_{00} - p_{10}p_{01}), \quad (25)$$

$$\phi = \frac{p_{11}p_{00} - p_{10}p_{01}}{\sqrt{p_A(1-p_A)p_B(1-p_B)}}, \quad (26)$$

$$\log \text{OR} = \log \frac{(n_{11} + \frac{1}{2})(n_{00} + \frac{1}{2})}{(n_{10} + \frac{1}{2})(n_{01} + \frac{1}{2})}, \quad (27)$$

$$Q = \frac{\text{OR} - 1}{\text{OR} + 1}, \quad (28)$$

where $p_A = p_{11} + p_{10}$ and $p_B = p_{11} + p_{01}$ are the marginal failure rates. These are, in order, the overlap coefficient of Jaccard (1912), the binary form of Kendall (1938)’s τ , the phi coefficient, the log odds ratio, and Yule (1900)’s Q . Equation (24) is the overlap of the two failure *sets*; the n_{00} cell is deliberately excluded so that missions neither agent failed cannot dilute the statistic. Equation (27) uses the Haldane–Anscombe $+\frac{1}{2}$ correction, which keeps the estimator finite when a cell is empty.

J , τ_a , and ϕ are all sensitive to the marginals: τ_a is twice the covariance of the two indicators and therefore cannot reach ± 1 unless both marginals equal $\frac{1}{2}$, and J charges every mission that exactly one agent failed to its denominator. $\log \text{OR}$ and Q are invariant to the marginals under row and column scaling. Section 10.2.3 shows this distinction is not pedantic: it decides whether the registered monotonicity claim survives.

Uncertainty. Confidence intervals are percentile bootstrap (Efron, 1979) with $B = 2000$ resamples at $\alpha = 0.05$, seeded at 20260813. The registration specifies a cluster bootstrap clustered by mission. In the logs `cluster_id` is one-to-one with `mission_id` (6000 distinct clusters over 6000 missions in each primary arm), so the cluster bootstrap reduces exactly to an i.i.d. bootstrap over missions. We state this rather than leave it implicit, and Section 10.7 measures the design effect (Kish, 1965) directly instead of assuming it: the empirical DEFF is 1.00–1.16.

Hypothesis H2 is a statement about *differences* between arms, so we test it on the bootstrap distribution of the arm contrast rather than by inspecting whether two intervals overlap. Overlapping marginal intervals do not imply a non-significant difference, and the contrast test is the correct instrument.

10.2 E1: model sharing induces correlated contract failures

10.2.1 Dependence is present, large, and positive in every arm

Table 2 gives the primary result. In every condition the two agents’ hard failures are strongly positively associated. The weakest association we measure anywhere in E1 is $\log \text{OR} = 2.92$ ($\text{OR} \approx 18.5$) in `parallel2/different_vendor`; the strongest is $\log \text{OR} = 6.66$ ($\text{OR} \approx 784$) in `series2/same_model`. Every interval excludes independence by a wide margin. H1 as registered concerns the `same_model` arm, and it is supported there. We note in passing that the association is also positive and significant in the other two arms; that is a descriptive observation, not a registered result, and it is a reminder that substituting a model reduces dependence without eliminating it.

The practical reading of `same_model` is blunt. Agent *A* fails 39.4% of missions and agent *B* fails 37.2%. Under independence the two would co-fail on 14.6% of missions; they co-fail on 36.3%. Of the 2418 missions on which at least one agent failed, both failed on 2177 — 90.0%. Replacing one agent with a second instance of the same model buys almost no independent evidence.

Table 2: E1 primary arms. Motif **series2**, pair (**node_a**, **node_b**), $n = 6000$ per arm. Cells are co-failure counts on the hard verdict. Brackets are percentile bootstrap 95% CIs, $B = 2000$.

	same_model	same_vendor	different_vendor
$n_{11}/n_{10}/n_{01}/n_{00}$	2177/189/52/3582	2289/66/757/2888	1987/289/225/3499
p_A	0.3943	0.3925	0.3793
p_B	0.3715	0.5077	0.3687
$ p_A - p_B $	0.0228	0.1152	0.0107
<i>marginal-sensitive</i>			
J	0.9003 [0.8885, 0.9120]	0.7355 [0.7200, 0.7508]	0.7945 [0.7795, 0.8099]
τ_a	0.4327 [0.4257, 0.4398]	0.3645 [0.3564, 0.3726]	0.3826 [0.3739, 0.3911]
ϕ	0.9161 [0.9057, 0.9261]	0.7465 [0.7318, 0.7613]	0.8173 [0.8026, 0.8321]
<i>marginal-free</i>			
log OR	6.6645 [6.3827, 7.0015]	4.8774 [4.6431, 5.1529]	4.6685 [4.4978, 4.8593]
Q	0.9975 [0.9966, 0.9982]	0.9849 [0.9809, 0.9885]	0.9814 [0.9780, 0.9846]

Table 3: E1 arm contrasts, all three motifs. Arms abbreviated **SM** **same_model**, **SV** **same_vendor**, **DV** **different_vendor**. Entries are the point difference with a 95% bootstrap CI on the difference; **SIG** marks an interval excluding zero. The SV – DV row is the registered comparison that does not survive. Only **series2** is confirmatory; the other two motifs are secondary.

motif	contrast	J	log OR	verdict
series2	SM – SV	0.1648 [0.1452, 0.1838]	1.787 [1.396, 2.202]	SIG
	SM – DV	0.1058 [0.0864, 0.1245]	1.996 [1.663, 2.361]	SIG
	SV – DV	−0.0589 [−0.0814, −0.0369]	0.209 [−0.089, 0.548]	conflict
quorum2of3	SM – SV	0.2079 [0.1612, 0.2552]	1.219 [0.579, 1.826]	SIG
	SM – DV	0.1896 [0.1439, 0.2360]	1.718 [1.222, 2.268]	SIG
	SV – DV	−0.0183 [−0.0673, 0.0279]	0.499 [−0.037, 1.115]	n.s.
parallel12	SM – SV	0.2124 [0.1759, 0.2521]	1.649 [1.154, 2.143]	SIG
	SM – DV	0.2230 [0.1844, 0.2619]	2.125 [1.707, 2.607]	SIG
	SV – DV	0.0106 [−0.0308, 0.0511]	0.476 [0.048, 0.943]	conflict

This is not merely interface propagation. **series2** is a handoff, so a failure of A degrades the input B receives, and co-failure there could in principle be explained without appeal to shared failure modes at all. Two features of the design separate the mechanisms. First, **parallel12** and **quorum2of3** are *not* handoffs — their model nodes receive the mission independently and never see one another’s output — and both reproduce the effect (Table 3), which propagation cannot explain. Second, the control pair of Section 10.2.4 sits at $\log \text{OR} \approx 5.1$ with no channel between its two workers. Propagation may well contribute to the **series2** magnitude and we do not claim to have partitioned the two mechanisms; the arm *contrasts*, which is what H2 concerns, are measured within identical topologies and so are unaffected by it.

10.2.2 The registered monotonicity claim splits

Table 3 tests H2 directly on the bootstrap distribution of each arm contrast. The verdict is not uniform, and we state it plainly: **H2 holds at the model-sharing level and fails at the vendor level.**

Two findings are robust. Sharing the *same model* produces significantly more co-failure than either alternative, on all five statistics, in all three motifs, with no exceptions. That is six independent significant contrasts in the registered direction.

The third comparison does not replicate. Moving from `same_vendor` to `different_vendor` produces no consistent change: in `quorum2of3` every statistic is non-significant; in `series2` and `parallel2` the marginal-sensitive and marginal-free statistics disagree in sign or significance. We therefore do not claim a three-level ordering. Vendor identity, as manipulated here, is not a reliable predictor of failure correlation once model identity differs.

Against the registered falsification criterion, the thesis survives: the `same_model` τ_a CI excludes zero, and τ_a decreases significantly from `same_model` (0.4327) to `different_vendor` (0.3826), difference 0.0500, CI [0.0389, 0.0615]. The registered falsification test was stated on exactly this contrast and it passes. The intermediate rung is where the ordering breaks, and the registration did not make the intermediate rung a falsification condition.

10.2.3 Why the marginal-sensitive statistics reverse

In `series2` the Jaccard overlap is *lower* for `same_vendor` (0.7355) than for `different_vendor` (0.7945), and the difference is significant. Read naively this says swapping to a different vendor increases correlated failure, which inverts the mechanism the experiment was built to test. It is an artifact of the marginals, and the artifact is measurable rather than conjectural.

`ministral-8b` is a weaker model than `mistral-small-24b` and fails far more often. In the `same_vendor` arm the two marginal failure rates are $p_A = 0.3925$ and $p_B = 0.5077$, a gap of 0.1152; in `different_vendor` the gap is 0.0107. That asymmetry lands in the n_{01} cell: agent B fails alone on 757 missions in `same_vendor` against 225 in `different_vendor`. Because J carries n_{01} in its denominator, an arm whose two agents fail at different rates is charged a penalty that has nothing to do with whether their failures are associated.

The pattern holds across motifs. The marginal gap $|p_A - p_B|$ is largest in `same_vendor` in every motif (0.115, 0.202, 0.197) and small in `same_model` (0.023, 0.006, 0.006), and it is exactly the arms with large gaps whose J , τ_a , and ϕ are depressed. Substituting a model changes two things at once — how often that agent fails, and how much its failures align with its partner’s — and only the second is the quantity H2 is about. The marginal-free statistics separate them: on log OR and Q the `series2` reversal disappears and the contrast is non-significant.

We therefore report H2 on log OR, and report J alongside it because J is the quantity a practitioner actually feels — it is the fraction of observed failures that a redundant agent fails to catch. Both are correct statistics; they answer different questions, and the paper needs both.

10.2.4 An internal negative control

The `quorum2of3` design contains an internal control. Only `worker_1` is substituted across arms: `worker_0` and `worker_2` both run `mistral-small-24b` in *all three* conditions. The pair (`worker_0`, `worker_2`) is therefore a same-model pair whose composition never changes while the arm label does. If the arm effects in Table 3 were driven by anything other than the model substitution — task mix drifting between runs, ordering, wall-clock effects, provider-side variation — this pair would move with the arm label. It must not move, and it does not.

Table 4 reports the result. Across five statistics and three pairwise contrasts — fifteen tests — not one interval excludes zero. The control pair’s dependence sits at $J \approx 0.83$ – 0.85 and log OR ≈ 5.07 – 5.32 in every arm, which is where the *manipulated* pair sits in its `same_model` condition ($J = 0.8251$, log OR = 4.957). A same-model pair measures as a same-model pair regardless of what the rest of the graph is doing. That is convergent validity for the estimator and, jointly with the fifteen nulls, evidence that the arm effects in Table 3 are caused by the model substitution rather than by any arm-level confound.

Table 4: Negative control. Pair (`worker_0`, `worker_2`) in `quorum2of3` is `mistral-small-24b` $\times$ `mistral-small-24b` in every arm. All 15 arm contrasts across the five statistics are non-significant.

	same_model	same_vendor	different_vendor
n	1281	1274	1388
$ p_A - p_B $	0.0008	0.0141	0.0072
J	0.8345 [0.8026, 0.8654]	0.8439 [0.8135, 0.8737]	0.8508 [0.8231, 0.8810]
log OR	5.068 [4.679, 5.536]	5.225 [4.820, 5.712]	5.324 [4.945, 5.833]
contrasts, J :	-0.009 [-0.051, 0.034]	-0.016 [-0.058, 0.026]	-0.007 [-0.049, 0.035] (all n.s.)
contrasts, log OR:	-0.157 [-0.740, 0.471]	-0.256 [-0.864, 0.362]	-0.099 [-0.727, 0.540] (all n.s.)

Table 5: E2 floor lifting. $m = 4$, $n = 717$, observed all-success 0.5328. The pre-allocated Bonferroni row spends the confidence budget over the larger family $J_{\max} = 14$ regardless of how many moments are constrained, which makes the certified floor monotone in the moment set by construction.

quantity	allocation	$J = 10$	$J = 14$
sharp point-moment interval	—	[0.5077, 0.5565]	[0.5272, 0.5342]
width		0.0488	0.0070
certified floor	used-set	0.2455	0.4116
certified floor	pre-allocated	0.2357	0.4116
certified upper	used-set	0.6087	0.5955

10.2.5 What E1 supports

Model sharing is the operative variable. Two instances of one model co-fail on 90% of the missions either one fails; substituting a different model reduces that overlap significantly and reduces the odds-ratio association by roughly 1.8–2.1 log units, consistently across three topologies. Vendor identity, holding model identity different, does not reliably matter. A redundant agent running the same model as the agent it is meant to check supplies far less independent evidence than a naive composition calculation credits it with, and Section 10.4 quantifies what that does to a reliability guarantee.

10.3 E2: enriching the moment set lifts the certified floor

E1 establishes that components fail together. A practitioner still has to certify a number. E2 measures what the moment-set certificate of Section 6 buys over the assumption-free bound on real data.

Data. The four-worker `quorum3of4` `same_model` arm, $m = 4$ scored workers over $n = 717$ missions. The arm was executed in two passes; we take the union of both logs deduplicated by mission identifier, with the later pass winning on the 212 missions present in both. The observed all-success rate is 0.5328.

Procedure. We compute two things at two moment sets. The *sharp point-moment interval* is the LP over all joint laws matching the empirical moments exactly — it isolates how much the moment family alone identifies, with no sampling uncertainty. The *certified floor* is the LP over the Clopper–Pearson box around those moments at family-wise $\eta_{\text{conf}} = 0.05$, which is what may actually be reported. The moment sets are $J = 10$ (marginals and pairwise co-success, $\binom{4}{1} + \binom{4}{2}$) and $J = 14$ (adding all four triple co-success moments).

Table 6: E3 coverage of the *true* all-success probability, as hits/200 with a 95% Clopper–Pearson interval. $n_{\text{boot}} = 500$, $\eta_{\text{conf}} = 0.05$, seed 20260812. Nominal coverage is 0.95. The Tier-1 floor covered in all 200 replications of every cell and is omitted from the interval columns.

	model floor				
n	adversarial ($\Delta = 0.0617$)		control ($\Delta = 0$)		Tier 1
250	72/200 = 0.36	[0.294, 0.431]	190/200 = 0.95	[0.910, 0.976]	1.00
500	14/200 = 0.07	[0.039, 0.115]	192/200 = 0.96	[0.923, 0.983]	1.00
1000	6/200 = 0.03	[0.011, 0.064]	192/200 = 0.96	[0.923, 0.983]	1.00
2000	2/200 = 0.01	[0.001, 0.036]	188/200 = 0.94	[0.898, 0.969]	1.00

Table 5 reports the result. Adding four triple moments narrows the sharp identified interval from width 0.0488 to 0.0070, an 85.7% reduction, and lifts the certified floor from 0.2455 to 0.4116 — 16.6 percentage points of reliability that the pairwise certificate cannot claim and the triple certificate can, on identical data.

Two details matter for honest reporting. First, the certified floor at $J = 14$ (0.4116) remains well below the observed rate (0.5328), because it is a distribution-free lower confidence bound over every joint law consistent with the moment box, not an estimate. It is a floor, and it is meant to be conservative. Second, the used-set allocation is tighter at $J = 10$ (0.2455 against 0.2357) but its monotonicity in the moment set is empirical rather than guaranteed: adding a moment widens every interval, so a richer moment family can in principle certify less. The pre-allocated allocation pays 0.0098 at $J = 10$ to buy monotonicity by construction. We report both rather than quietly choosing the flattering one.

10.4 E3: coverage of a model-based floor collapses under misspecification

Theorem 4.2 predicts that a bootstrap bound on a fitted model’s functional stops covering the truth once the model is misspecified. E3 measures that against a control, because the claim is not that the Gaussian floor is always wrong — it is that the floor cannot tell you when it is.

Two arms. Both sample $m = 3$ components at marginal 0.6. The *control* arm draws from the equicorrelated Gaussian one-factor law itself, for which the fitted family is correctly specified and $\Delta = 0$; the floor should hold at nominal there, and if it did not, any collapse in the other arm would indict the estimator rather than the misspecification. The *adversarial* arm draws from the witness of Example 4.2: the LP-minimising law, which has *identical* marginals and *identical* pairwise co-success moments and an all-success probability of 0.330475 against the Gaussian law’s 0.392144, so $\Delta = 0.061670$. Its five non-zero cell masses are listed in Example 4.2. Both constants come from the same deterministic quadrature and LP used in Example 4.2, not from simulation. No procedure restricted to marginal and pairwise data can distinguish the two.

Each arm runs 200 replications at each of four sample sizes, computing the Gaussian model floor with 500 bootstrap resamples at $\eta_{\text{conf}} = 0.05$ and recording how often it lies at or below the *true* all-success probability. The Tier-1 floor of Theorem 5.2 is computed on the same draws.

In the control arm the model floor holds between 0.94 and 0.96 at every sample size, with all four intervals covering the nominal 0.95 — a correctly specified model behaving correctly across a factor of eight in n . In the adversarial arm coverage falls from 0.36 at $n = 250$ to 0.01 at $n = 2000$, an upper confidence limit of 0.036: by two thousand missions the certificate sits above the true reliability in 198 of 200 replications. The Tier-1 floor covers in all 1600 replications across both arms.

We report the counts rather than the rounded rates because a small count is not a zero. Two hits in two hundred replications bounds coverage above by 0.036; it does not establish that

Table 7: E4 empirical type-I crossing rate under the null. $p_0 = 0.8$, $\alpha = 0.05$, 8000 streams $\times$ 500 missions, seed 42. The three-sigma Monte Carlo band on a proportion of size α at this n_{streams} is 0.0073.

λ	0.125	0.3125	0.625	0.875	1.125	1.2375
crossing rate	0.0013	0.0387	0.0471	0.0447	0.0441	0.0442
mean first crossing	403.9	246.3	84.0	40.2	19.1	14.6
$\leq \alpha$?	yes	yes	yes	yes	yes	yes

coverage *is* zero, and a run reporting 0/200 would bound it above by 0.018 rather than prove exact failure.

Three points follow. The two arms are *indistinguishable* on the evidence the model consumes, so the collapse is not a case a diagnostic on marginals or pairwise moments could detect and exclude. Coverage *decreases* as n grows, inverting the usual reassurance that more data makes an interval safer: the interval narrows, but around the fitted functional rather than the truth. We report that decrease as an empirical observation — Theorem 4.2 claims only the limit, since tightness does not entail monotonicity (Section A.6). And the Tier-1 floor is conservative here, 1.00 against a nominal 0.95; that conservatism is the price of the immunity, as Corollary 4.3 describes.

A correction to the previous version of this experiment. The v2 preprint reported a collapse from a simulation that drew from the Gaussian law. That design cannot exhibit the effect: with the model correctly specified $\Delta = 0$, Theorem 4.2 does not apply, and re-running it returns the control column above. Obtaining Table 6 required a law that is misspecified *and* moment-indistinguishable, which Theorem 6.1 supplies as the LP minimiser. We record the correction rather than quietly restate the earlier number.

10.5 E4: the anytime-valid certificate holds its type-I error

A team that watches a reliability dashboard and stops when it first looks good is running an optional-stopping experiment, and a fixed- n confidence interval does not survive that. The e-process of Section 7 is valid under arbitrary stopping. E4 measures whether the shipped implementation attains the guarantee.

Procedure. We simulate 8000 independent null streams of 500 missions each at the null boundary $p_{\text{true}} = p_0 = 0.8$, with $\alpha = 0.05$ and seed 42, and record the fraction of streams whose wealth ever reaches $1/\alpha$. Ville’s inequality bounds that fraction by α for *every* predictable betting fraction λ , not merely for a tuned one, so we sweep the whole admissible range $\lambda \in (0, 1/p_0)$ and report the curve. A single λ would not distinguish a valid certificate from a lucky one.

The worst rate over the six prespecified fractions is 0.0471, below the nominal 0.05, and every individual rate is inside the Monte Carlo band. Six points cannot establish a supremum over a continuum; the guarantee for every predictable λ is Theorem 7.1, and this experiment checks that the shipped implementation attains it at fractions spanning the admissible range. The bound is attained most tightly at $\lambda = 0.625$, which is exactly the SPRT-optimal bet $\lambda^* = (p_1 - p_0)/(p_0(1 - p_0))$ for the alternative $p_1 = 0.9$ — the certificate is tightest where the betting fraction is best matched to the alternative, and conservative elsewhere, which is the expected shape.

SPRT recovery. Setting $\lambda^* = (p_1 - p_0)/(p_0(1 - p_0))$ makes the betting factor identical to the Bernoulli likelihood ratio. We verify this against the shipped `from_sprt` constructor rather than against a second closed form: across five (p_0, p_1) pairs and three mission sequences, the

Table 8: E5 cross-backend replication, motif **series2**. The upstream agent changes provider; the downstream agent is **mistral-small-24b** throughout. Marginal-sensitive and marginal-free statistics move in *opposite* directions.

arm	n	p_A / p_B	J [95%]	log OR [95%]
different_vendor (primary)	6000	0.3793 / 0.3687	0.7945 [0.7786, 0.8093]	4.669 [4.499, 4.854]
different_vendor_meta	635	0.0016 / 0.0157	0.100 [0.000, 0.333]	5.286 [3.758, 6.550]
different_vendor_grok	1901	0.0174 / 0.0594	0.292 [0.214, 0.381]	7.306 [6.898, 7.704]

maximum absolute deviation between the accumulated log-wealth and the accumulated Bernoulli log-likelihood-ratio is 0.0 — exact in floating point, not merely close.

Monotone latch. Certification is a statement about the path, not the present: once the wealth has crossed $1/\alpha$, no subsequent evidence retracts it. We drive a process across the threshold at mission 14 (peak log-wealth 13.27) and then feed it 340 consecutive failures, collapsing log-wealth to -1552.49 . The certificate remains issued throughout. This is correct — the crossing happened, and Ville bounds the probability that it ever happens under the null — and it is a property implementations get wrong by recomputing certification from current wealth.

10.6 E5: the dependence result replicates across backends

The three primary arms of E1 all run through a single inference provider. If the measured association were an artifact of that provider’s serving stack — batching, caching, a shared sampler — it would not appear when the upstream agent runs somewhere else entirely. The preregistration therefore fixed two cross-backend breadth arms, both on the **series2** motif, in which the upstream agent is served by a different provider and the downstream agent remains **mistral-small-24b**. The Meta-backed arm runs **muse-spark-1.2-contributor** upstream; the Grok-backed arm runs **grok-4.5**. Both arms change the upstream model as well as the serving backend, so we label them by backend as shorthand, not as a claim that the backend is the only thing varying.

Registered deviation, and attrition. Both arms were registered at $n = 2000$ and both yield fewer scored missions: 635 for the Meta arm and 1901 for the Grok arm. The shortfall is *output attrition*, and its mechanism matters more than its size. Meta returned 1365 unusable responses — 1352 empty completions under the frozen reasoning and token settings of Section 10.1, plus 13 rate limits; Grok returned 99, almost all rate limits.

Meta’s 68.25% attrition may therefore be *outcome-dependent*: a model returning no text is plausibly one that would also have failed the contract, so the surviving 635 missions are a selected subset and their complete-case estimate is descriptive only. Grok’s 5% loss is transport error and carries no such selection. We read Grok as usable secondary evidence and Meta as directional; both were registered as breadth arms rather than confirmatory tests in any case.

Positive association appears on both breadth arms, and on the Grok arm it appears cleanly. That arm scores 1901 of its 2000 registered missions, loses the remainder to transport errors that carry no outcome selection, and returns log OR = 7.306 with a 95% lower bound of 6.898 against 0 under independence. A serving-stack artifact confined to one provider cannot produce that. The Meta arm points the same way, with a lower bound of 3.758, but its attrition is selective and it corroborates rather than replicates.

We do *not* read the breadth arms as strengthening the sharing story, and the reason is worth stating because the table invites the opposite reading. The Grok arm’s log OR of 7.306 is the largest association anywhere in this paper and it is a *cross-vendor* pair, nominally exceeding **same_model**’s 6.665. That comparison is not admissible. The marginal failure rates differ by

more than an order of magnitude between the breadth and primary arms (0.017 and 0.059 against 0.379 and 0.369), the upstream model is different, and the Grok estimate rests on 33 concordant and 80 discordant missions with $n_{10} = 0$, so it depends on the Haldane–Anscombe correction. Odds-ratio invariance to row and column scaling does not make estimates from such different regimes comparable in magnitude. The breadth arms establish that the association *exists* on other backends; they do not rank sharing conditions, and any cross-arm magnitude comparison with Table 2 should be resisted — including one that appears to favour our thesis.

The Jaccard column tells the opposite story — 0.100 and 0.292 against the primary arm’s 0.794 — and the reason is the mechanism of Section 10.2.3 appearing again on independent data. The upstream agents here are far stronger: `grok-4.5` fails 1.74% of missions and `muse-spark-1.2` fails 0.16%, against `mistral-small-24b`’s 37.9% in the primary arm. Because `series2` is a handoff, a cleaner upstream output makes the downstream task easier, and the downstream agent’s failure rate falls from 36.9% to 5.9% and 1.6% respectively. Both marginals collapse, and J — which is bounded by the marginals — collapses with them, while the marginal-free log OR rises. The two statistics are not in conflict; they are measuring different things, exactly as Section 10.2.3 argued, and E5 is an independent replication of that methodological point on data from three different providers.

Two limits are worth stating rather than burying. The Meta arm records a single co-failure ($n_{11} = 1$), so its Jaccard confidence interval is $[0.000, 0.333]$ and the point estimate carries essentially no information; only its log OR is usable, and even that rests on nine discordant missions. And $n_{10} = 0$ in both arms — the upstream agent never failed without the downstream agent also failing. With only 1 and 33 upstream failures observed this is unsurprising under any model, so we do not read deterministic failure propagation into it. It does mean the odds ratio in these arms depends on the Haldane–Anscombe correction, which is why we report it with the correction stated rather than as an unadorned ratio.

Because mission identifiers embed the condition name and the task is a fixed hash of the identifier, the arms share no missions: task draws are independent across arms, from the same generator distribution. The comparison is between distributions, not between paired tasks.

What E5 establishes is narrow, and it is solid. On the Grok arm — 1901 missions, non-selective attrition, an interval nowhere near independence — cross-agent failure dependence is present on infrastructure the primary arms never touched, and the marginal-sensitivity mechanism of Section 10.2.3 reproduces there on data the primary arms did not generate. One clean external replication and one directional corroboration is what two breadth arms can honestly buy, and it is enough to rule out the single-provider explanation.

10.7 E6: the i.i.d. assumption, tested rather than assumed

Every interval in E1 and every floor in E2 assumes missions are independent. That assumption is testable, and E6 tests it two ways instead of asserting it.

Serial dependence. For each of the 19 agent-arm combinations we compute the lag- k autocorrelation of the failure indicator in execution order and the induced design effect $\text{DEFF} = 1 + 2 \sum_{k=1}^{10} (1 - k/n) \rho_k$. The largest first-order autocorrelation anywhere is $|\rho_1| = 0.0546$ and the largest design effect is $\text{DEFF} = 1.1482$; many arms return $\text{DEFF} < 1$, indicating mild negative serial dependence. At $\text{DEFF} = 1.15$ the E1 confidence intervals are optimistic by a factor of $\sqrt{1.15} \approx 1.07$, which does not change any verdict in Table 3: the smallest significant contrast there has a margin far exceeding 7% of its width.

Floor sensitivity. We then concede a design effect we do not measure — $\text{DEFF} = 1.5$ — and ask what the certified floor costs. The measurement has to isolate the width of the Clopper–Pearson intervals: subsampling missions to $n/1.5$ would also perturb the empirical moments,

Table 9: E6 certified-floor sensitivity at a conceded $\text{DEFF} = 1.5$, $\eta_{\text{conf}} = 0.05$. Movement scales inversely with n , as Clopper–Pearson width must.

case	m	$n \rightarrow n_{\text{eff}}$	$\text{floor}(n)$	$\text{floor}(n_{\text{eff}})$	Δ (pp)
series2/same_model	2	6000 $\rightarrow$ 4000	0.5817	0.5782	0.35
series2/same_vendor	2	6000 $\rightarrow$ 4000	0.4658	0.4622	0.36
series2/different_vendor	2	6000 $\rightarrow$ 4000	0.5678	0.5641	0.37
quorum2of3/same_model	3	1281 $\rightarrow$ 854	0.4262	0.4000	2.62
quorum2of3/same_vendor	3	1274 $\rightarrow$ 849	0.2609	0.2364	2.44
quorum2of3/different_vendor	3	1388 $\rightarrow$ 925	0.3067	0.2827	2.40
quorum3of4/same_model	4	717 $\rightarrow$ 478	0.4116	0.3847	2.69

Table 10: Single-agent results carried forward from the v1 framework paper (Bhardwaj, 2026): 1980 sessions, 200 scenarios, 7 models, 6 vendors. These are not re-measured in this work.

quantity	result	basis
soft violations detected per session	5.2–6.8	missed entirely by uncontracted baselines $p < 0.0001$, Cohen’s $d = 6.7$ –33.8
hard-constraint compliance	88–100%	across all 7 models
behavioral drift over extended sessions	$D^* < 0.27$	bounded, per the drift theorem
recovery success	100% frontier	17–100% across all models
enforcement overhead	$< 10\text{ ms}$	per action

and the resulting movement would be the sum of interval widening and sampling noise reported as if it were the former. We instead hold the joint cell distribution fixed, rescale the cell counts to $n_{\text{eff}} = n/1.5$ by largest-remainder rounding, and recompute. Only n changes.

The floor moves at most 2.69 percentage points under a conceded design effect of 1.5, which is $1.31\times$ the largest design effect measured (1.148). The movement is larger at smaller n — 0.35 pp at $n = 6000$ against 2.69 pp at $n = 717$ — consistent with Clopper–Pearson width shrinking as n grows, though these rows vary m and the moment family as well as n and so do not isolate a rate. The certificate is not sensitive to the i.i.d. assumption at the magnitudes the data support.

10.8 Carried-forward single-agent evidence

E1–E6 measure composition. They take for granted that a behavioral contract can be specified, enforced, and measured on a *single* agent at acceptable cost, which is the result established in the v1 framework paper (Bhardwaj, 2026) and not re-run here. We restate it because the compositional claims are meaningless without it: a dependence-aware bound over components whose individual contracts could not be enforced would certify nothing.

The v1 evaluation ran 1980 sessions over AGENTCONTRACT-BENCH, a benchmark of 200 scenarios across 7 models from 6 vendors. Four results carry forward.

The effect sizes in the first row are large enough to warrant a caution rather than a boast. Cohen’s d between 6.7 and 33.8 reflects a comparison in which the baseline detects approximately zero soft violations by construction — an uninstrumented agent has no mechanism to report a soft-constraint deviation — so the contrast measures the value of instrumentation, not a narrow improvement over a competing detector. We restate it in those terms here because the v1 abstract’s phrasing invites the stronger reading.

Carrying prior-version evidence forward rather than re-running it is deliberate and bounded: Table 10 is cited, not claimed, and no result in Section 10.2–Section 10.7 depends on re-deriving it.

Table 11: Summary of the six experiments, their registered status, and their verdicts. “Confirmatory” means the hypothesis, sample size, and analysis were fixed before outcomes existed.

	status	verdict
E1	confirmatory (series2)	Registered H1 — positive dependence in same_model — is supported, and positive dependence holds in every other arm and motif as well. Registered H2’s full three-level ordering is not supported : same_model exceeds both substituted conditions in 6/6 contrasts, but same_vendor does not exceed different_vendor . The sharing-level/vendor-level split is an exploratory decomposition of a registered claim that fails as a whole. An unmanipulated same-model control pair returns 15/15 nulls.
E2	descriptive	Enriching the moment set from $J=10$ to $J=14$ narrows the identified interval by 85.7% and lifts the certified floor by 16.6 pp.
E3	descriptive	Under a moment-indistinguishable misspecification the model floor’s coverage falls $0.36 \rightarrow 0.01$ as n grows; a correctly-specified control holds at 0.94–0.96. Tier 1 covers in all 1600 replications.
E4	descriptive	Empirical type-I $\leq \alpha$ for every admissible λ , worst case 0.0471. SPRT recovered exactly; certification latch holds.
E5	secondary breadth (registered, under-ran)	Positive association with intervals excluding independence on two further backends; magnitudes are not comparable across arms. Jac-card collapses because the marginals collapse — independent replication of the Section 10.2.3 mechanism.
E6	ablation	Worst measured DEFF = 1.148; at a conceded DEFF = 1.5 the certified floor moves ≤ 2.69 pp.

10.9 What the evaluation establishes

Three claims survive the evaluation.

Shared models produce correlated contract failures, and the effect is large. Two instances of one model co-fail on 90.0% of the missions on which either fails. The association is significant in every arm of every motif, and it replicates on two additional inference backends. A composition bound that treats such components as independent is not making a small approximation.

The operative variable is the model, not the vendor. Substituting a different model reduces association significantly and consistently — six of six contrasts across three topologies. Substituting a different *vendor* while the model already differs does not, and we report that null rather than presenting a three-level ordering the data do not support. For a practitioner choosing redundancy, this says the useful axis is model identity; sourcing two different models from one vendor is not obviously worse than sourcing from two.

Certification can be tightened without assuming a copula. Enriching the moment set lifts the certified floor by 16.6 percentage points on identical data, and the resulting certificate is insensitive to the i.i.d. assumption at the magnitudes the data support.

Two limits belong here rather than only in Section 11.3. The substitution manipulation

confounds model *identity* with model *capability*: `ministral-8b` is both a different model and a weaker one, so E1’s arm effects cannot separate “different inductive bias” from “different competence.” Section 10.2.3 shows this is what drives the marginal-sensitive reversal, and the marginal-free statistics are the correct instrument, but no design here isolates the two. And every arm draws its own missions — identifiers embed the condition name, and the task is a fixed hash of the identifier — so all comparisons are between independent draws from a shared generator distribution, not between paired tasks. At $n = 6000$ per primary arm that is adequate; for the under-run breadth arms of Section 10.6 it is a real limit on precision.

11 Discussion

11.1 What the practitioner should do differently

Three consequences follow directly from Section 10.

Do not multiply reliabilities across components that share a model. Proposition 4.1 makes the error signed, and Section 10.2 makes it large: at $\phi = 0.92$ the joint-failure probability of a two-agent series is far above the independent product. A pipeline certified by multiplication is certified against a dependence structure that the data reject.

Choose redundancy on model identity, not vendor. Six of six contrasts show that substituting a different model reduces association significantly. The `same_vendor` versus `different_vendor` contrast does not replicate in any motif. We state the negative result and stop there: this evaluation substituted exactly one same-vendor model and one cross-vendor model, so it cannot support a procurement policy about vendor diversity in general. It supports only the narrow claim that vendor identity did not predict failure correlation here once model identity already differed.

Report the moment set alongside the floor. A certified floor is meaningless without the family it was computed over: the same data yield 0.2455 at $J=10$ and 0.4116 at $J=14$ (Section 10.3). Enriching the family is the cheapest available tightening, and Proposition 6.2 shows how to do it without losing monotonicity.

11.2 Limitations

The LP is exponential in m . Equation (18) has 2^m variables. Every motif here has $m \leq 4$ and the program solves in milliseconds, but a fifty-stage pipeline is out of reach and would require certifying sub-compositions and composing certificates, at a cost in tightness we have not quantified.

Aggregators are assumed deterministic. Merge and quorum nodes are code, not models, so they carry no contract and contribute no failure correlation. A system whose aggregator is itself an LLM — an increasingly common design — introduces a third correlated component that this analysis does not model.

Drift dynamics are carried, not tested. Theorem 3.1 is stated and proved, and it is inherited from v1 rather than re-measured here. No experiment in Section 10 estimates α , γ , or σ from data, so the design criterion $\gamma > \alpha$ is a theoretical statement in this paper.

The certificate is per-distribution. By Definition 5.1 a guarantee holds over the mission distribution, model versions, and topology under which it was obtained. Model upgrades invalidate it. Nothing here says how quickly a certificate decays as a deployment drifts from its certification conditions.

11.3 Threats to Validity

Model identity is confounded with model capability. The substitution replaces `mistral-small-24b` with `ministral-8b` or `gemma-3-12b-it`, which differ in both training

lineage and competence. An arm effect could therefore reflect differing failure *rates* rather than differing failure *modes*. This is the most serious threat to E1’s interpretation, and it is the mechanism behind the reversal diagnosed in Section 10.2.3. Two things limit it. The marginal-free statistics are invariant to the marginal rates by construction, and the qualitative conclusion — `same_model` dominates — holds on those. And the negative control of Section 10.2.4 holds capability fixed while the arm label varies, returning fifteen nulls. Neither device separates identity from capability outright; a design that did would require two distinct models matched on failure rate, which we did not run.

Arms do not share missions. Mission identifiers embed the condition name and the task is a fixed hash of the identifier, so each arm draws its own tasks from the shared generator distribution. Comparisons are between independent draws, not paired. At $n = 6000$ per primary arm the resulting variance is absorbed by the bootstrap; for the under-run breadth arms of Section 10.6 it is a material limit, and the Meta arm’s single co-failure makes its overlap estimate uninformative.

All arms ran inside one window. The full campaign executed on a single day across roughly four hours. A provider-side change — a model version rollout, a serving-stack change, a load excursion — occurring mid-campaign would confound arms with time, since arms ran sequentially rather than interleaved. The negative control again bounds this: a pair whose composition never changed shows no arm effect, which a global temporal shift would have produced. We nonetheless regard sequential execution as a design weakness and would interleave arms in a replication.

Contracts and gold code share an author. The contracts under test and the deterministic scoring code were written by the same authors. A contract inadvertently written to match what the scorer checks would inflate measured compliance uniformly. It would not, however, generate the *differential* co-failure structure E1 reports, since the same contracts and scorer are used in all arms; the threat is to absolute compliance levels, which this paper does not claim, rather than to the arm contrasts, which it does.

Two task domains only. Missions come from six generators spanning retail and financial workflows, scored deterministically. Whether the dependence magnitudes transfer to code generation, research synthesis, or open-ended dialogue is untested. We expect the direction to transfer — shared weights imply shared failure modes on any distribution — and make no claim about the magnitudes.

One statistic was chosen after seeing the data. The preregistration fixed τ_a as the primary estimator. The decision to lead H2 on the log odds ratio was made after observing the reversal described in Section 10.2.3. Choosing an estimator after seeing outcomes is a researcher degree of freedom of exactly the kind that inflates false positives (Simmons et al., 2011; Gelman and Loken, 2014), so we disclose it rather than present log OR as preregistered. Three considerations bear on it: the preregistered falsification test was evaluated on τ_a as registered and passes (Section 10.2.2); all five statistics are reported for every arm and contrast, so nothing is hidden by the choice; and the reason for preferring a marginal-free statistic is structural rather than result-dependent, since τ_a ’s marginal bound is a property of its definition (Definition 3.17). A reader who insists on the registered statistic alone reaches the same verdict on H1 and on the `same_model` contrasts, and a *stronger* rejection of the three-level ordering.

11.4 Broader impact

The failure this paper documents is quiet. A composition bound computed under independence returns a number, the number looks reassuring, and nothing in the pipeline signals that the assumption behind it is false. Systems certified this way will be deployed in exactly the settings — financial screening, clinical triage, content moderation — where correlated failure is most costly, because those are the settings where redundant review is mandated. Making the dependence measurable, and making the certificate degrade honestly when it is unmeasured, is the contribution we regard as most consequential.

There is a countervailing risk. A certificate is a number that invites over-trust, and Definition 5.1 restricts it narrowly: to one mission distribution, one set of model versions, one topology. A team that treats a Tier-1 floor as a property of their system rather than of their measurement will be wrong the first time they upgrade a model. We have made the scope explicit in the artifact rather than only in this paper, and we regard tooling that silently carries a certificate across a model upgrade as an anti-pattern this work should not be used to justify.

12 Contributions and their standing

Each contribution below is labelled *lead* (we know of no prior work establishing it), *co-lead* (concurrent work reaches a related result by different means; we claim the specific form given here), or *supporting* (the technique is established and the application is what is new).

C1 — Preregistered, manipulated measurement of cross-agent failure dependence. (Co-lead.) That agents sharing a base model can fail together is established: McDonnell et al. (2026) quantify correlated failure in a three-learner triage ensemble, Huang et al. (2026) observe shared blind spots within model families, and Bengio et al. (2026) names it as a concern. We do not claim the phenomenon. What is new is the design: model sharing as a *manipulated* variable at three levels, preregistered before outcomes existed, with deterministic gold scoring and no LLM judge. The registration is confirmatory for the **series2** motif over 18,000 missions; the two further topologies are secondary replications within a 30,820-mission campaign and we label them so. That converts an observation about particular systems into a contrast attributable to substitution, and it is what licenses the causal reading of Table 3. It also yields the finding we did not expect: the ordering is real at the model level and *absent* at the vendor level (Section 10.2.2).

C2 — A finite-sample, copula-agnostic reliability certificate for composed agent pipelines. (Lead.) Theorem 5.2 combines an exact Clopper–Pearson moment box with the extremal LP of Theorem 6.1 to give a $(1 - \eta_{\text{conf}})$ lower bound on composed all-success reliability that assumes no dependence structure whatever. Neither the correlated-failure literature above nor the agent-evaluation literature produces a bound of any kind; measuring that failures correlate does not tell an operator what they may certify. The LP is classical (Boole, 1854; Hailperin, 1965; Bertsimas and Popescu, 2005); its use as a finite-sample agent certificate, and Proposition 6.2’s allocation that makes the floor monotone in the moment family by construction, are ours.

C3 — Coverage collapse of model-based reliability floors. (Supporting.) Theorem 4.2 proves that a bootstrap lower bound on a *fitted* dependence model’s functional loses coverage of the true reliability as $n \rightarrow \infty$, because the identification gap is $O(1)$ while the bootstrap haircut is $O(n^{-1/2})$. The practical content is uncomfortable and specific: past some finite sample size such a certificate is wrong with probability approaching one, and nothing in the interval signals it. This is our reason for reporting the Gaussian floor as a diagnostic and never as a guarantee

(Remark 5.2), and Section 10.4 exhibits the collapse on a witness whose misspecification is invisible to marginal and pairwise data.

C4 — Anytime-valid certification of graph reliability. (Co-lead.) The betting e-process and its validity under optional stopping are established (Ville, 1939; Shafer and Vovk, 2019; Shafer, 2021; Ramdas et al., 2023; Waudby-Smith and Ramdas, 2024; Grünwald et al., 2024). Applying them to composed agent reliability is what we claim, and the payoff is specific to this paper’s problem: the null in Definition 7.1 constrains only a conditional mean, so the sequential certificate requires no independence assumption across missions, components, or shared-model shocks — it is immune to precisely the failure that Section 4 documents for the composition bound. Proposition 7.1 records that the SPRT is recovered exactly, so the anytime guarantee is free against a known alternative.

C5 — Marginal sensitivity as a measurement hazard. (Supporting.) That Jaccard, ϕ , and τ_a are bounded by the marginals is textbook. The contribution is the demonstration that this is decisive in practice: Section 10.2.3 shows a significant *reversal* of the apparent condition ordering driven purely by a difference in marginal failure rates, and Section 10.6 replicates the mechanism on two further backends where the marginal-free statistic rises while the marginal-sensitive one collapses. The consequence generalises beyond this paper: correlated-failure results that lead with ϕ — including McDonnell et al. (2026) — are sensitive to marginal imbalance between the compared agents, and reporting a marginal-free companion statistic is cheap insurance.

C6 — An internal negative control, and the artifact. (Supporting.) Section 10.2.4 identifies a pair whose composition is held constant across all three arms. Fifteen of fifteen contrasts are non-significant with point estimates near zero, and in every arm the pair’s dependence sits where a same-model pair should sit. Two things follow. The estimator returns the same answer for the same composition regardless of what the rest of the graph is doing, which is convergent validity. And the leading alternative explanations for the E1 effect — task-mix drift between runs, ordering, wall-clock variation, provider-side variation — would each have moved this pair, and none did. It is not an equivalence test: it bounds no confound against a prespecified margin, and a confound acting only on the substituted worker would leave the pair untouched. The design was nonetheless falsifiable at this point, and was not falsified. The contracts, generators, scoring code, analysis scripts, and preregistration are released (Section C).

12.1 What we do not claim

We do not claim that shared models cause correlated failures for the first time; Section 2.3 attributes that. We do not claim a three-level sharing ordering: the vendor-level contrast fails to replicate and we report the null. We do not claim the drift dynamics of Theorem 3.1 are validated here — they are carried from v1 and not re-measured. We do not claim the certificate scales past small m ; Remark 6.1 states the exponential cost. And we do not claim frontier-model results: the evaluation runs mid-sized open models, and whether the dependence magnitudes hold at the frontier is untested.

13 Future Work

Five directions follow directly from what this paper could not settle.

Separating model identity from model capability. The confound of Section 11.3 is the most consequential open problem for C1. The design that resolves it pairs two models matched on marginal failure rate but differing in training lineage — for instance, two distinct architectures tuned to equal accuracy on the mission distribution. Any residual association is then attributable to shared inductive bias rather than to shared competence. We did not run it: matching on failure rate requires a prior calibration pass over the candidate models, which the registration did not provide for and which cannot be added after outcomes exist without forfeiting the confirmatory status of the arms it would inform.

Certifying large compositions. Equation (18) is exponential in m (Remark 6.1). Two routes are worth pursuing: certifying sub-compositions and composing the certificates, which is sound but lossy and whose loss we have not quantified; and column generation over the 2^m cells, which would keep sharpness for structured moment families. Neither is attempted here.

Model-agent aggregators. Merge and quorum nodes are deterministic code throughout this evaluation, so they contribute no correlated failure. Systems in which the aggregator is itself an LLM — a judge, a router, a critic — introduce a third dependent component whose failures may correlate with the workers it is adjudicating. That is the common design in deployed systems and the analysis here does not cover it.

Certificate decay. By Definition 5.1 a certificate holds over one mission distribution and one set of model versions. Nothing here says how fast it decays as a deployment drifts, which is the quantity an operator actually needs to know when deciding how often to recertify. The anytime-valid machinery of Section 7 is the natural tool: a certificate that is continuously re-earned rather than periodically re-issued.

Broader task distributions. Missions here span retail and financial workflows with deterministic gold scoring, which is what makes the measurement trustworthy and also what limits it. Whether the dependence magnitudes transfer to code generation, research synthesis, or open-ended dialogue — domains where scoring is itself contested — is untested. We expect the direction to hold and make no claim about the magnitudes.

14 Conclusion

Compositional reliability bounds for multi-agent systems multiply component reliabilities, and that step assumes the components fail independently. A preregistered confirmatory evaluation of 18,000 two-agent-handoff missions, within a 30,820-mission campaign, measures what the assumption costs. Two instances of one model co-fail on 90.0% of the missions on which either fails (log OR = 6.66, 95% CI [6.38, 7.00]). Substituting a different model reduces the association significantly in the confirmatory motif and in both secondary topologies. Substituting a different vendor, with the model already different, does not — and we report that null rather than the tidier three-level ordering we registered.

The consequence is not that the bound is slightly loose. By Proposition 4.1 the error is signed, and it runs against the operator: positive dependence inflates joint failure above the independent product, so a redundant design is over-credited exactly when its components share a model. Dropping the assumption entirely does not help, because the assumption-free floor is zero whenever mean component reliability falls below $1 - 1/m$. Fitting a dependence model is worse than either: Theorem 4.2 proves its coverage of the true reliability tends to zero as the sample grows, with no visible symptom.

What works is to constrain the joint with measured co-execution moments and optimise over everything consistent with them. That certificate is sound without any dependence assumption

(Theorem 5.2), sharp for the information supplied (Theorem 6.1), and tightens as the moment family grows: enriching from ten functionals to fourteen narrows the identified interval by 85.7% and lifts the certified floor from 0.2455 to 0.4116 on identical data. It is insensitive to the i.i.d. assumption at the magnitudes the data support — at a conceded design effect 1.31 times the largest measured, the floor moves at most 2.69 percentage points. And where a team wants to watch a dashboard and stop when it looks good, the anytime-valid certificate holds its type-I error at 0.0471 or below across every admissible betting fraction.

None of this makes correlated failure go away. It makes it measurable, and it makes the resulting guarantee degrade honestly instead of silently. That distinction is the whole of the contribution: a system whose certificate is merely loose can be shipped with known margin, whereas a system whose certificate is confidently wrong cannot be shipped safely at all — and the second is what the independence assumption produces.

A Full Proofs

Every numbered result in Section 4–Section 7 is proved here in full. Nothing is left as a sketch.

A.1 Proof of Theorem 3.1

Proof. Put $e(t) = D(t) - \alpha/\gamma$. Substituting into Definition 3.8,

$$de = dD = (\alpha - \gamma D)dt + \sigma dW = -\gamma e dt + \sigma dW,$$

so e is an Ornstein–Uhlenbeck process with mean reverting to zero. Applying Itô’s formula to $f(t, e) = e^{\gamma t} e$ gives $d(e^{\gamma t} e) = \sigma e^{\gamma t} dW$, and integrating from 0 to t ,

$$e(t) = e(0)e^{-\gamma t} + \sigma \int_0^t e^{-\gamma(t-s)} dW(s). \quad (29)$$

(v). The Itô integral in Equation (29) has mean zero and, by the Itô isometry, variance $\sigma^2 \int_0^t e^{-2\gamma(t-s)} ds = \frac{\sigma^2}{2\gamma}(1 - e^{-2\gamma t})$. The integrand is deterministic and adapted, and by hypothesis $e(0)$ is independent of the driving Wiener process with $\mathbb{E}[e(0)^2] < \infty$; the cross term therefore vanishes, $\mathbb{E}[e(0) \int_0^t e^{-\gamma(t-s)} dW(s)] = \mathbb{E}[e(0)] \cdot 0 = 0$, and squaring Equation (29) gives

$$\mathbb{E}[e(t)^2] = \mathbb{E}[e(0)^2] e^{-2\gamma t} + \frac{\sigma^2}{2\gamma}(1 - e^{-2\gamma t}).$$

Note e is mean-zero only when $\mathbb{E}[e(0)] = 0$; in general $\mathbb{E}[e(t)] = \mathbb{E}[e(0)]e^{-\gamma t}$, which decays to zero at rate γ and does not affect (i)–(iv).

(i), (iii). The stochastic integral in Equation (29) is a Wiener integral of a deterministic kernel, hence Gaussian. Letting $t \rightarrow \infty$ in Equation (29), the transient $e(0)e^{-\gamma t} \rightarrow 0$ and the variance converges to $\sigma^2/(2\gamma)$, so $e(t) \Rightarrow \mathcal{N}(0, \sigma^2/(2\gamma))$.

For stationarity, take $e(0) \sim \mathcal{N}(0, \sigma^2/(2\gamma))$ independent of W — the integrand in Equation (29) is adapted and the increments $\{W(s) - W(0)\}_{s>0}$ are independent of $\mathcal{F}_0$, so the two terms of Equation (29) are independent and their variances add:

$$\text{Var}(e(t)) = e^{-2\gamma t} \frac{\sigma^2}{2\gamma} + \frac{\sigma^2}{2\gamma}(1 - e^{-2\gamma t}) = \frac{\sigma^2}{2\gamma}$$

for every t , while the mean stays 0. Hence the law is unchanged in t and $\pi_D = \mathcal{N}(\alpha/\gamma, \sigma^2/(2\gamma))$ is stationary.

(ii). Immediate from (i): $\mathbb{E}_\pi[D] = \alpha/\gamma$, which is < 1 exactly when $\gamma > \alpha$.

(iv). Under π , $Z = (D - \alpha/\gamma)/\sqrt{\sigma^2/(2\gamma)}$ is standard normal, and the standard Gaussian tail bound $\mathbb{P}(Z > z) \leq e^{-z^2/2}$ for $z > 0$ gives, with $z = \eta/\sqrt{\sigma^2/(2\gamma)}$,

$$\mathbb{P}_\pi(D > \alpha/\gamma + \eta) \leq \exp\left(-\frac{\eta^2}{2} \cdot \frac{2\gamma}{\sigma^2}\right) = \exp\left(-\frac{\gamma\eta^2}{\sigma^2}\right). \quad \square$$

Remark A.1. Definition 3.8 models D on all of $\mathbb{R}$ while Definition 3.4 confines it to $[0, 1]$. The stationary law therefore assigns non-zero mass outside $[0, 1]$; the mass below 0 is $\Phi(-\alpha\sqrt{2/(\gamma\sigma^2)})$, which is negligible for the parameter ranges in which the model is used but is not zero. We report this rather than suppress it: the bound in (iv) is a bound for the unconstrained process, and it remains an upper bound for the constrained one since truncation to $[0, 1]$ only removes mass from the upper tail.

A.2 Proof of Proposition 3.1

Proof. Fix a turn with state-action pair (s_t, a_t) and let $k = |\mathcal{I} \cup \mathcal{G}|$ be the number of constraints and $|A|$ the action-vocabulary size.

Constraint evaluation. By Definition 3.2 the evaluator is stateless: each constraint c is a predicate on (s_t, a_t) evaluated independently of other constraints and of prior turns. Evaluating all of them and forming the two ratios is k predicate evaluations plus $O(1)$ arithmetic, hence $O(k)$ under the standard assumption that each predicate costs $O(1)$.

Drift update. By Definition 3.5 the distributional term requires $\text{JSD}(P_{\text{obs}} \| P_{\text{ref}})$ over the action alphabet. Updating the sliding-window histogram on one new action is $O(1)$ (one increment, one decrement at the window boundary), and evaluating the divergence is a sum over the support, hence $O(|A|)$.

Aggregation. By Definition 3.4, combining the two components is a fixed convex combination, $O(1)$.

Summing, the per-action cost is $O(k) + O(|A|) + O(1) = O(k + |A|)$. The measured constant is small: for $k < 100$ and $|A| < 50$ the v1 implementation records under 10 ms per action. $\square$

A.3 Proof of Proposition 4.1

Proof. Write $T_k = \prod_{i=k}^m h_i$ for $1 \leq k \leq m$, with $T_{m+1} = 1$, so $Y_G = T_1$ and $\mathbb{P}(Y_G = 1) = \mathbb{E}[T_1]$ since T_1 is a $\{0, 1\}$ variable. For each k ,

$$\mathbb{E}[T_k] = \mathbb{E}[h_k T_{k+1}] = \text{Cov}(h_k, T_{k+1}) + p_k \mathbb{E}[T_{k+1}].$$

Applying this identity at $k = 1$ and then recursively substituting for $\mathbb{E}[T_{k+1}]$ gives, after $m - 1$ steps,

$$\mathbb{E}[T_1] = \sum_{k=1}^{m-1} \left(\prod_{i=1}^{k-1} p_i \right) \text{Cov}(h_k, T_{k+1}) + \left(\prod_{i=1}^{m-1} p_i \right) \mathbb{E}[T_m],$$

where the empty product at $k = 1$ equals 1. Since $T_m = h_m$ we have $\mathbb{E}[T_m] = p_m$, so the final term is $\prod_{i=1}^m p_i$. Rearranging,

$$\mathbb{P}(Y_G = 1) - \prod_{i=1}^m p_i = \sum_{k=1}^{m-1} \left(\prod_{i=1}^{k-1} p_i \right) \text{Cov}\left(h_k, \prod_{i=k+1}^m h_i\right),$$

which is the stated expression with the $k = 1$ term written out separately. For $m = 2$ the sum has the single term $\text{Cov}(h_1, h_2)$.

For the directional claim at $m = 2$: $\mathbb{P}(Y_G = 0) = 1 - p_1 p_2 - \text{Cov}(h_1, h_2)$, while the independence calculation returns $1 - p_1 p_2$. Under positive dependence $\text{Cov}(h_1, h_2) > 0$, so the true probability that the series fails is *smaller* than the independent calculation — the independence product is conservative for a series system, in both directions. $\square$

A.4 Proof of Corollary 4.1

Proof. Let $F_i = 1 - h_i$. Expanding the covariance of the two failure indicators,

$$\mathbb{P}(F_1 = F_2 = 1) = \mathbb{E}[F_1 F_2] = \text{Cov}(F_1, F_2) + \mathbb{E}[F_1] \mathbb{E}[F_2] = \text{Cov}(F_1, F_2) + (1 - p_1)(1 - p_2).$$

Because $F_i = 1 - h_i$ is an affine function of h_i with slope -1 , covariance is preserved: $\text{Cov}(F_1, F_2) = \text{Cov}(1 - h_1, 1 - h_2) = (-1)(-1) \text{Cov}(h_1, h_2) = \text{Cov}(h_1, h_2)$. Hence under positive dependence $\text{Cov}(h_1, h_2) > 0$ and $\mathbb{P}(F_1 = F_2 = 1) > (1 - p_1)(1 - p_2)$: the probability that both redundant paths fail together strictly exceeds the independence product.

The two statements are therefore compatible and concern different events. Series *any*-failure is over-estimated by the independence calculation; redundant *joint* failure is under-estimated by it. A redundant design is bought to control the second. $\square$

A.5 Proof of Theorem 4.1

Proof. Let $A_i = \{h_i = 1\}$, so $\mathbb{P}(A_i) = p_i$.

Upper bound. $\bigcap_i A_i \subseteq A_j$ for each j , hence $\mathbb{P}(\bigcap_i A_i) \leq \min_j p_j$.

Lower bound. By De Morgan and the union bound,

$$\mathbb{P}\left(\bigcap_i A_i\right) = 1 - \mathbb{P}\left(\bigcup_i A_i^c\right) \geq 1 - \sum_{i=1}^m (1 - p_i) = \sum_{i=1}^m p_i - (m - 1),$$

and the probability is non-negative, giving the maximum with 0.

Attainment of the upper bound. Let $U \sim \text{Unif}[0, 1)$ and set $h_i = \mathbf{1}[U < p_i]$. Each h_i has the correct marginal, and $\bigcap_i \{h_i = 1\} = \{U < \min_i p_i\}$, which has probability $\min_i p_i$.

Attainment of the lower bound. Let $s = \sum_i p_i - (m - 1)$ and note $\sum_i (1 - p_i) = m - \sum_i p_i = 1 - s$.

If $s > 0$ then $\sum_i (1 - p_i) = 1 - s < 1$, so we may choose pairwise disjoint sets $B_i \subseteq [0, 1)$ with Lebesgue measure $1 - p_i$. Put $h_i = \mathbf{1}[U \notin B_i]$. Then $\mathbb{P}(h_i = 1) = p_i$, and $\bigcap_i \{h_i = 1\} = \{U \notin \bigcup_i B_i\}$ has measure $1 - \sum_i (1 - p_i) = s$, attaining the bound.

If $s \leq 0$ then $\sum_i (1 - p_i) = 1 - s \geq 1$, and we construct a cover explicitly rather than assert one. Identify $[0, 1)$ with the circle $\mathbb{R}/\mathbb{Z}$. Put $L_i = 1 - p_i$, $S_0 = 0$, and $S_k = \sum_{i \leq k} L_i$, and lay the arcs end to end with wrap-around:

$$B_k = [S_{k-1} \bmod 1, (S_{k-1} + L_k) \bmod 1) \subseteq \mathbb{R}/\mathbb{Z}.$$

Each B_k has Lebesgue measure $L_k = 1 - p_k$, so $\mathbb{P}(h_k = 1) = p_k$ as required. The arcs are laid consecutively without gaps starting at 0 and have total length $\sum_k L_k = 1 - s \geq 1$, so their union wraps at least once around the circle and $\bigcup_k B_k = [0, 1)$. Hence $\bigcap_k \{h_k = 1\} = \{U \notin \bigcup_k B_k\} = \emptyset$ and $\mathbb{P}(\bigcap_k \{h_k = 1\}) = 0 = \max(0, s)$.

(For $s > 0$ the same construction has total length $1 - s < 1$, the arcs stay disjoint, no wrap occurs, and the uncovered remainder has measure s — which is the previous case.)

Both bounds are therefore attained by laws with the prescribed marginals, so neither can be improved using marginal information alone. $\square$

Proof of Corollary 4.2. Immediate from Theorem 4.1: the lower bound is 0 iff $\sum_i p_i \leq m - 1$, i.e. iff the mean marginal $\bar{p} \leq 1 - 1/m$. At $m = 4$, $1 - 1/4 = 0.75$, so $\bar{p} = 0.75$ gives a floor of exactly 0. $\square$

A.6 Proof of Theorem 4.2

Proof. Write $G_n = R_{\mathcal{F}}(\mu^*) - \hat{L}_n$ for the bootstrap haircut. By hypothesis $G_n = O_p(n^{-1/2})$, i.e. the family $\{\sqrt{n} G_n\}$ is bounded in probability: for every $\varepsilon > 0$ there exist M_ε and N_ε with $\mathbb{P}(\sqrt{n} |G_n| > M_\varepsilon) < \varepsilon$ for all $n \geq N_\varepsilon$.

Coverage of the true value fails exactly when $\hat{L}_n > R^*$, so

$$\mathbb{P}(\hat{L}_n \leq R^*) = \mathbb{P}(R_{\mathcal{F}}(\mu^*) - G_n \leq R^*) = \mathbb{P}(G_n \geq R_{\mathcal{F}}(\mu^*) - R^*) = \mathbb{P}(G_n \geq \Delta),$$

using Equation (15). Since $\Delta > 0$ is a fixed constant not depending on n ,

$$\mathbb{P}(G_n \geq \Delta) = \mathbb{P}(\sqrt{n} G_n \geq \sqrt{n} \Delta).$$

Fix $\varepsilon > 0$ and take M_ε as above. For all n large enough that $\sqrt{n}\Delta > M_\varepsilon$ — which holds for $n > M_\varepsilon^2/\Delta^2$ — we have

$$\mathbb{P}(\sqrt{n}G_n \geq \sqrt{n}\Delta) \leq \mathbb{P}(\sqrt{n}|G_n| > M_\varepsilon) < \varepsilon.$$

As ε was arbitrary, $\mathbb{P}(\hat{L}_n \leq R^*) \rightarrow 0$.

We claim convergence to zero and nothing stronger. Tightness of $\{\sqrt{n}G_n\}$ gives $\mathbb{P}(G_n \geq \Delta) \rightarrow 0$, but it does *not* entail that the sequence is monotone: G_n may oscillate while remaining tight, so coverage need not decrease at every n . The interpretation that survives is the limit — past some finite sample size the interval sits above R^* with probability approaching one — together with the mechanism, that the haircut shrinks while the target does not move. $\square$

Proof of Corollary 4.3. Let μ^* be the true moments and $Q^* \in \mathcal{M}(\mu^*)$ the true law. By Definition 4.1, $\underline{R}(\mu^*)$ is an infimum over a set containing Q^* , hence $\underline{R}(\mu^*) \leq Q^*(\bigwedge_i h_i = 1) = R^*$. The inequality uses only membership of Q^* in $\mathcal{M}$, which holds by definition of μ^* and requires no parametric assumption. Therefore Equation (15) cannot hold with $R_{\mathcal{F}}$ replaced by $\underline{R}$, and the argument of Section A.6 does not apply. $\square$

A.7 Proof of Theorem 5.1

Proof. Let $X = \sum_{r=1}^n \mathbf{1}[Y_G^{(r)} = 1] \sim \text{Bin}(n, \theta)$ with $\theta = \mathbb{P}(Y_G = 1)$, which is the exact law of X when missions are i.i.d. and Y_G is observed on each. The Clopper–Pearson lower limit is

$$\hat{L}_0(X) = \begin{cases} 0, & X = 0, \\ \inf\{\theta' \in [0, 1] : \mathbb{P}_{\theta'}(X' \geq X) > \eta_{\text{conf}}\}, & X \geq 1, \end{cases}$$

where $X' \sim \text{Bin}(n, \theta')$. For fixed $x \geq 1$ the map $\theta' \mapsto \mathbb{P}_{\theta'}(X' \geq x)$ is continuous and strictly increasing on $(0, 1)$, so $\hat{L}_0(x)$ is the unique root of $\mathbb{P}_{\theta'}(X' \geq x) = \eta_{\text{conf}}$ and $\hat{L}_0(x) > \theta$ holds iff $\mathbb{P}_\theta(X' \geq x) < \eta_{\text{conf}}$.

Let $x^* = \min\{x : \mathbb{P}_\theta(X \geq x) < \eta_{\text{conf}}\}$ (with $x^* = n + 1$ if no such x exists). Then $\{\hat{L}_0 > \theta\} = \{X \geq x^*\}$ and, by minimality of x^* ,

$$\mathbb{P}_\theta(\hat{L}_0 > \theta) = \mathbb{P}_\theta(X \geq x^*) < \eta_{\text{conf}}.$$

Hence $\mathbb{P}_\theta(\hat{L}_0 \leq \theta) \geq 1 - \eta_{\text{conf}}$ for every n and every θ , with no asymptotic approximation. The bound is conservative rather than exact because X is discrete. $\square$

A.8 Proof of Theorem 5.2

Proof. For each $S \in \mathcal{J}$ let I_S be the two-sided Clopper–Pearson interval for $\mu_S = Q^*(\bigwedge_{i \in S} h_i = 1)$ built from the n i.i.d. indicator observations $\mathbf{1}[\bigwedge_{i \in S} h_i^{(r)} = 1]$, each tail at level $\eta_{\text{conf}}/(2J)$. By the argument of Section A.7 applied to each tail, $\mathbb{P}(\mu_S \notin I_S) \leq 2 \cdot \eta_{\text{conf}}/(2J) = \eta_{\text{conf}}/J$.

Let $\mathcal{E} = \bigcap_{S \in \mathcal{J}} \{\mu_S \in I_S\}$ be the event that the whole box covers. By the union bound over the J moments,

$$\mathbb{P}(\mathcal{E}^c) \leq \sum_{S \in \mathcal{J}} \mathbb{P}(\mu_S \notin I_S) \leq J \cdot \frac{\eta_{\text{conf}}}{J} = \eta_{\text{conf}}.$$

On $\mathcal{E}$ the true moment vector lies in $B(\hat{\mu})$, so Q^* is a feasible point of the program in Equation (17): it is a probability distribution on $\{0, 1\}^m$ whose $\mathcal{J}$ -moments lie in the box. Since $\hat{L}_1$ is the *minimum* of the objective over the feasible set and Q^* is feasible,

$$\hat{L}_1 \leq Q^*\left(\bigwedge_i h_i = 1\right).$$

Therefore $\mathbb{P}(\hat{L}_1 \leq Q^*(\bigwedge_i h_i = 1)) \geq \mathbb{P}(\mathcal{E}) \geq 1 - \eta_{\text{conf}}$.

Note the argument never referenced the dependence structure of Q^* : it used only that Q^* satisfies its own moment constraints, which is a tautology. That is the precise sense in which Tier 1 is copula-agnostic. $\square$

A.9 Proof of Proposition 5.1

Proof. Let $\mathcal{A}_0 = \{\text{i.i.d. missions}\} \cup \{Y_G \text{ directly observed}\}$ and $\mathcal{A}_1 = \{\text{i.i.d. missions}\}$ be the assumption sets of Theorem 5.1 and Theorem 5.2. Then $\mathcal{A}_1 \subseteq \mathcal{A}_0$.

Suppose the flag asserting end-to-end execution is mis-set. There are two cases. If it is set when it should not be, the certificate reports $\hat{L}_0$, whose validity requires $\mathcal{A}_0$; the second element of $\mathcal{A}_0$ fails, so the guarantee is unsound. If it is unset when it could have been set, the certificate reports $\hat{L}_1$, whose validity requires only $\mathcal{A}_1$, which holds; the guarantee is sound.

Hence selecting Tier 1 by default and Tier 0 only on explicit assertion places the unsound case behind a deliberate action rather than behind an omission. The claim is exactly this asymmetry; by Remark 5.1 it is not a numerical ordering of $\hat{L}_0$ and $\hat{L}_1$.

Finally, the failure in the first case is undetectable from the data: two runs producing identical pass matrices, one executed end-to-end and one assembled from per-stage measurements, are indistinguishable to any function of the matrix, while only the first satisfies $\mathcal{A}_0$. No validation of the input can substitute for the default. $\square$

A.10 Proof of Theorem 6.1

Proof. Soundness. Q^* , viewed as a vector $x^* \in \mathbb{R}^{2^m}$, is feasible for Equation (18): it is non-negative, sums to 1, and satisfies $a_S^\top x^* = \nu_S$ for $S \in \mathcal{J}$ by hypothesis. The minimum over a feasible set containing x^* is at most the objective at x^* , giving $\underline{R} \leq Q^*(\bigwedge_i h_i = 1)$. The upper bound is symmetric.

Attainment. The feasible set $\mathcal{X} = \{x \geq 0 : \mathbf{1}^\top x = 1, a_S^\top x = \nu_S\}$ is the intersection of the probability simplex in $\mathbb{R}^{2^m}$ with finitely many hyperplanes, hence closed and bounded, hence compact; and it is non-empty because $x^* \in \mathcal{X}$. The objective $a_{\{1..m\}}^\top x$ is linear and therefore continuous, so by the extreme value theorem it attains its minimum and maximum on $\mathcal{X}$ at some $x_-, x_+ \in \mathcal{X}$. Every element of $\mathcal{X}$ is a probability vector on $\{0, 1\}^m$ with the prescribed $\mathcal{J}$ -moments, so x_- and x_+ are laws of the required kind, achieving $\underline{R}$ and $\bar{R}$.

Consequence. The set of achievable values of $Q(\bigwedge_i h_i = 1)$ over $Q \in \mathcal{M}(\nu)$ is the image of the connected set $\mathcal{X}$ under a continuous map, hence an interval; combined with attainment of both endpoints it equals exactly $[\underline{R}, \bar{R}]$. Any valid bound using only the moments in $\mathcal{J}$ must hold for every $Q \in \mathcal{M}(\nu)$, in particular for x_- , so it cannot exceed $\underline{R}$. $\square$

A.11 Proof of Proposition 6.1

Proof. Let $\mathcal{X}(\mathcal{J})$ and $\mathcal{X}(\mathcal{J}')$ be the feasible sets for the two families, with moment values agreeing on $\mathcal{J}$. Every $x \in \mathcal{X}(\mathcal{J}')$ satisfies all constraints indexed by $\mathcal{J}' \supseteq \mathcal{J}$, in particular those indexed by $\mathcal{J}$, so $\mathcal{X}(\mathcal{J}') \subseteq \mathcal{X}(\mathcal{J})$. Minimising a fixed objective over a subset cannot yield a smaller value: $\underline{R}(\mathcal{J}) = \min_{\mathcal{X}(\mathcal{J})} \leq \min_{\mathcal{X}(\mathcal{J}')} = \underline{R}(\mathcal{J}')$. The maximisation statement is symmetric. $\square$

A.12 Proof of Proposition 6.2

Proof. (i) Validity of the used-set allocation is Section A.8 verbatim with $J = |\mathcal{J}|$. For the failure of monotonicity, observe that the Clopper–Pearson interval for a fixed count widens as its tail level decreases. Enlarging $\mathcal{J}$ to $\mathcal{J}'$ decreases the per-tail level from $\eta_{\text{conf}}/(2|\mathcal{J}|)$ to $\eta_{\text{conf}}/(2|\mathcal{J}'|)$, so every interval in the box strictly widens. The feasible set for $\mathcal{J}'$ therefore gains constraints (from the new moments) and loses them (from the widened old ones), and neither set need contain the other. Hence the conclusion of Proposition 6.1 does not transfer.

(ii) *Validity of the pre-allocated allocation.* Fix $\mathcal{J}_{\max} \supseteq \mathcal{J}$ and give every constrained moment per-tail level $\eta_{\text{conf}}/(2|\mathcal{J}_{\max}|)$. The union bound over the $2|\mathcal{J}|$ tails actually used gives miscoverage at most $2|\mathcal{J}| \cdot \eta_{\text{conf}}/(2|\mathcal{J}_{\max}|) = \eta_{\text{conf}}|\mathcal{J}|/|\mathcal{J}_{\max}| \leq \eta_{\text{conf}}$, so the certificate is valid at the stated level.

Monotonicity under (ii). Each interval's width now depends only on its own count and on $|\mathcal{J}_{\max}|$, not on $|\mathcal{J}|$. Hence for $\mathcal{J} \subseteq \mathcal{J}' \subseteq \mathcal{J}_{\max}$ the box constraints indexed by $\mathcal{J}$ are identical under both families, and $\mathcal{J}'$ merely adds further constraints. The argument of Section A.11 applies unchanged, giving $\underline{R}(\mathcal{J}) \leq \underline{R}(\mathcal{J}')$ by construction. $\square$

A.13 Proof of Lemma 7.1

Proof. Let $\mathcal{F}_R$ be the σ -algebra generated by $y_1, \dots, y_R$.

Non-negativity. Since $y_r \in \{0, 1\}$ we have $y_r - p_0 \geq -p_0$, so $1 + \lambda_r(y_r - p_0) \geq 1 - \lambda_r p_0 > 0$ because $\lambda_r \in [0, 1/p_0]$. A product of strictly positive factors is positive, so $E_R > 0$ for all R .

Supermartingale property. λ_R is predictable, hence $\mathcal{F}_{R-1}$ -measurable, and E_{R-1} is $\mathcal{F}_{R-1}$ -measurable. Therefore

$$\mathbb{E}[E_R \mid \mathcal{F}_{R-1}] = E_{R-1} \left(1 + \lambda_R(\mathbb{E}[y_R \mid \mathcal{F}_{R-1}] - p_0) \right).$$

Under H_0 we have $\mathbb{E}[y_R \mid \mathcal{F}_{R-1}] \leq p_0$ by hypothesis: this is exactly the content of the sequential null Equation (19), and it is the step at which a merely marginal null would fail, because a bound on $\mathbb{P}(Y_G = 1)$ says nothing about the conditional mean given the past. With $\lambda_R \geq 0$ and $E_{R-1} > 0$, the bracket is at most 1 and $\mathbb{E}[E_R \mid \mathcal{F}_{R-1}] \leq E_{R-1}$. Taking expectations and iterating from $E_0 = 1$ gives $\mathbb{E}[E_R] \leq 1$ for every R . $\square$

A.14 Proof of Theorem 7.1

Proof. Set $a = 1/\alpha$ and let $T = \inf\{R \geq 1 : E_R \geq a\}$, with $T = \infty$ if no crossing occurs. T is a stopping time with respect to $(\mathcal{F}_R)$ because $\{T \leq R\}$ is determined by $E_1, \dots, E_R$.

Fix $n \in \mathbb{N}$. The stopped process $(E_{T \wedge R})_{R \leq n}$ is a non-negative supermartingale by Lemma 7.1 and the optional stopping theorem for bounded stopping times, so

$$\mathbb{E}[E_{T \wedge n}] \leq \mathbb{E}[E_0] = 1.$$

On the event $\{T \leq n\}$ we have $E_{T \wedge n} = E_T \geq a$ by definition of T . Since $E_{T \wedge n} \geq 0$ everywhere,

$$1 \geq \mathbb{E}[E_{T \wedge n}] \geq \mathbb{E}[E_{T \wedge n} \mathbf{1}_{\{T \leq n\}}] \geq a \mathbb{P}(T \leq n),$$

so $\mathbb{P}(T \leq n) \leq 1/a = \alpha$ for every n . The events $\{T \leq n\}$ increase to $\{T < \infty\} = \{\sup_{R \geq 1} E_R \geq a\}$, so by continuity from below

$$\mathbb{P}\left(\sup_{R \geq 1} E_R \geq 1/\alpha\right) = \lim_{n \rightarrow \infty} \mathbb{P}(T \leq n) \leq \alpha.$$

Because the bound is on the supremum over the entire path, it holds simultaneously for all stopping rules: any rule that issues a certificate does so only on the event $\{T < \infty\}$, whose probability is at most α under H_0 , whatever the rule. $\square$

A.15 Proof of Proposition 7.1

Proof. With $\lambda^* = (p_1 - p_0)/(p_0(1 - p_0))$, evaluate the betting factor at each outcome.

For $y = 1$:

$$1 + \lambda^*(1 - p_0) = 1 + \frac{(p_1 - p_0)(1 - p_0)}{p_0(1 - p_0)} = 1 + \frac{p_1 - p_0}{p_0} = \frac{p_1}{p_0}.$$

For $y = 0$:

$$1 - \lambda^* p_0 = 1 - \frac{(p_1 - p_0)p_0}{p_0(1 - p_0)} = 1 - \frac{p_1 - p_0}{1 - p_0} = \frac{(1 - p_0) - (p_1 - p_0)}{1 - p_0} = \frac{1 - p_1}{1 - p_0}.$$

Both agree with $(p_1/p_0)^y((1 - p_1)/(1 - p_0))^{1-y}$, which is the likelihood ratio of $\text{Bern}(p_1)$ to $\text{Bern}(p_0)$ at y . Substituting into Equation (20) makes E_R the product of per-observation likelihood ratios, i.e. the SPRT statistic.

Admissibility: $\lambda^* \geq 0$ since $p_1 > p_0$, and $\lambda^* < 1/p_0$ because $(p_1 - p_0)/(1 - p_0) < 1$ whenever $p_1 < 1$. $\square$

A.16 Proof of Proposition 7.2

Proof. Supermartingale. Each E^λ is a non-negative supermartingale by Lemma 7.1. For non-negative weights summing to one,

$$\mathbb{E}[E_R^{\text{mix}} \mid \mathcal{F}_{R-1}] = \sum_{\lambda} \pi(\lambda) \mathbb{E}[E_R^\lambda \mid \mathcal{F}_{R-1}] \leq \sum_{\lambda} \pi(\lambda) E_{R-1}^\lambda = E_{R-1}^{\text{mix}},$$

where exchanging expectation and the finite sum is immediate. Non-negativity is inherited, and $E_0^{\text{mix}} = \sum_{\lambda} \pi(\lambda) = 1$. Hence Theorem 7.1 applies verbatim.

Regret. All terms are non-negative, so for any fixed λ , $E_R^{\text{mix}} \geq \pi(\lambda) E_R^\lambda$. Taking logarithms, $\log E_R^{\text{mix}} \geq \log E_R^\lambda - \log(1/\pi(\lambda))$, and maximising the right-hand side over $\lambda \in \Lambda$ gives the claim. The bound is pathwise: it holds for every realisation, not merely in expectation. $\square$

B Formula Catalogue

The v1 framework was published with several formulas withheld under a patent claim. That claim is withdrawn. This appendix is the complete disclosure: every formula, its v1 equation number, where it is stated in this paper, and whether it is implemented in the released library. Nothing in this table is redacted, and the “theory only” entries are marked as such rather than left ambiguous.

B.1 Corrections to the v1 statements

Three discrepancies between the v1 paper and its supporting documents are worth recording, because a reader reconstructing the framework from either could adopt the wrong form.

F2, condition (ii). Supporting documents state the soft guarantee as the deterministic bound $\max_t |C_{\text{soft}}(t) - 1| \leq \delta$. The v1 paper’s eq. 4, restated as Equation (7), is a *probabilistic* guarantee about recoverable compliance. These are different conditions and the deterministic one is strictly stronger. Definition 3.6 follows the paper.

F5, the condition list. Supporting documents present “C1–C5” as a single list. The v1 paper uses C1–C4 for the deterministic composition theorem (Def. 4.7, Thm. 4.9) and adds C5, conditional independence, only for the probabilistic theorem (Thm. 4.11). Collapsing them obscures which result needs which assumption — and C5 is the one this paper is about. Definitions 3.9 and 3.10 keep them separate.

F8, the fourth term. Supporting documents describe α_4 as weighting a “recovery success rate”. The v1 paper’s Def. 3.20 defines the fourth term as the stress resilience index $S = \mathbb{E}[C(t) \mid \text{stressed}] / \mathbb{E}[C(t) \mid \text{baseline}]$ (eq. 12). Recovery success and stress resilience are distinct quantities. Definition 3.7 follows the paper and notes the conflation.

Table 12: The F1–F12 catalogue. “v1 ref” cites the framework paper (Bhardwaj, 2026); “here” points into this paper; “code” names the module in the released library or records that the item is theoretical.

	quantity	v1 ref	here	code
F1	composite drift $D(t)$	Def. 3.12, eq. 7–9	Definition 3.4	<code>metrics/drift.py</code>
F2	(p, δ, k) -satisfaction	Def. 3.7, eq. 3–4	Definition 3.6	<code>certification/satisfaction.py</code>
F3	OU drift dynamics	Def. 4.1, eq. 14	Definition 3.8	theory only
F4	Lyapunov stationary law	Thm. 4.3, eq. 19,22	Theorem 3.1	theory only
F5	compositional guarantee	Thm. 4.11, eq. 28–29	Definition 3.11	<code>certification/composition.py</code> (bound only; C1–C5 unverified)
F6	SPRT certification	§7	Definition 3.12	<code>certification/sprt.py</code>
F7	Hoeffding fixed-sample	— (baseline)	Definition 3.13	<code>certification/sprt.py</code>
F8	reliability index Θ	Def. 3.20, eq. 13	Definition 3.7	<code>metrics/theta.py</code>
F9	contract tuple	Def. 3.1	Definition 3.1	<code>models.py</code>
F10	JSD distributional drift	eq. 9	Definition 3.5	<code>metrics/drift.py</code>
F11	constraint evaluation scores	Def. 3.6, eq. 1–2	Definition 3.2	<code>evaluator/engine.py</code>
F12	per-action complexity	Prop. 4.15	Proposition 3.1	measured, <code>benchmarks/</code>

B.2 Scope of the disclosure

F3 and F4 are stated and proved (Definition 3.8, Theorem 3.1, Section A.1) but are not implemented and not measured in this paper: no experiment here estimates α , γ , or σ . F5’s bound is implemented, but conditions C1–C5 are not machine-verified by the library — a caller composing two contracts is responsible for discharging them, and Section 10 exists because C5 in particular frequently cannot be discharged. We flag both rather than let the catalogue imply that a listed formula is an enforced one.

C Reproduction and Artifact

C.1 What is released

The library, the contracts, the mission generators, the deterministic scoring code, the preregistration, the simulation benchmarks, and the analysis scripts are released under AGPL-3.0. Every number in Section 10 is *regenerated* by those scripts rather than transcribed by hand: each E1 and E5 statistic is printed output of `scripts/e1_final.py`, and each E3, E4, and E6 table is printed output of the corresponding benchmark.

The per-mission logs are the one artifact not in the repository. They carry the full model output for every mission in the corpus, and we release them on request rather than by default. Everything needed to *regenerate* them is public — the generators, the frozen sampling configuration of Section 10.1, the preregistration, and the scoring code — so the pipeline reproduces end to end without them. What they save a replicator is the inference spend, not the method.

C.2 Reproducing each experiment

E1, E5 — dependence and cross-backend replication.

`scripts/e1_final.py` reads the per-arm JSONL logs and emits every cell count, estimator, bootstrap interval, and arm contrast in Tables 2 to 4 and 8. Its defaults are the published settings — seed 20260813, $B = 2000$, failure defined as `¬hard_ok`, tables keyed on `component_id` — so a bare invocation reproduces the paper. Cell counts and point estimates reproduce exactly; percentile bootstrap endpoints are resampling-dependent and reproduce to within Monte Carlo error at this B , which does not move any reported verdict.

E2 — floor lifting.

`benchmarks/` reconstructs the $m=4$ pass matrix and calls `moment_cp_box_floor` and `moment_lp_all_success_bounds` at both moment families and both Bonferroni allocations.

E3 — coverage collapse.

Table 6 is produced by exactly two invocations of `benchmarks/coverage_collapse_sim.py`: `--witness adversarial` and `--witness gaussian`. The script’s defaults are the published settings (200 replications, $n_{\text{boot}} = 500$, seed 20260812), so a bare run reproduces the table; it prints its parameters on every run, so a reduced-fidelity result cannot be mistaken for the published one. The Gaussian moments, the LP minimiser and Δ are computed by deterministic quadrature and linear programming inside the same script, so every constant in Example 4.2 and Section 10.4 has a single source.

E4 — anytime validity.

`benchmarks/eprocess_type1_sim.py`, defaulting to the paper settings ($p_0 = 0.8$, $\alpha = 0.05$, 8000 streams of 500 missions, seed 42) and sweeping the admissible range of λ .

E6 — ablation.

`benchmarks/deff_ablation.py` measures lag- k autocorrelation and design effects, then recomputes the certified floor at a conceded design effect.

C.3 Two reproduction hazards

Two ways of computing the wrong number from the logs produce plausible output rather than an error. Both are recorded here.

Components are identified by `component_id`, not `role`. Every model-backed node carries `role = "worker"`. Keying a co-failure table on `role` silently compares an agent *to itself* and returns a degenerate table — Jaccard exactly 1.0 with both off-diagonal cells zero, in every arm. The identifiers are `node_a/node_b` for `series2`, `worker_0..2` plus `aggregator` for `quorum2of3`, and `branch_a/branch_b` plus `merge` for `parallel2`. Deterministic nodes carry `scored = false`.

An exception is not a data point. An earlier version of the E3 script read a field name that did not exist on the result object and caught the resulting `AttributeError` in a broad `except Exception`, scoring it as a coverage miss. It reported “coverage 0.00” for runs in which the estimator never executed. The lesson generalises beyond this script: in simulation code, catch only the exception that models a real degenerate case, and let everything else fail loudly.

The failure field is `hard_ok`. Component records have both `hard_ok` and `soft_ok`. All dependence estimates in this paper are on the hard verdict. Reading a differently-named field yields all-zero tables rather than an error.

C.4 Preregistration

PREREGISTRATION.md is committed in the repository and git-timestamped before any confirmatory outcome was generated. It fixes the hypotheses, the five conditions and their sample sizes, the frozen sampling parameters, the primary estimator and bootstrap, the stopping rule, and the falsification criteria. Two registered deviations are disclosed in this paper: the breadth arms under-ran their registered n (Section 10.6), and H2 is reported on a marginal-free statistic chosen after observing the reversal (Section 11.3). The registration was not posted to an external registry before submission, so its timestamp rests on the repository history rather than on a third party; we regard that as weaker than an external registration and state it plainly.

C.5 Scale

The full campaign is 30,820 scored missions across 12 arms, recorded in 13 execution logs. Each arm resumes from its own log, so an interrupted run never repeats completed missions.

Availability and Licensing

The AGENTASSERT library, contracts, mission generators, scoring code, preregistration, simulation benchmarks, and analysis scripts are released under AGPL-3.0 at <https://github.com/qualixar/agentassert-abc>. The per-mission logs are available on request; see Section C. The v1 framework paper is Bhardwaj (2026).

Author Contributions

V.P.B. designed the study, wrote the preregistration, implemented the library and the experimental harness, ran the campaign, performed the analysis, and wrote the paper. G.S. and A.P.B. contributed to contract design, reviewed the experimental protocol, and reviewed the manuscript. All authors approved the final version.

Use of AI Assistance

An AI assistant was used for writing and for reviewing the manuscript. All experiments, the code repository, the analysis, the testing, and the review were carried out by the authors, who take full responsibility for the contents.

Conflicts of Interest

AGENTASSERT is developed by Qualixar, with which V.P.B. is affiliated. The evaluation measures the library’s own certificates, so the authors are not disinterested parties. Three mitigations are in place and we state them so a reader can weigh them: the confirmatory hypotheses and analysis were preregistered before outcomes existed; all scoring is by deterministic gold code rather than by author judgement or an LLM judge; and the analysis scripts are released, so that every reported number is the printed output of code a reader can inspect rather than a figure transcribed by an author. The patent claim referenced in the v1 paper has been withdrawn, and no patent application covering the methods in this paper is pending.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

- Richard E. Barlow and Frank Proschan. *Statistical Theory of Reliability and Life Testing*. Holt, Rinehart and Winston, 1975.
- Mike Barnett, K. Rustan M. Leino, and Wolfram Schulte. The Spec# programming system: An overview. In *CASSIS*, pages 49–69, 2004.
- Yoshua Bengio et al. International AI safety report 2026. *arXiv preprint arXiv:2602.21012*, 2026.
- Frédéric Berdoz, Leonardo Rugli, and Roger Wattenhofer. Can AI agents agree? *arXiv preprint arXiv:2603.01213*, 2026.
- Dimitris Bertsimas and Ioana Popescu. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization*, 15(3):780–804, 2005.
- Varun Pratap Bhardwaj. Agent behavioral contracts: Formal specification and runtime enforcement for reliable autonomous AI agents. *arXiv preprint arXiv:2602.22302*, 2026.
- Carlo E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.
- George Boole. *An Investigation of the Laws of Thought*. Walton and Maberly, 1854.
- Harrison Chase. LangChain. <https://github.com/langchain-ai/langchain>, 2022.
- Edmund M. Clarke, Orna Grumberg, and Doron A. Peled. *Model Checking*. MIT Press, 1999.
- C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.
- Patrick Cousot and Radhia Cousot. Abstract interpretation: A unified lattice model for static analysis of programs by construction or approximation of fixpoints. In *POPL*, pages 238–252, 1977.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *ICML*, 2024.
- Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.
- Dominik M. Endres and Johannes E. Schindelin. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860, 2003.
- Michael D. Ernst, Jeff H. Perkins, Philip J. Guo, Stephen McCamant, Carlos Pacheco, Matthew S. Tschantz, and Chen Xiao. The Daikon system for dynamic detection of likely invariants. *Science of Computer Programming*, 69(1–3):35–45, 2007.
- Maurice Fréchet. Sur les tableaux de corrélation dont les marges sont données. *Annales de l’Université de Lyon, Section A*, 14:53–77, 1951.
- Andrew Gelman and Eric Loken. The statistical crisis in science. *American Scientist*, 102(6):460–465, 2014.
- Peter Grünwald, Rianne de Heide, and Wouter Koolen. Safe testing. *Journal of the Royal Statistical Society B*, 86(5):1091–1128, 2024.
- Theodore Hailperin. Best possible inequalities for the probability of a logical function of events. *The American Mathematical Monthly*, 72(4):343–359, 1965.

- Lewis Hammond et al. Multi-agent risks from advanced AI. *arXiv preprint arXiv:2502.14143*, 2025.
- C. A. R. Hoare. An axiomatic basis for computer programming. *Communications of the ACM*, 12(10):576–580, 1969.
- Wassily Hoeffding. Maßstabinvariante Korrelationstheorie. *Schriften des Mathematischen Instituts und des Instituts für Angewandte Mathematik der Universität Berlin*, 5:179–233, 1940.
- Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.
- Sirui Hong, Mingchen Zhuge, Jonathan Chen, et al. MetaGPT: Meta programming for a multi-agent collaborative framework. In *ICLR*, 2024.
- 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.
- Jiatan Huang, Mingchen Li, Ziming Li, Sunjae Kwon, Hong Yu, and Chuxu Zhang. Counterfactual graph for multi-agent LLM calibration. *arXiv preprint arXiv:2605.30653*, 2026.
- Paul Jaccard. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50, 1912.
- Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *ICLR*, 2024.
- Maurice G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1–2):81–93, 1938.
- Leslie Kish. *Survey Sampling*. John Wiley and Sons, New York, 1965.
- Leslie Lamport. *Specifying Systems: The TLA+ Language and Tools for Hardware and Software Engineers*. Addison-Wesley, 2002.
- K. Rustan M. Leino. Dafny: An automatic program verifier for functional correctness. In *LPAR*, pages 348–370, 2010.
- Tian Liang, Zhiwei He, Wenxiang Jiao, et al. Encouraging divergent thinking in large language models through multi-agent debate. In *EMNLP*, 2024.
- Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.
- Xiao Liu, Hao Yu, Hanchen Zhang, et al. AgentBench: Evaluating LLMs as agents. In *ICLR*, 2024.
- Shay Seiya McDonnell, Avantika Singh, Quoc-Viet Pham, Vratislav Havlik, and Gregory M. P. O’Hare. Harnessing disagreement: Detecting correlated agreement blindness in multi-agent triage. *arXiv preprint arXiv:2607.19899*, 2026. Accepted, PAAMS 2026.
- Bertrand Meyer. Applying “design by contract”. *Computer*, 25(10):40–51, 1992.
- Roger B. Nelsen. *An Introduction to Copulas*. Springer, 2nd edition, 2006.

- Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018.
- Long Ouyang, Jeffrey Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.
- Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *UIST*, 2023.
- Chen Qian, Wei Liu, Hongzhang Liu, et al. ChatDev: Communicative agents for software development. In *ACL*, 2024.
- Hezhe Qiao, Hanghang Tong, Ee-Peng Lim, Bing Liu, and Guansong Pang. VerifyMAS: Hypothesis verification for failure attribution in LLM multi-agent systems. *arXiv preprint arXiv:2605.17467*, 2026.
- Xue Qin, Simin Luan, John See, Zeyd Boukhers, Cong Yang, and Zhijun Li. Governed capability evolution: Lifecycle-time compatibility checking and rollback for AI-component-based systems. *arXiv preprint arXiv:2604.08059*, 2026.
- Md Nakhla Rafi, Md Ahasanuzzaman, Dong Jae Kim, Zhijie Wang, and Tse-Hsun Chen. FALAT: Tracing failures in LLM agent trajectories via dependency-guided search. *arXiv preprint arXiv:2606.00765*, 2026.
- Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.
- Traian Rebedea, Razvan Dinu, Makesh Sreedhar, Christopher Parisien, and Jonathan Cohen. NeMo guardrails: A toolkit for controllable and safe LLM applications with programmable rails. In *EMNLP System Demonstrations*, 2023.
- Herbert Robbins. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409, 1970.
- Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society A*, 184(2):407–431, 2021.
- Glenn Shafer and Vladimir Vovk. *Game-Theoretic Foundations for Probability and Finance*. Wiley, 2019.
- Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *NeurIPS*, 2023.
- Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, 2011.
- Abe Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l’Institut de Statistique de l’Université de Paris*, 8:229–231, 1959.
- David Slepian. The one-sided barrier problem for Gaussian noise. *Bell System Technical Journal*, 41(2):463–501, 1962.
- Jean Ville. *Étude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939.

- Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, et al. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023.
- Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society B*, 86(1):1–27, 2024.
- Qingyun Wu, Gagan Bansal, Jieyu Zhang, et al. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. In *COLM*, 2024.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *ICLR*, 2023.
- G. Udny Yule. On the association of attributes in statistics: With illustrations from the material of the childhood society. *Philosophical Transactions of the Royal Society A*, 194:257–319, 1900.
- Lifan Zheng, Jiawei Chen, Qinghong Yin, Jingyuan Zhang, Xinyi Zeng, and Yu Tian. Rethinking the reliability of multi-agent system: A perspective from Byzantine fault tolerance. *arXiv preprint arXiv:2511.10400*, 2025.
- Shuyan Zhou, Frank F. Xu, Hao Zhu, et al. WebArena: A realistic web environment for building autonomous agents. In *ICLR*, 2024.