PUBLIC

Precedents, policy, and commitment

A disambiguation study of governance-context effects in AI decision agents

Sam Carter, MeshQu

MRP-2026-04 · v1.0 · STABLE · Published 2026-06-09

predictions-lock df838f7 · corpus a08570709f70…
decision-assurance · ai-governance · procurement · cross-model · disambiguation

MeshQu

Contents

CLASSIFICATION	PUBLIC
VERSION	1.0
PREDICTIONS-LOCK	df838f7
CORPUS SHA	a08570709f70daceaae8e87e48b74bc4
GENERATED	2026-06-09

Abstract

E3 is the third experiment in a pre-registered programme studying governance-context effects in AI decision agents. E2 surfaced three unresolved questions: whether the L3 commitment break was caused by precedent receipts or by accumulated governance context; whether the L4 backoff was driven by the anti-sycophancy nudge or by the policy text itself; and whether inversion-blindness reflected a model-specific behaviour or a broader task-class property.

E3 was designed to separate these effects directly. Three L3 decomposition arms disentangled verdict signal, informational concreteness, and prompt density. An L4-without-nudge variant isolated the contribution of the anti-sycophancy clause. A scaled n=100 diagnostic re-tested inversion-blindness at corpus scale, and a cross-model replication arm introduced Claude Opus 4.7 alongside the original GPT-5.4 configuration.

The corpus falsified the leading interpretations carried forward from E2. The L3 commitment break collapses when verdict-bearing precedents are removed, indicating that accumulated governance context, rather than precedent verdicts themselves, is the dominant driver of commitment. The L4 backoff persists without the anti-sycophancy clause, indicating that policy exposure rather than the nudge language is responsible for most of the effect. The inversion-blindness pattern reproduces across both diagnostic arms and both model families, suggesting a task-class behaviour rather than a model-specific artefact.

The strongest cross-model finding is a separation between reasoning and verdict behaviour. Both models exhibit the same inversion-blind reasoning pattern, yet produce materially different verdict distributions — a distribution-shape comparison rather than per-record equivalence — with GPT-5.4 remaining review-oriented (23% commit rate) and Claude Opus 4.7 committing substantially more often (80%). The result suggests that inversion-blind reasoning may be a task-class property while verdict commitment remains model-specific. If that separation holds in further work, it has direct implications for how cross-model AI deployment in regulated decisioning should be evaluated — **reasoning evaluations may generalise; verdict evaluations may not.**

Across six pre-registered predictions, one was confirmed, three were falsified, and two were under-tested due to confirmation-only lock criteria. All predictions, thresholds, and artefacts remained unchanged from the v0.3 pre-registration lock (`ba4ebfb`).

EXECUTIVE SUMMARY

VERSION	1.0
CLASSIFICATION	PUBLIC
STATUS	STABLE
PREDICTIONS LOCK	v0.3-predictions-locked · ba4ebfb3233b819e · 2026-05-28
RELEASE TAG	v1.0-mrp-2026-04 · df838f777b0c4803 · 2026-06-09
CORPUS RUN	phase-2-20260529T092611-Z
SUBSTRATE	UK Contracts Finder OCDS (cached E1/E2 phase-2 fixture, n=283)
PRIMARY MODEL	gpt-5.4-2026-03-05, temperature=0
CROSS MODEL	claude-opus-4-7, effort=low (no temperature)
ARMS	L3 decomposition (A/B/C) · L4-without-nudge · diagnostic_primary + diagnostic_claude (n=100)

Executive summary. E3 tested whether governance artefacts actually influence AI decision-making. Four findings emerged:

1. Precedents matter less than expected. A treatment that provided only precedent receipts (no policy text, no rules) produced DENY decisions on just 3.5% of records — against a 20% confirmation floor. Governance effects appear to emerge from accumulated context rather than isolated examples.

2. Policy text matters more than prompts. Removing the anti-sycophancy instruction barely changed agent behaviour (a 3.7 percentage-point delta). The policy text itself drove the outcome; the nudge clause was incidental.

3. Reasoning behaviour appears portable across models. GPT-5.4 and Claude Opus 4.7 exhibited the same inversion-blind reasoning pattern, with Cat 2 ("reasons solely against rule intent") at 93% and 100% respectively. The reasoning surface generalises across model families on this corpus.

4. Verdict behaviour does not transfer in the same way. On the same record-matched corpus, GPT-5.4 committed on 23% of records and Opus on 80% — a 57-percentage-point gap in decisive rate. This separation — reasoning portability without verdict portability — is the result with the most direct implication for cross-model AI deployment in regulated decisioning.

Implication. Governance cannot live inside the model alone. The surrounding decision environment — policy text, accumulated context, evaluation methodology — is what carries the load. AI evaluation in regulated contexts should test both the verdict an agent emits and the reasoning that produced it; the receipt-anchored evaluation methodology supports both from a single corpus.

> **The methodology is portable across domains; the substrate findings are not.**

1. Programme context

E3 is the third experiment in a pre-registered programme on governance-context effects in AI decision agents, conducted on a UK public procurement substrate. E1 (MRP-2026-02) established the baseline — a single-condition evaluation on 283 OCDS records, validating the Decision Receipt primitive at corpus scale. E2 (MRP-2026-03) introduced a five-rung additive ladder from L0 baseline through L4 full policy, and produced two structural findings the additive design could not mechanistically isolate: an L3 commitment break and an inversion-blindness signal on an n=14 Permuted-Policy diagnostic. E3's job was to separate those effects directly.

The methodological discipline carries forward from E2 unchanged. Every AI decision in the corpus is bound to a cryptographically signed Decision Receipt — Ed25519 signature, public Sigstore transparency-log anchor, schema-versioned envelope — together with the policy snapshot, prompt SHA, and reasoning text. Predictions are pre-registered with locked confirmation and falsification bands. P-series predictions and F-series findings are reported using their pre-defined disposition vocabularies. Anti-claims are first-class output, reported alongside findings and aggregated in §8. The methodology itself is documented separately in the companion methods note, Receipt-Anchored Evaluation. The programme's question-shape, summarised in Fig. 2, carries the same methodological substrate across all four experiments.

> **Why receipts matter.** *The methodology rests on a single substrate primitive — every AI decision is bound to a signed Decision Receipt at evaluation time. Receipts give the corpus policy provenance (which policy the agent was shown), auditability (any reader can recompute the integrity hash offline), reproducibility (the corpus is re-derivable from on-disk bundles without re-running the agent), and cross-model comparability (token-level provenance is model-independent). Without receipts, this discipline would not scale to a 1,332-record corpus.*

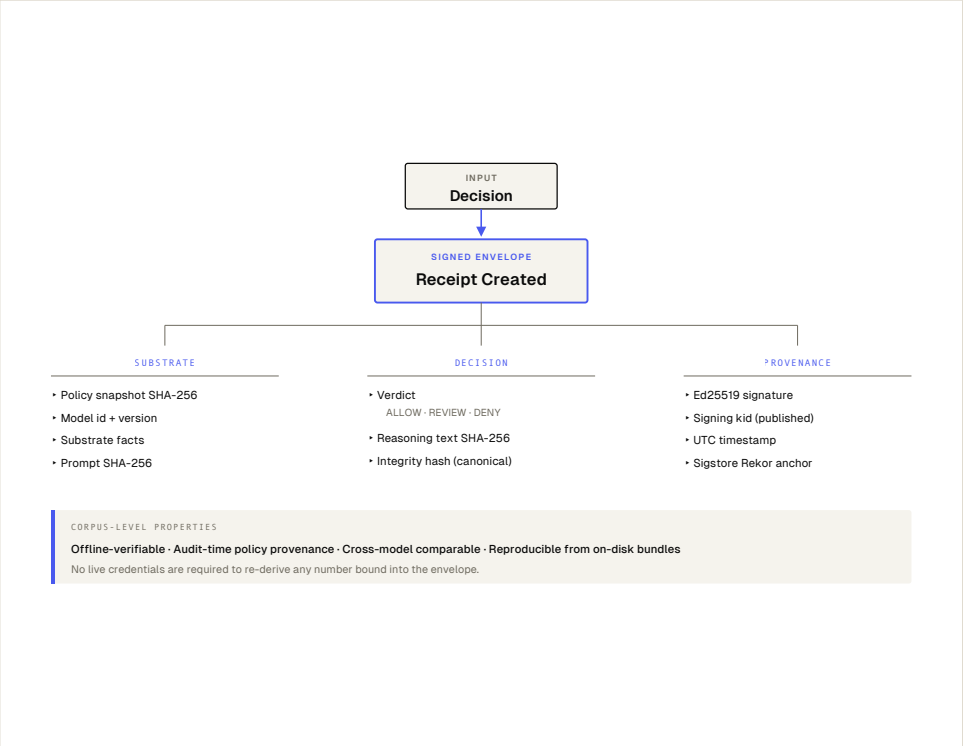


Fig. 1 — Schematic diagram of the signed envelope produced at every AI decision in the corpus. The receipt binds, into one cryptographically signed object: the policy snapshot SHA-256 (which policy the agent was shown), the agent's model identity and version, the substrate facts the agent saw, the agent's verdict (ALLOW / REVIEW / DENY), the reasoning text SHA-256, the integrity hash over the canonical envelope, the Ed25519 signature from the published signing kid meshqu-experiment-procurement-2026-05, a UTC timestamp, and a public Sigstore Rekor transparency-log anchor. Any reader can download a bundle, recompute the envelope bytes, verify the signature against the published key, and check the Rekor anchor — all offline, no credentials. The receipt is the operational primitive that makes the rest of the methodology (pre-registration, anti-claims, cross-model comparison, audit-time policy provenance, cost extrapolation, F-series register) tractable at corpus scale.

> **Key takeaway** – *E3 is the third paper in a four-experiment programme. Each experiment narrows the questions the previous one raised; the methodology substrate – signed receipts, locked predictions, anti-claims, and the disposition vocabulary – carries forward unchanged across all four.*

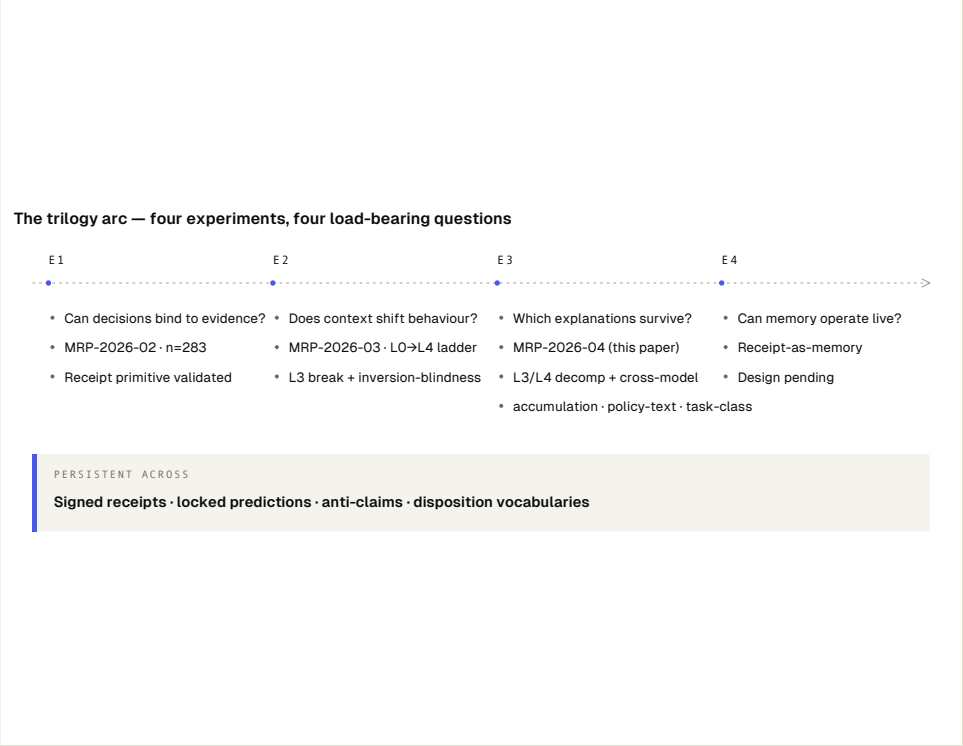


Fig. 2 — Each experiment in the four-experiment programme narrows the questions raised by its predecessor. Columns name the load-bearing question of each; the methodological substrate carries forward across all four.

2. E3 design recap

Predictions-lock
v0.3-predictions-locked ba4ebfb
2026-05-28

The pre-registration was locked on 2026-05-28 as `v0.3-predictions-locked` (commit `ba4ebfb`). The locked content includes the Arm C density-control payload, the Arm B precedent-no-verdict format, the L4-without-nudge prompt variant, the hand-coded rubric protocol, the diagnostic subset selection rule (sha256(ocid) sort, first 100 records), and the Claude version pin (`claude-opus-4-7`, no `temperature`, `effort: low`). The substrate is inherited from E2 unchanged: the same 283-record OCDS corpus, the same policy snapshot (`5d7d800186…`), the same primary agent configuration (`gpt-5.4-2026-03-05`, temperature 0), and the same signing kid (`meshqu-experiment-procurement-2026-05`).

The design consists of three experimental pieces, each targeted at one of E2's unresolved structural readings. Piece 1 — L3 decomposition — runs Arm A (precedents-only), Arm B (precedents-no-verdict), and Arm C (density-control) at n=283 each. It isolates whether precedents drove the L3 commitment break, whether any sufficiently rich governance context would have produced the same effect, and whether verdict exemplars are load-bearing or informational concreteness alone is sufficient. This directly tests the governance-memory interpretation proposed in E2 §10.

Piece 2 — L4 decomposition — runs an L4-without-nudge variant at n=283. It isolates whether the explicit anti-sycophancy clause drove E2's L3→L4 backoff or whether the policy text itself was responsible for most of the effect.

Piece 3 — the scaled Permuted-Policy diagnostic and cross-model arm — runs `diagnostic_primary` and `diagnostic_claude` at n=100 each on a record-matched subset. It evaluates whether the inversion-blindness signal observed in E2 survives at corpus scale and whether the pattern is model-specific or a property of the task class itself.

Predictions P1–P6 are specified at the segment level, incorporating the primary methodological lesson from E2. Several predictions carry forward the directional readings E2 leaned toward, but the design is constructed to distinguish among competing explanations regardless of which direction the corpus ultimately supports.

2.1 Predictions table

Each prediction is reported below against its locked confirmation and (where present) falsification bands. The disposition vocabulary — Confirmed / Falsified / Inverted / Refuted / Deferred / Under-tested — is locked at pre-registration; no "partial confirmations" are permitted.

Table 1. Pre-registered predictions vs. corpus outcomes (P1–P6).

ID	Pre-registered claim	Falsification criterion	Outcome (corpus)	Disposition
P1	Arm A DENY ≥ 20% AND Arm C DENY ≤ 12% (precedents drive commitment, not raw volume)	Arm C DENY ≥ 20% OR Arm A DENY < 20%	Arm A DENY 3.5% (10/283), Arm B DENY 0% (0/283), Arm C DENY 0% (0/283)	**Falsified** — Arm A 3.5% trips the locked falsification clause "Arm A < 20%"; the third-reading interpretation (accumulation amplifies) is starker than the directional 20–30% band anticipated
P2	Arm A DENY rate – Arm B DENY rate ≥ 15pp (verdict exemplars are load-bearing for DENY-anchoring)	locked spec specifies confirm band only; no falsification band locked	A – B DENY gap = 3.5pp (3.5% – 0%)	**Under-tested** — gap is below the 15pp confirm band; no falsification band locked. See §3.2 — the underlying finding (Arm A is the only arm producing any DENY at all) is substantively richer than P2's binary captured
P3	L4-without-nudge retention of E2's L3-DENY set ≥ 80% (nudge load-bearing, Framing A.1)	retention ≤ 65% (locked falsification band — policy text alone, Framing A.2)	60.7% retention (65/107)	**Falsified** — retention 60.7% ≤ 65% locked falsification floor. See §3.4 for the Framing A.2 reading that follows by pre-registered design.
P4	Scaled Permuted-Policy diagnostic: ≥90% same-as-unperturbed L4 verdict (n=100, primary model)	locked spec specifies confirm band only; no falsification band locked	88% same-as-unperturbed (88/100)	**Under-tested** — 88% is 2pp below the 90% confirm floor; no falsification band locked. See §3.5 for the substantive reading; see §4 for the methods-note observation.

ID	Pre-registered claim	Falsification criterion	Outcome (corpus)	Disposition
P5	Hand-coded rubric: Cat 2 ("reasons solely against rule intent") ≥ 60% modal AND Cat 1 ("names the inversion") ≤ 15%	Cat 1 > 25%	Primary reconciled: Cat 2 93% / Cat 1 7%. Claude reconciled: Cat 2 100% / Cat 1 0%	**Confirmed** on both arms (κ between blind agent and reconciled sheet = +1.0000 on both arms — rubric-anchored adjudication, not independent re-coding; see §7 for the protocol scope)

ID	Pre-registered claim	Falsification criterion	Outcome (corpus)	Disposition
P6	Claude's same-as-unperturbed rate within 15pp of primary's rate (task-class, not model-specific)	gap > 15pp (locked falsification band — pre-registered the model-specific outcome as "a strong finding, not a failure")	gap = 46pp (primary 88%, claude 42%)	**Falsified** — 46pp gap far outside locked 15pp band; the cross-model finding is the substantive cross-model contribution (see §3.7)

The disposition vocabulary draws on the **F-series** — the experiment's numbered post-data findings register (F013 onward, continuing from E2's F007–F012), each finding carrying a status label, an evidence block with denominators, and its own anti-claims. The F-series is the post-data analog of the pre-registered P1–P6 predictions: predictions are locked before the data; findings are registered after it. F013 (Piece 1 verdict-distribution refinement), F014 (Cross-model verdict-style divergence), F015 (arm_c verdict-count anchor drift), F016 (diagnostic_claude verdict-mix anchor drift), and F017 (κ-check coder drift catch) are all introduced in §3 and §4 below.

2.2 Questions in plain English

Reading the corpus through the questions the predictions were designed to answer — rather than through the formal P1–P6 dispositions alone — produces a complementary view of what E3 closed:

Pre-registered questions vs. corpus answers — complementary view to the disposition table above.

Question	What we expected	What the corpus shows
Did precedents alone drive the L3 commitment break?	Mostly yes (Reading A)	Only partially — Arm A produces every observed DENY, but at 3.5% rather than the 20%+ rate the prediction required
Did raw content density alone drive the L3 break?	Possibly (Reading B)	No — Arm C produces zero DENYs
Did the anti-sycophancy nudge drive the L4 backoff?	Yes (Framing A.1)	No — `l4_without_nudge` retention is within noise of E2's nudge condition
Does inversion-blindness reproduce at corpus scale?	Open at lock	Yes — 88% same-as-unperturbed on the primary, with the reasoning-axis pattern dominating in both diagnostic arms
Is inversion-blindness model-specific on the reasoning axis?	Open at lock; possibly model-specific	Appears not — Cat 2 dominates on both GPT-5.4 and Claude Opus 4.7
Is verdict-commitment style model-specific?	Not specifically pre-registered	Appears yes — Opus 80% decisive vs GPT-5.4 23% on identical records

The formal disposition table reports against the locked pre-registration; this table reports against the underlying questions. The two complement each other and diverge only where the locked vocabulary forces Under-tested calls (P2, P4) on observations that nonetheless point at a clear corpus-level answer.

3. Results

What changed from E2. Read against E2's two structural findings and the predictions designed to disambiguate them:

- Precedents alone do not explain the L3 commitment break — Arm A produces 3.5% DENY against the locked 20% confirmation floor.
- Policy text, not the anti-sycophancy clause, drives most of the L4 backoff — `l4_without_nudge` retention sits within noise of E2's nudge condition.
- Inversion-blindness reproduces at n=100 — the diagnostic primary arm shows 88% same-as-unperturbed verdicts and Cat 2 dominance at 93%.
- The reasoning pattern reproduces across model families — Cat 2 dominates in both GPT-5.4 (93%) and Claude Opus 4.7 (100%).
- Verdict commitment remains model-specific — Opus 80% commit rate vs GPT-5.4 23%, with the comparison at the distribution-shape level rather than per-record.

Each bullet restates a finding established below; nothing in this summary anticipates a result the corpus does not support.

3.1 P1 falsified: Arm A DENY rate far below the 20% commit floor

Arm A produces a DENY commitment rate of 3.5% (10 records out of 283), an order of magnitude below the 20% locked confirmation threshold. The locked falsification clause — Arm A DENY < 20% — is therefore triggered. Arm C produces zero DENYs (0/283), comfortably below the 12% upper bound for the density-control arm.

The simplest reading is that precedents alone are not sufficient to produce DENY commitment at anything close to the magnitude observed at E2's L3 rung. The corpus supports a more specific interpretation, anticipated in the predictions' interpretive note as a third reading. E2's L3 commitment rate emerged under accumulated governance context: L0 baseline, L1 prose framing, L2 named rules, and L3 precedent receipts operating together. When isolated, precedents produce only 3.5% DENY commitment. Raw prompt density performs even less strongly. Both Reading A (precedents alone drive the L3 break) and Reading B (content density alone drives the L3 break) are falsified by the corpus, with Arm B carrying the principal evidence against Reading B — precedents-no-verdict, full token-parity with Arm A, zero DENY commitment. Arm C corroborates within the documented parity-asymmetry caveat (§3.3). The evidence instead supports an accumulation effect in which multiple governance-context layers combine to produce the commitment behaviour observed in E2.

The falsification of P1 should not be read as a refutation of the precedent mechanism. Arm A still produces every DENY observed across the Piece 1 decomposition — 10 records, compared with zero in Arm B and zero in Arm C. Precedents are doing measurable work on the verdict axis; they simply do not account for the full magnitude of the L3 break when evaluated in isolation.

The three-arm verdict distribution therefore reveals a finding sharper than the binary P1 disposition captures. The important distinction is not whether precedents matter, but how much of the E2 effect they explain. That question is

examined directly in §3.2 alongside the under-tested P2 result. The practical implication is that governance behaviour appears to emerge from interacting governance layers rather than any single artefact in isolation.

> **Key takeaway** — *Precedents matter, but they do not explain the L3 commitment break on their own. The corpus supports an accumulation reading: governance behaviour appears to emerge from interacting governance layers rather than any single artefact in isolation.*

3.2 P2 under-tested; Piece 1 mechanism refinement: verdict-bearing precedents enable directional commitment (mostly ALLOW)

P2 was pre-registered as a DENY-rate anchoring prediction: if verdict-bearing precedents were load-bearing for DENY commitment, Arm A's DENY rate would exceed Arm B's by at least 15 percentage points. The observed gap is 3.5 percentage points (Arm A 3.5%, Arm B 0%), well below the 15pp confirmation threshold. Because the locked specification did not define a lower-band falsification condition, the appropriate disposition under the pre-registered vocabulary is Under-tested. P2 is one of two predictions where the locked-band asymmetry is itself a methods observation — examined in §4.

The verdict distribution across the three arms (table below; visualised in Fig. 3) reveals a finding sharper than P2's DENY-rate binary captured.

Verdict distribution per L3 decomposition arm (n=283 per arm).

arm	ALLOW	REVIEW	DENY	any commitment (ALLOW + DENY)
arm_a (precedents with verdicts)	34	239	10	44 (15.5%)
arm_b (precedents-no-verdict)	9	274	0	9 (3.2%)
arm_c (density-control)	10	273	0	10 (3.5%)
all three arms combined	53	786	10	63 (7.4%)

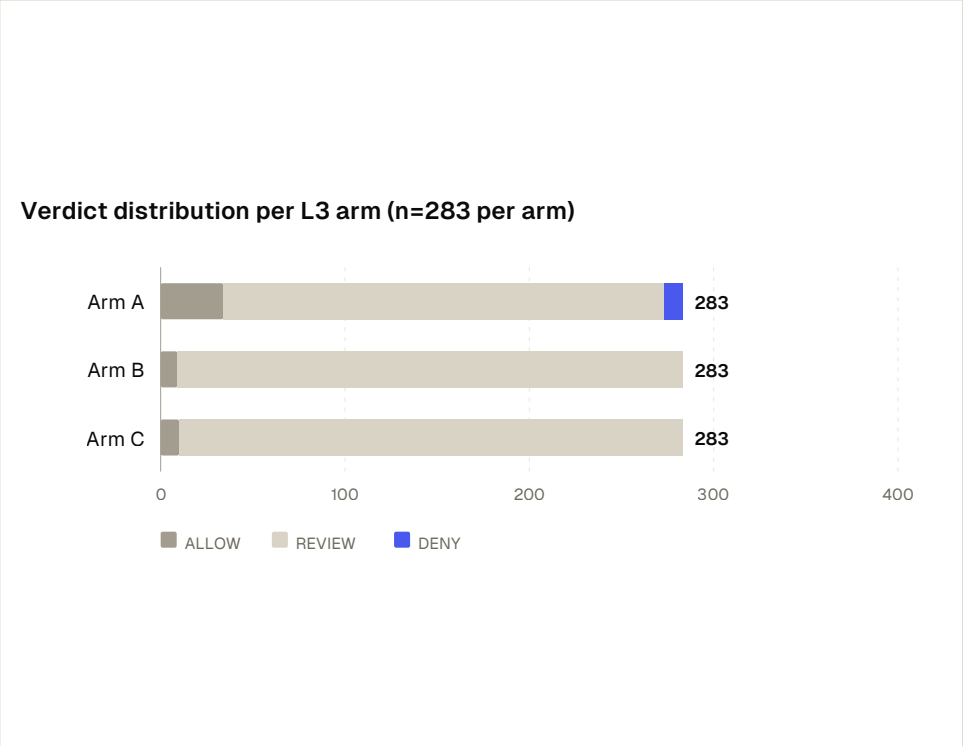


Fig. 3 — Arm A is the only condition producing any DENY commitment. The full pattern: 10/10 DENYs concentrate in Arm A; ALLOW commitment dominates DENY commitment 3.4× across all three arms combined. · Strip the verdict signal from precedents and the agent never commits to DENY.

Strip the verdict signal from precedents, and the agent never commits to DENY in this experimental condition.

Arm A is the only condition producing any DENY commitment across the three arms: 10 records from precedents-with-verdicts, compared with zero from precedents-no-verdict and zero from density-control. Across all three arms combined, ALLOW commitment dominates DENY commitment by 3.4× (53 ALLOW vs 10 DENY). Any-direction commitment — ALLOW or DENY — rises from 3.2% in Arm B to 15.5% in Arm A: a 4.8× lift, well above what the DENY-only slice captured.

The data are consistent with a mechanism P2 did not anticipate: verdict-bearing precedents increase directional commitment in either direction, with ALLOW the dominant direction in this corpus and policy combination. The finding refines the

governance-memory hypothesis by identifying verdict-bearing precedent as the active component while rejecting the stronger claim that the effect operates primarily through DENY anchoring.

P2's locked-band asymmetry collapsed a two-dimensional commitment pattern onto a single DENY-rate axis. Under the locked vocabulary the disposition remains Under-tested; the substantive result is therefore recorded as F013, a Piece 1 mechanism refinement rather than a post-hoc rescue of P2.

> **Key takeaway** — *Verdict-bearing precedents enable directional commitment in either direction — ALLOW or DENY — with ALLOW dominant in this corpus and policy combination. F013 refines the governance-memory interpretation: it is the verdict field that carries the load, not the polarity.*

- **Status**: Discovered
- **Evidence (n=283 per arm, all three L3 arms)**: Arm A produces 100% of observed DENYs across the three L3 arms (10 of 10). Arms B and C produce zero DENYs. Across all three arms, ALLOW commitment dominates DENY commitment 3.4× (53 ALLOW vs 10 DENY)
- **Mechanism**: verdict-bearing precedents enable directional commitment in both directions; ALLOW is the dominant direction in this corpus + policy combination
- **Anti-claim**: F013 does NOT show the §10 governance-memory interpretation from E2 is wrong — it shows the same interpretation is consistent with the corpus, but operates symmetrically rather than DENY-asymmetrically as P2 anticipated
- **E4 design implications**: the operational receipt-as-memory experiment should anticipate ALLOW-bearing precedents anchoring ALLOW commitments at least as strongly as DENY-bearing precedents anchor DENYs

3.3 Arm C asymmetric-control caveat (methods reading)

The Arm C density-control payload was 16.43% shorter than the Arm A precedents-with-verdicts payload, a parity asymmetry locked at PR #93 and pre-registered as a documented methods caveat. The asymmetry eliminates one potential confound — Arm C cannot have failed to commit because it contained more content than Arm A — but introduces another: Arm C may have failed to commit because it contained less content than Arm A.

The asymmetry therefore limits what Arm C can establish regarding exact volume equivalence. It does not, however, affect the pre-registered test criterion for Reading B. P1's locked falsification condition specified that Reading B (raw content density drives commitment) would be rejected if Arm C's DENY rate was 12% or lower, irrespective of whether perfect payload parity had been achieved.

Arm C's observed DENY rate is 0%, comfortably below the 12% threshold. Under the pre-registered criterion, Reading B is therefore falsified.

The caveat remains important for interpretation. Arm C cannot establish that volume has no effect whatsoever on commitment behaviour; it can only establish that the observed commitment pattern is not explained by raw content density at the magnitude required by Reading B. The asymmetry is recorded explicitly so that subsequent comparisons involving Arm C remain clear about both their evidentiary strength and their limits.

3.4 P3 falsified, Framing A.2 confirmed: the policy text drove the backoff

P3 was pre-registered with both confirmation and falsification bands. Under Framing A.1, the anti-sycophancy nudge clause was hypothesised to be load-bearing for E2's L3→L4 backoff. If so, L4-without-nudge retention on the 107-record L3-DENY set would equal or exceed 80%. Under Framing A.2, the backoff was hypothesised to arise from the policy text itself — the explicit rule clauses, threshold tests, and field expectations introduced at L4. If so, retention would fall at or below 65%.

The observed retention is 60.7% (65/107), below the falsification threshold and close to E2's L4-with-nudge retention of 57.0% (61/107). The difference between the two conditions is approximately 3.7 percentage points (Fig. 4), far smaller than would be expected if the anti-sycophancy clause were carrying the primary causal load.

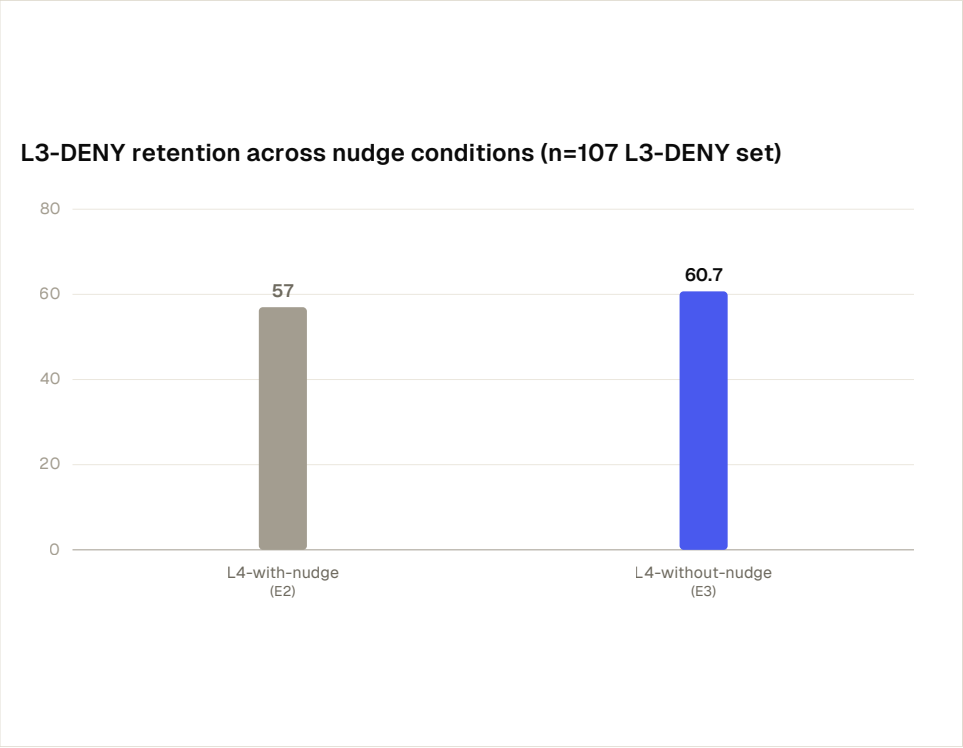


Fig. 4 — L4 retention without the nudge is within noise of L4 with the nudge. The 3.7pp delta between E2's L4-with-nudge (57.0%) and E3's L4-without-nudge (60.7%) is too small for the anti-sycophancy clause to be carrying the structural backoff work; locked bands at 80% confirm / 65% falsify. · The policy text drove the backoff; the nudge clause is incidental.

P3 is therefore falsified under the locked criterion. By the pre-registered design, this result confirms Framing A.2: the policy text itself drove the L3→L4 backoff. Explicit rule clauses, threshold tests, and structured field expectations were sufficient to produce the reduction in DENY commitment observed at L4.

This result should not be interpreted as evidence that the anti-sycophancy clause is ineffective. Rather, it rejects the stronger claim that the clause was load-bearing for the backoff observed in E2. The clause may still provide behavioural discipline

in adversarial or edge-case settings not represented in the diagnostic corpus. What the experiment indicates is that the structural effect was already present in the policy layer.

The magnitude of that policy effect remains substantial. `l4_without_nudge` produces DENY commitment on 27.2% of records (77/283), compared with 3.5% in Arm A (10/283). Even with the nudge removed, the policy condition continues to produce an order-of-magnitude increase in DENY commitment relative to precedents alone. The policy text is doing the structural work.

> **Key takeaway** — *The L3→L4 backoff is policy-text-driven, not nudge-driven. With the nudge removed, the policy condition still produces an order-of-magnitude increase in DENY commitment relative to precedents alone.*

3.5 P4 under-tested: inversion-blindness substantively holds at scale; locked spec specifies confirm band only

On 88 of 100 records in the n=100 Permuted-Policy diagnostic, the agent reached the same verdict whether or not the policy operators had been inverted. The inversion-blindness pattern is overwhelmingly present in the corpus.

P4 was pre-registered with a confirmation band only. If inversion-blindness reproduced robustly at scale, the same-as-unperturbed-L4 verdict rate would equal or exceed 90%. The locked specification did not register a falsification band on the lower side. The observed rate is 88% — two percentage points below the

90% confirmation floor. Because the locked specification did not define a lower-band falsification condition, the appropriate disposition under the pre-registered vocabulary is Under-tested.

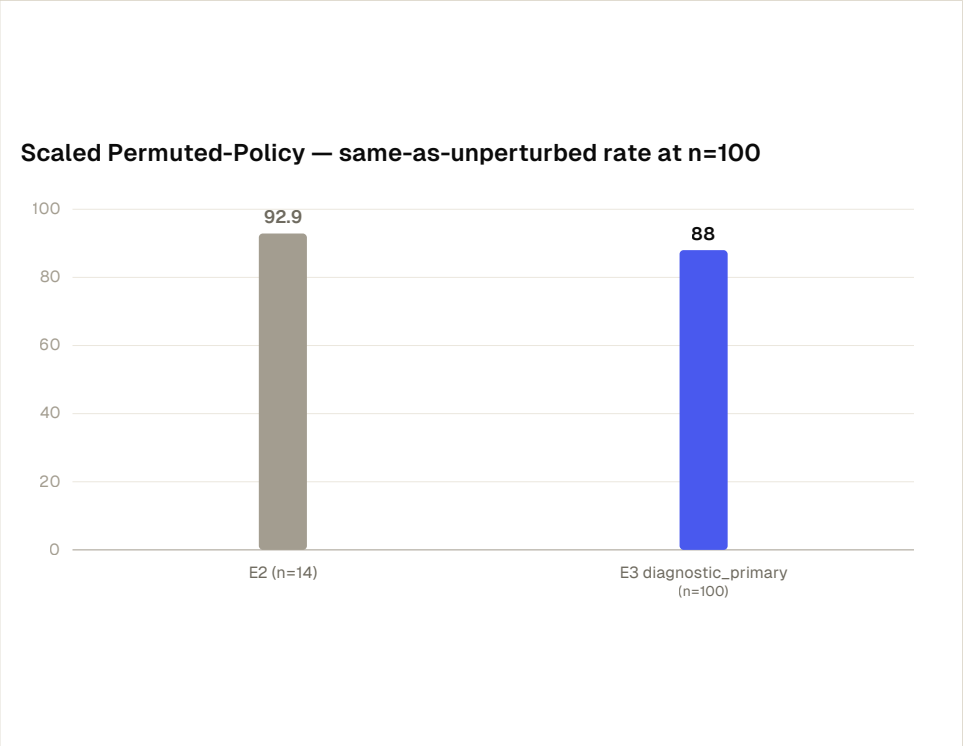


Fig. 5 — 88% of records produce the same verdict under inversion as under unperturbed L4. The observation sits 2pp below the 90% locked confirmation band (no falsification band locked); E2's n=14 anchor result of 92.9% shown as reference. · Pattern overwhelmingly intact at scale; confirmation-band-only force Under-tested on near-misses.

The pre-registered ≥90% threshold was set with reference to E2's n=14 result of 92.9%. The n=100 corpus reduces the observed magnitude slightly (Fig. 5) while leaving the underlying pattern overwhelmingly intact. The reasoning-axis evidence — examined in §3.6 — reinforces this directly: on the 12 records where verdicts shifted, the rubric coding records the same Cat 2 "reasons against rule intent" pattern that dominates both diagnostic arms. The phenomenon is present in the reasoning trace whether or not the verdict happens to shift.

The discipline observation that follows — confirmation-band-only predictions force Under-tested dispositions on near-misses — is examined in §4 as a methods-note refinement carried forward for future predictions.

3.6 P5 confirmed on both diagnostic arms: rubric Cat 2 dominance

On both diagnostic arms, the same Cat 2 reasoning pattern — "reasons solely against rule intent" — dominates the n=100 distribution. P5 is the only prediction in the experiment confirmed against its locked bands, and it is confirmed under two model-protocols rather than one.

P5 was pre-registered with both confirmation and falsification bands. Confirmation required Cat 2 ≥ 60% and Cat 1 ≤ 15% per arm; falsification required Cat 1 > 25%. The locked bands were derived from E2's n=14 result, which produced no contradiction-naming under the lexicon-strict reading and the structural inversion-blind pattern under the v1.1 reading.

On `diagnostic_primary` (GPT-5.4), the rubric distribution is Cat 1 = 7%, Cat 2 = 93%, Cat 3 = 0%. On `diagnostic_claude` (Claude Opus 4.7), it is Cat 1 = 0%, Cat 2 = 100%, Cat 3 = 0% (Fig. 6). Both arms confirm P5 under the locked criterion.

The result is particularly notable because confirmation occurs under two independent model-protocols: GPT-5.4 and Claude Opus 4.7. The inversion-blindness pattern therefore appears to be a property of the task class rather than an artefact of either individual model. The Cat 1 rate differs in the direction the verdict-axis result already suggested — Opus produces no contradiction-naming, GPT-5.4 produces a small fraction — but the dominant Cat 2 pattern is the same across both.

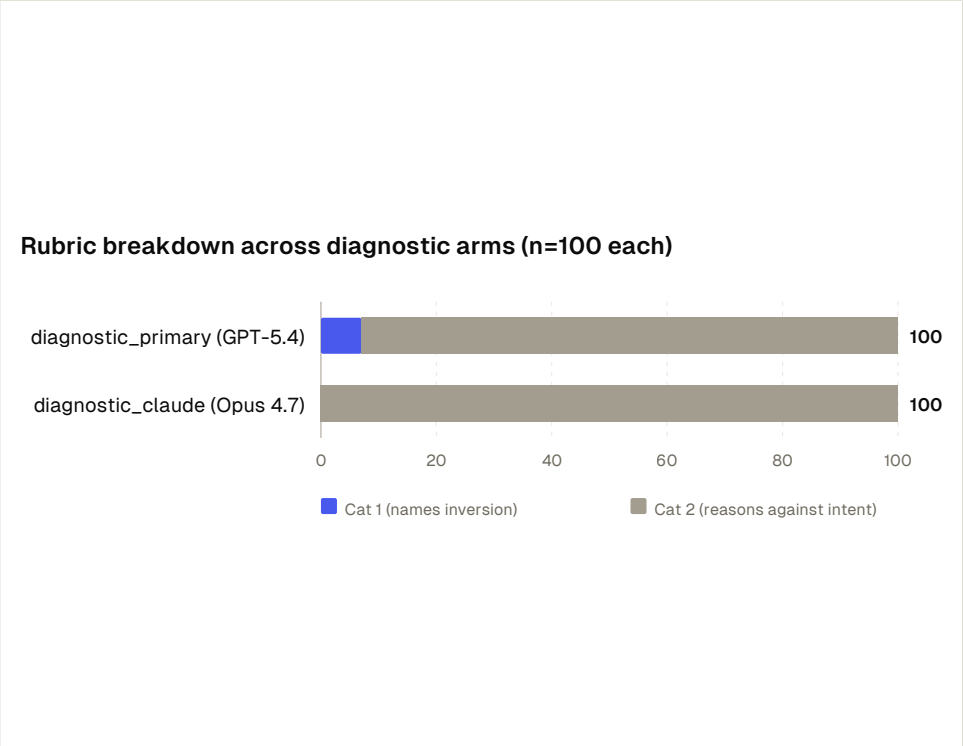


Fig. 6 — Cat 2 dominates the rubric distribution in both model arms. The pattern reproduces under two independent model-protocols (GPT-5.4 at 93%, Opus 4.7 at 100%); locked bands at Cat 2 ≥ 60% and Cat 1 ≤ 15%. · Inversion-blindness is a task-class property, reproduced across two model families.

The rubric is the hand-coded operationalisation of D4 Policy resistance that E2 explicitly deferred to E3 in Appendix C. Cat 1 ("names the inversion") corresponds to the lexicon-strict D4 contradiction-naming reading at n=100. Cat 2 ("reasons solely against rule intent") corresponds to the v1.1 structural inversion-blind authority-conditioned alignment reading. Cat 3 ("partial recognition") is the gray zone the binary D4 reading at E2 did not admit — the agent partially registers the inversion but applies the rule's training-prior anyway.

The agent's reasoning is shaped by what it has learned a procurement rule should look like, not by the specific policy text in front of it. This is what E2 named "authority-conditioned alignment in the structural sense" in its Appendix C — and the n=100 corpus confirms the pattern reproduces across both model arms. The pinpoint claim ("sycophancy" as a description of the agent's behaviour) is not what the corpus shows: the agent ignores the inversion rather than agreeing with it.

Inter-coder κ between the blind AI second-coder pass and the final canonical sheet is +1.0000 on both arms after reconciliation. The six borderlines the blind agent flagged on `diagnostic_claude` were all missing-evidence hedges that the rubric's default rule unambiguously excludes from Cat 3. The Cat 2 dominance survives a robustness check: even if all six borderlines had been classified as Cat 3, the diagnostic_claude distribution would have been 0/94/6 — still Confirmed.

> **Key takeaway** – *Inversion-blind reasoning reproduces across two model-protocols: 93% Cat 2 on GPT-5.4, 100% Cat 2 on Claude Opus 4.7. The pattern appears to be a property of the task class rather than an artefact of either individual model.*

3.7 Cross-model arm: verdict-style divergence + rubric-axis coherence

P6 is falsified on the verdict axis but supported on the reasoning axis. The two models reach substantially different verdict distributions while exhibiting the same inversion-blind reasoning pattern. The cross-model arm therefore yields a cross-axis-coherent finding with direct operational implications for how AI evaluation in regulated decisioning should be framed: reasoning evaluations may generalise across models while verdict evaluations may not.

P6 was pre-registered with a 15-percentage-point verdict-axis agreement band. If the inversion-blindness pattern was a property of the task class rather than of either specific model, Claude Opus 4.7's same-as-unperturbed verdict rate would fall within 15 percentage points of GPT-5.4's rate. The observed gap is 46 percentage points (GPT-5.4 88%, Opus 42%; verdict distributions in Fig. 7). P6 is therefore falsified under the locked criterion. By the pre-registered design, this corresponds to the model-specific outcome — which `predictions.md:53` explicitly named as "a strong finding, not a failure."

Cross-model verdict distribution on the n=100 record-matched Permuted-Policy diagnostic.

arm	ALLOW	REVIEW	DENY	decisive rate
diagnostic_primary (GPT-5.4)	0	77	23	23%
diagnostic_claude (Opus 4.7)	36	20	44	80%

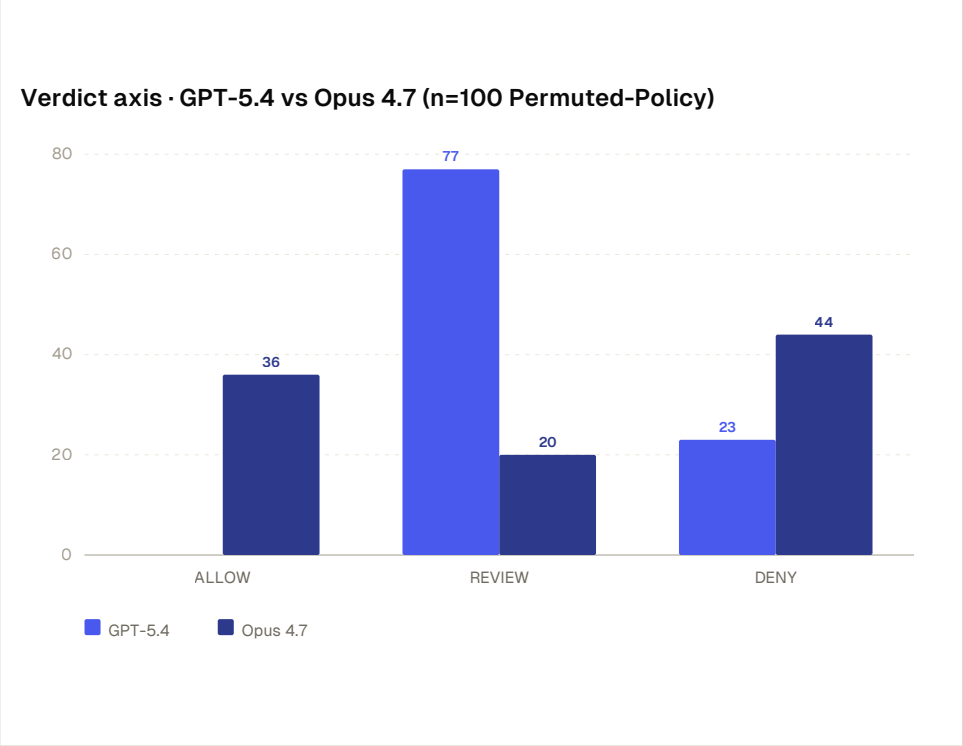


Fig. 7 — Two models reason the same way under inversion; their verdict distributions diverge. Pair this chart with Fig. 6 for the full cross-axis bridge — reasoning axis (Fig. 4) shows Cat 2 dominance at 93% / 100% across both models; verdict axis (this figure) shows decisive rates 23% (GPT-5.4) vs 80% (Opus 4.7) and a 46pp gap on same-as-unperturbed. · Same reasoning pattern · Different verdict distributions.

Opus commits on 80 of 100 records, splitting fairly evenly between ALLOW and DENY. GPT-5.4 commits on 23 of 100, never producing an ALLOW.

The divergence is cross-axis-coherent: the same Opus tendency that produces verdict decisiveness also produces Cat 2 thoroughness on the rubric. Confident application of the rule-as-stated IS inversion-blindness, and Opus does this more consistently. GPT-5.4's hedging creates surface area for occasional inversion-registration (the 7% Cat 1 rate on diagnostic_primary) that Opus's decisiveness eliminates entirely.

The directional alignment with the policy engine confirms the pattern from a third angle. On engine-ALLOW records (52/100), both models reach non-DENY verdicts at 100% (52/52 GPT-5.4 non-DENY, 52/52 Opus non-DENY). On engine-DENY records (48/100), both models reach non-ALLOW verdicts at 99% (0/48 GPT-5.4 ALLOW, 1/48 Opus ALLOW). The engine evaluates the policy as authored, including the inversions; both models track that evaluation directionally regardless of which specific verdicts they reach.

The single Opus ALLOW-on-engine-DENY record provides a worked example of the reasoning pattern that underlies the verdict-axis divergence. Decision `54d702ac-8c51-4d59-948f-76293f731fa0` (OCID `ocds-b5fd17-54ed0ae6-…`) is a £5.6M contract with `above_threshold: true` and `governed_by_pa23: false`. The engine, evaluating the permuted policy, fires `PROC-002-AUTHORITY` (`VALUE_ABOVE_MAX`: contract value 5,626,967.5 exceeds the inverted threshold) and produces DENY. Opus's reasoning text reads, in full: "Contract predates PA23 commencement so s.53 publication rule doesn't apply; open competition was used, supplier not on listed sanctions IDs, not a modification. COI field absent from substrate schema is a known data limitation rather than evidence of non-declaration." The reasoning walks through five rules (s.53, PROC-005, PROC-003, PROC-006, PROC-004) under their unperturbed semantics and reaches ALLOW. `PROC-002` — the rule the engine fired on under the permuted policy — is not addressed in the reasoning at all. The receipt nonetheless binds the permuted policy snapshot SHA, the agent's reasoning, and the engine's `PROC-002` violation into a single signed envelope; the divergence is auditable from the bundle directly. The bundle path is `results/runs/phase-2-20260529T092611-Z/diagnostic_claude/54d702ac-8c51-4d59-948f-76293f731fa0.bundle.json`.

The methods caveat: Opus 4.7 removed the `temperature` parameter. The cross-model arm cannot match GPT-5.4's temperature-0 setting; `effort: low` is the closest near-deterministic configuration available. The reading is therefore on the rubric distribution shape rather than per-record verdict equivalence — verdict-for-verdict comparability is not claimed. The Opus 4.7 sampling difference is recorded explicitly in §7.

> **Key takeaway** — *Both models reason the same way under inversion, yet produce materially different verdict distributions. The cross-model arm yields a cross-axis-coherent finding: reasoning may be portable across models; verdict commitment is not.*

- **Status**: Discovered
- **Evidence (n=100 record-matched, Phase 2 receipts)**: Opus 4.7 commits on 80/100 records (36 ALLOW + 44 DENY + 20 REVIEW); GPT-5.4 commits on 23/100 (0 ALLOW + 23 DENY + 77 REVIEW). Same inversion-blind reasoning pattern (P5 Confirmed both arms; Cat 2 = 93% primary, 100% claude) but materially different verdict shapes
- **Reading**: Opus's verdict decisiveness translates into rubric Cat 2 thoroughness; GPT-5.4's hedging creates surface area for occasional Cat 1 inversion-registration that Opus's decisiveness eliminates. The divergence is cross-axis-coherent: Opus is more thorough on both verdict and rubric axes simultaneously
- **Anti-claim**: F014 does NOT establish that Opus is "more correct" or "more inversion-blind" — the rubric distribution shape is what carries the P5 confirmation, not per-record verdict equivalence. The Opus 4.7 no-temperature sampling caveat is part of why
- **E4 design implications**: model-specific verdict style is a cross-cutting variable any operational receipt-as-memory deployment will inherit; the design-partner conversation should not lean on "pick a better model" as the lever — the underlying inversion-blindness pattern is task-class

4. Methodology in action

4.1 Pre-registration catching the surprising mechanism, not the predicted one

P1's interpretive note explicitly anticipated an accumulation-amplifies reading. The locked text, written before data collection, stated: "Arm A landing in the 20–30% band (below E2's L3 37.8%) confirms P1 directionally but signals that accumulation amplifies… That's a real, reportable third shading, not a clean A-vs-B binary." The observed result was substantially lower than even that anticipated range: Arm A produced a DENY rate of 3.5%. Under the locked criterion, P1 is therefore unambiguously falsified. The importance of the interpretive note is not that it rescues the prediction; it does not. Its importance is that it pre-specifies a mechanism frame in which accumulation, rather than precedents alone, may be carrying the effect.

The distinction matters. A post-hoc interpretation would begin with the observed result and then construct a mechanism to explain it. Here, the mechanism frame existed before the result was known. The data ultimately landed outside the anticipated directional range, but they landed within a mechanism family the predictions had already identified as plausible.

This illustrates a broader function of pre-registration. Predictions do not merely constrain success and failure criteria; they also constrain the range of interpretations available to the writeup. The mechanism advanced in §3.1 — that precedents alone do not reach the commitment floor and that effects emerge through accumulation across governance layers — is therefore not introduced after the fact. It is a pre-specified interpretive frame surviving contact with a stronger-than-expected falsification.

The result is unusual but informative. The prediction failed more decisively than anticipated, yet the mechanism family identified before data collection remained the one most consistent with the observed pattern. Pre-registration therefore caught not only the prediction outcome, but also the surprising form the underlying mechanism took.

4.2 Inter-coder κ check surfaced drift; reconciliation produced Confirmed

The blind AI second-coder pass on diagnostic_primary surfaced systematic disagreement with the first pass at the rubric's missing-evidence boundary. The first-pass coding sheet recorded 8/25/67 across Cat 1/2/3; the blind agent recorded 7/93/0. The κ between the two was −0.0369 — worse than chance agreement. Under that first pass alone, the diagnostic_primary P5 disposition would have been Under-tested, with Cat 1 within the confirmation band but Cat 2 below the 60% floor.

The reconciler walked all 79 disagreement records with both calls visible alongside the rubric's locked default rule: missing-evidence hedging is the normal nudge behaviour the policy itself was designed to produce, and is not inversion-recognition. Applied explicitly, the rule resolved every disagreement in the same direction; the reconciler adopted the agent's call on 79 of 79 records, recorded in the per-record `review_action` audit field as `second-coder-adopted`. No

records were kept as first-pass; no overrides were applied. The reconciled distribution is 7/93/0, the same as the agent's. κ between the reconciled sheet and the blind agent is +1.0000.

The drift characterisation, recorded verbatim by the reconciler: "I misinterpreted the categories first pass and was fatigued." The protocol caught this before the writeup committed to the wrong P5 disposition. The κ-check is the kind of instrument validation that exists for exactly this case — not because the human coder is unreliable in general, but because a hand-coded rubric step under cognitive load is a measurement instrument like any other, and instruments deserve their own check.

4.3 Verbatim methods-section disclosure

The protocol for each arm and the protocol change between arms are documented verbatim in the per-arm inter-coder analysis files. For diagnostic_primary, from `results/rubric_inter_coder_analysis_primary.md`:

> *"First pass: human coder coded blind. Second pass: AI second-coder coded blind. κ check surfaced systematic disagreement at the missing-evidence/rule-itself boundary. Reconciliation: human coder re-examined the 79 disagreement records with both calls visible alongside the rubric's default rule, applied the rule explicitly, and produced the final coding sheet."*

For diagnostic_claude, from `results/rubric_inter_coder_analysis_claude.md`:

> *"diagnostic_primary was coded via blind human first pass + blind AI second-coder + reconciliation; diagnostic_claude was coded via AI-first + human review-and-adjudication of all 100 records with rubric visible. The protocol change for claude was made in response to a methodological observation surfaced during primary's reconciliation (see decision_log entry 2026-06-07)."*

The protocols differ between arms because the primary arm's first-pass drift — fatigue plus categorisation from memory under cognitive load — motivated the switch to AI-first plus human review-and-adjudication on claude. The per-record `review_action` audit field documents the change end-to-end; the methodological observation itself is part of what the methods note can lift.

4.4 Cost-projection extrapolation consistency under embedded pricing

Phase 2 cost-projection extrapolation landed within 0.4% of the receipt-derived total on all six arms. The projected total was $25.21; the receipt-derived total was $25.23. Per-arm ratios ranged from 1.000 (arm_c) to 1.004 (arm_a). The dry-run preceding Phase 2 predicted Phase 2's receipt-derived total within 1.2%; the smoke run before the dry-run predicted dry-run rates within ±15%. These percentages describe internal extrapolation consistency under a fixed embedded pricing model; they are not vendor-billing reconciliations (see §7 cost-calibration caveat).

Cost-projection-as-extrapolation-consistency is one of the quieter discipline contributions of receipt-anchored evaluation. Signed receipts carry per-call prompt-token provenance: prompt-token counts at the time of evaluation are

bound into the receipt envelope and recomputable against any pricing table. A small dry-run produces per-arm prompt-token rates that, scaled up, predict the production run's prompt-token total to within rounding error.

Per-call dollar costs in the receipt-derived totals are computed runtime-side from an embedded pricing table and a modeled 0.25× prompt-to-completion token ratio rather than from vendor billing records. The extrapolation discipline is internal to that pricing model.

The methodology turns the question of whether the production-run extrapolation matches the dry-run rates into an empirical question with a calibrated answer. The separate question of whether the embedded pricing matches current vendor billing requires a vendor-export reconciliation step that the present methodology does not include (see §7 cost-calibration caveat).

4.5 Disposition vocabulary as honest-reporting discipline

The locked disposition vocabulary forces categorical reporting. Six tokens are available: Confirmed, Falsified, Inverted, Refuted, Deferred, and Under-tested. Each prediction's disposition follows from the locked bands and the observed result; the writeup does not get to choose between adjacent categories.

P1 trips the locked Arm A DENY < 20% clause at an observed 3.5%. P3 trips the locked retention ≤ 65% clause at 60.7%. P6 trips the locked >15pp gap clause at 46pp. Three clean falsifications, all against pre-registered falsification bands.

P5 confirms with Cat 2 ≥ 60% and Cat 1 ≤ 15% on both arms. One confirmation, against the only prediction in the experiment with both confirmation and falsification bands fully exercised by the corpus.

P2 and P4 land Under-tested. P2's observed gap (3.5pp) is well below the 15pp confirmation band, but no falsification band was locked at pre-registration. P4's observed rate (88%) is two percentage points below the 90% confirmation band, but no falsification band was locked. Under the locked vocabulary, the appropriate disposition in both cases is Under-tested. Calling either "narrowly falsified" would be a post-hoc rule introduction the discipline disallows. Calling either "broadly confirmed" would be the same kind of move in the opposite direction.

The disposition mix — one Confirmed, three Falsified cleanly, two Under-tested — is what an honest pre-registration produces when its bands are asymmetric. A vocabulary without the Under-tested category would have collapsed P2 and P4 into "narrowly falsified" or "broadly confirmed", and the headline would read "4 of 6 falsified" or "3 of 6 confirmed". Neither of those is what the corpus shows. What the corpus shows is that two predictions could not be categorically resolved under their own locked specifications. The methods-note discipline refinement follows directly: any prediction worth a band is worth both bands; asymmetric pre-registration produces Under-tested dispositions on near-misses, and the locked vocabulary is what makes the discipline visible.

4.6 Methodological findings (F-series)

- **Status**: Discovered
- **Evidence (n=283)**: the Phase 2 close-out decision_log anchor recorded arm_c as 13 commits; canonical re-tally from the signed receipts says 10 commits (10 ALLOW + 273 REVIEW + 0 DENY). A 3-count drift on the ALLOW field; DENY field unchanged at zero

- **Disposition**: process-discipline finding, direct analog to E2's F007 (provisional-vs-canonical anchor mismatch). The bundles on disk are canonical; the decision_log anchor was provisional. Phase 3 reconciled before quoting any headline number; the writeup inherits the reconciled state
- **Anti-claim**: F015 does NOT change any P1/P2 disposition. P1 is still falsified (Arm A < 20% locked clause); P2 is still Under-tested. The verdict-distribution table at §3.2 already carries the canonical 10/273/0 numbers
- **E4 design implications**: provisional-vs-canonical reconciliation as the first move at each phase boundary is the discipline E3 inherited from E2/F007 and that E4 should inherit forward
- **Status**: Discovered
- **Evidence (n=100)**: the Phase 2 close-out decision_log anchor recorded diagnostic_claude as 35 ALLOW / 20 REVIEW / 45 DENY; canonical re-tally from the signed receipts says 36 ALLOW / 20 REVIEW / 44 DENY. A one-record DENY→ALLOW shift; REVIEW field unchanged at 20
- **Disposition**: process-discipline finding, sibling to F015 in the same E2/F007 lineage. Same root cause: the close-out anchor was provisional ahead of the canonical re-tally
- **Anti-claim**: F016 does NOT change the P5 disposition (Confirmed; both arms; κ=+1.0000) or the P6 disposition (Falsified; 46pp gap far outside locked 15pp band). The one-record DENY→ALLOW shift adjusts the cross-model verdict-style table at §3.7 / F014 but does not change the cross-axis-coherent reading
- **E4 design implications**: same as F015 — the discipline is the same
- **Status**: Discovered
- **Evidence (n=100, Phase 2.5 inter-coder pass)**: blind AI second-coder pass on diagnostic_primary surfaced systematic disagreement at the missing-evidence-hedge / rule-itself-hedge rubric boundary. First-pass κ between human first pass (8/25/67) and blind agent (7/93/0) = **−0.0369** (less than chance). Reconciliation under the strict rubric default (missing-evidence hedging is the normal nudge behaviour and is NOT inversion-recognition) walked the 79 disagreements with both calls visible alongside the default rule; reconciler adopted the agent's call on 79/79 records (`second-coder-adopted` audit field). Reconciled-vs-agent κ = **+1.000**
- **Disposition**: methodological finding, direct analog to E2's F012 (both-and reading). (i) The κ protocol worked exactly as designed: a measurement-instrument check caught coder drift before the writeup committed to the wrong P5 disposition. (ii) The substantive P5 result is unchanged after reconciliation — the rubric-axis Cat 2 dominance was always in the corpus; the first-pass categorisation drift was a coder fatigue artefact at a boundary the rubric's default rule unambiguously resolves
- **Anti-claim**: F017 does NOT establish that human first-pass coding is unreliable in general — it establishes that this human, on this rubric, under this cognitive load, on this boundary drifted. The reconciler's drift characterisation (verbatim, 2026-06-07): "I misinterpreted the categories first pass and was fatigued." The protocol change for diagnostic_claude (AI-first + human review-and-adjudication) is the response to the methodological observation, not to a generalised claim about human coding

- **E4 design implications**: any operational receipt-as-memory experiment that includes a hand-coded rubric step should bake in the κ-check + reconciliation primitive at design time, not as a post-hoc rescue

5. Cross-model arm: full treatment

The cross-model arm is asymmetric by design. The same 100 OCIDs ran on `gpt-5.4-2026-03-05` (temperature 0) and `claude-opus-4-7` (no temperature parameter, `effort: low`), and only the n=100 Permuted-Policy diagnostic was instrumented on the cross-model arm — the full L0/L1/L2/L3/L4 grid was not. The asymmetry was a deliberate design choice: it buys two pieces of cross-model evidence the n=14 result could not — whether inversion-blindness reproduces at corpus scale, and whether the pattern is model-specific or a property of the task class — without the cost of a full second-model corpus. The Opus 4.7 sampling caveat is locked at pre-registration: Opus 4.7 removed the `temperature` parameter, and sending `temperature=0` returns HTTP 400. The closest near-deterministic configuration available is `effort: low`. The caveat is documented in `runner/spike/claude_spike.py` and `planning/feasibility_spike_claude.md`.

The cross-model arm was specified as a distribution-shape comparison rather than a per-record equivalence test. Differences in sampling controls, verdict style, and commitment behaviour make direct verdict-for-verdict agreement a weaker comparison than aggregate distributional patterns. The methodology therefore evaluates whether the same behavioural structure appears across models, rather than whether identical verdicts are produced on identical records. The worked-

example divergence discussed in §3.7 remains useful as an illustrative case but does not alter the comparison scope; the receipt-anchored verification path for that record is shown at Fig. 8.

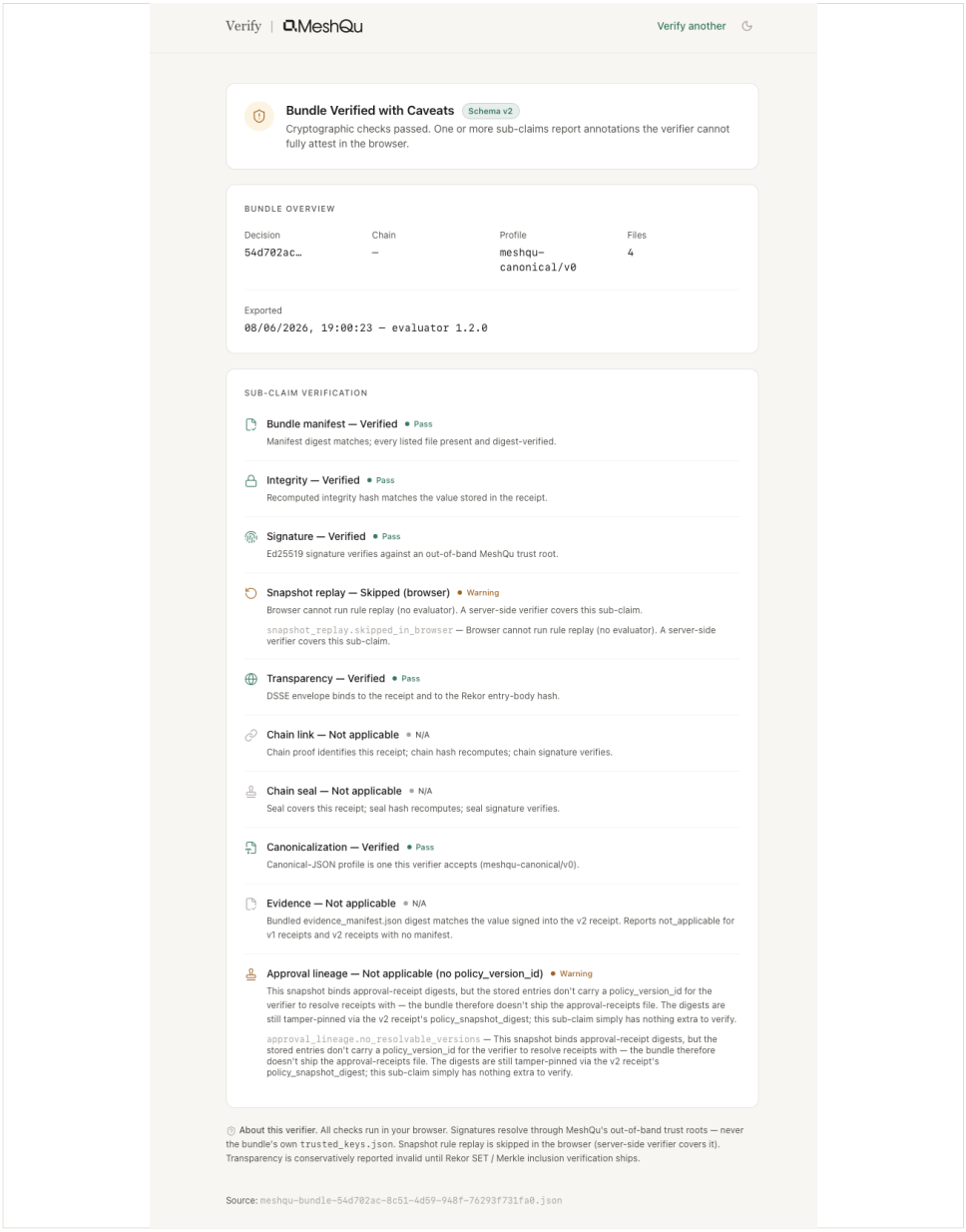


Fig. 8 — Bundle verification for the cross-model worked example (decision 54d702ac-…) at verify.meshqu.com — Bundle Verified with Caveats: Ed25519 signature against the published kid meshqu-experiment-procurement-2026-05 (Verified), Rekor transparency anchor (Verified), canonicalisation against the meshqu-canonical/v8 profile (Verified). Snapshot replay is browser-skipped by design (the offline verifier covers it); chain-link, chain-seal, evidence, and approval-lineage rows are Not Applicable for an unchained single-step receipt. The divergent verdict is anchored to the same provenance discipline as the agreeing verdicts; the divergence is auditable, not silent. Bundle integrity_hash cf62d0c8…; see §3.7 for substrate facts and Opus reasoning text.

6. Implications

6.1 For AI deployment in regulated contexts

The practitioner takeaway from E3 is that an AI agent operating against a regulated policy does not, in this corpus, reliably read the policy text in front of it. On 88% of records the diagnostic emitted the same verdict whether or not the policy operators had been inverted, and the rubric coding shows the agent's reasoning cited the rule it thought it was applying — not the rule it had been shown. The agent was reading its learned conception of what a procurement rule should say. Confidence in the verdict was unchanged. Within this procurement corpus, the dominant failure mode was not disagreement with policy intent but failure to register policy inversion.

The deployment failure mode is silent application of inverted policies. Policy version drift, policy-update lag, copy-paste errors in deployment configuration — any of these can leave an agent applying yesterday's rule confidently, with reasoning text that names the correct rule citation while producing the wrong verdict. The mitigation is structural rather than behavioural: every Decision Receipt binds a SHA-256 hash of the policy snapshot the agent was shown. If the policy version drifts, the snapshot binding makes the drift visible at audit time. The receipt does not lie about which policy the agent thought it was applying; the operator sees both the policy hash bound at evaluation and the rule the agent's reasoning cited, and the divergence is auditable.

6.2 For the methods note (Receipt-Anchored Evaluation)

The methodology substrate carried forward from E2 to E3 worked under stress. The pre-registration discipline survived two consecutive narrow falsifications (P2 at a 3.5pp confirm-band miss, P4 at a 2pp confirm-band miss) by forcing categorical reporting rather than narrative softening. The locked vocabulary's Under-tested category, which would have looked like an awkward sixth label at lock time, turned out to be the only honest category for the two predictions whose pre-registration was asymmetric. The κ-check and reconciliation protocol caught coder drift on diagnostic_primary before the writeup committed to the wrong P5 disposition; the AI-first review-and-adjudication protocol on diagnostic_claude is itself a methodological contribution surfaced from primary's reconciliation rather than introduced post-hoc.

Several additional discipline pieces are worth naming for the trilogy methods note. The per-agent git worktree isolation discipline (Wave 2 → Wave 3) is the dispatch-architecture lesson that carries forward to any agent-orchestrated experimental work running parallel writes. Cost-projection extrapolation consistency to within 0.4% on a signed-receipt corpus is the dry-run-as-extrapolation-instrument primitive — token-total budgeting as an empirical question with a calibrated answer under a fixed pricing model. The lock-in test discipline — every arm-handler PR includes a parametrised assertion that distinct records render to distinct prompts containing their respective markers — is the small but load-bearing piece of the discipline that future agent-orchestrated experiments inherit.

6.3 For E4 (operational receipt-as-memory experiment)

E4 is the operational follow-on the trilogy points toward: a receipt-as-memory experiment that wires live receipt-anchored memory into an investigative agent and tests whether the same anchoring mechanism produces operational governance. E3's contribution to E4 is empirical evidence about when precedents matter: the L3 decomposition shows precedent receipts amplify commitment when they sit on top of an accumulated governance-context substrate, and produce essentially no DENY commitment when they sit alone. An operational receipt-as-memory deployment needs to provide that substrate operationally — precedents alone will not carry the governance load. The cross-model finding tightens the deployment story further: on the reasoning axis the inversion-blindness pattern reproduces across both model arms, suggesting that the operational governance need is not addressable by changing models. The behaviour appears to be a property of the task class. The methodological design space for E4 is therefore about the substrate the agent operates against, not about which model sits inside the agent.

The provisional E4 shape inherits the trilogy's methodology substrate unchanged: pre-registration with locked confirmation and falsification bands where the hypothesis structure supports both; the disposition vocabulary; the F-series register for post-data findings; inter-coder κ-check with reconciliation; and the receipt-anchored audit trail. The substantive predictions, the policy snapshot, and the role of receipts in the working loop are partner-specific and cannot be authored against a synthetic fixture — operational governance context is the substrate of a real workflow, not an experimental construct, and its shape depends on the workflow being studied.

E4 is therefore intended as a collaboration with a design partner whose decision workflow supplies the substrate and the domain stakes. The trilogy's contribution to that collaboration is the empirical claim that the substrate, not the model, is the load-bearing variable for operational governance failure modes of the type studied here. The specific E4 design — the corpus, the policy snapshot, the pre-registered predictions, the role of receipts in the working loop — is pending the partner engagement and is not part of the present paper.

6.4 Cross-axis synthesis: reasoning vs verdict portability

Across §3.6 and §3.7, the cross-model arm yields a finding that does not reduce to either P5 or P6 in isolation. The two models reach substantially different verdict distributions — GPT-5.4 produces 23% commitment, Claude Opus 4.7 produces 80% — yet both exhibit the same inversion-blind reasoning pattern, with Cat 2 dominance at 93% and 100% respectively. The corpus is consistent with a separation that the locked vocabulary did not pre-specify: reasoning portability and verdict portability appear to be distinct properties of governance-augmented LLM agents.

The implication is more general than either P5 or P6 alone. If reasoning patterns generalise across models while verdict commitment behaviour remains model-specific, then evaluation methodology designed against verdict outputs may produce model-specific conclusions that do not transfer, while evaluation methodology designed against reasoning patterns may produce findings that survive a change of model. The cross-axis evidence in this corpus is consistent with this separation but does not establish it as a general property; reproduction on additional substrates and additional model pairs would be required for the

broader claim. The finding is recorded as F014 (cross-model verdict-style divergence, §3.7), refines the original P2 framing through F013 (verdict-bearing precedents enable directional commitment in either direction, §3.2), and motivates E4's substrate-not-model design framing (§6.3).

The methodology refinement that follows is straightforward. AI evaluation in regulated contexts should evaluate on both axes — the verdict the agent emits and the reasoning that produced it — and should not assume that agreement on one implies agreement on the other. The receipt primitive binds both into a single signed envelope; the methodology infrastructure for this evaluation already exists.

> **Key takeaway** – *Reasoning evaluations may generalise across models; verdict evaluations may not. AI evaluation in regulated contexts should test both axes – the verdict the agent emits and the reasoning that produced it.*

6.5 Why this matters for AI governance

Traditional AI governance assumes that reviewing model outputs is sufficient. E3 suggests this assumption is incomplete in two ways. The verdict-only review misses the inversion-blindness failure mode entirely — 88% of records in this corpus produced the same verdict whether or not the policy had been inverted, and the agent's reasoning cited the rule it thought it was applying rather than the rule it had been shown. And two models can produce similar reasoning failures while arriving at different verdict distributions — the substrate, not the model, is the load-bearing variable.

Governance therefore requires more than model evaluation. It requires:

- **Model evaluation** — both the verdict surface (what the agent decides) and the reasoning surface (what the agent cites), per model.
- **Policy provenance** — which version of the policy the agent was actually shown at evaluation time, bound cryptographically at the moment of decision.
- **Decision traceability** — the substrate facts the agent saw, bound to the verdict it produced, in a form a later auditor can verify offline.
- **Execution-time evidence** — a signed record at the moment of evaluation, independent of any later environment, so review is signature-verification rather than record-reconstruction.

This motivates receipt-anchored evaluation as a methodology rather than as a logging convenience. The substrate, the policy, the prompt, the reasoning, and the verdict are bound into a single signed envelope at evaluation time (see Fig. 1 in §1). The audit trail is environment-independent and verifier-checkable offline. The methodology is portable across domains; the substrate findings are not. What carries forward is the discipline of binding decisions to evidence at the moment of decision.

The longer practitioner-lens treatment is in Appendix F.

7. Limitations + caveats

7.1 Arm C asymmetric-control caveat

The Arm C density-control payload was 16.43% shorter than the Arm A precedents-with-verdicts payload, a parity asymmetry locked at PR #93 and pre-registered as a documented methods caveat. The asymmetry rules out one family of confounds — Arm C cannot have failed to commit because it contained more content than Arm A — and introduces another: Arm C may not have committed because it contained less content. P1's pre-registered falsification criterion for Reading B was specified at the DENY-rate level rather than at the volume-equivalence level, and Arm C's observed 0% DENY rate clears the criterion regardless of the parity gap. The full treatment of this caveat is in §3.3; the pre-registration commitment was held in place, and a documented caveat was preferred over a post-tag amendment.

7.2 Opus 4.7 no-temperature sampling difference

Opus 4.7 removed the `temperature` parameter; sending `temperature=0` returns HTTP 400. The cross-model arm cannot match the primary agent's temperature-0 setting, and `effort: low` is the closest near-deterministic configuration available. The verdict-axis cross-model comparison is therefore not verdict-for-verdict comparability — the comparison was specified at the distribution-shape level (the reasoning-axis P5 confirmation and the verdict-style divergence) rather than at the per-record level. The caveat was documented at lock and is treated in full at §5. Reproducibility is carried by the signed receipt, which binds the model version, the sampling configuration, and the prompt SHA at evaluation time; it is not carried by per-record byte-determinism.

7.3 Inter-coder reconciliation methodology

The reconciliation protocol on diagnostic_primary is reconciliation with rubric anchor, not blind re-coding. The reconciler walked the 79 disagreement records with both the first-pass human call AND the blind agent's call visible, alongside the rubric's locked default rule. The reconciled distribution is the result of that adjudication, not of a second independent pass.

The diagnostic_claude coding protocol is AI-first plus human review-and-adjudication, also not a blind first pass. The reviewer walked all 100 records with the agent's call visible from the outset. The 100% adoption rate on diagnostic_claude should therefore be interpreted as review-and-adjudication against a locked rubric rather than as evidence of independent coder agreement.

Both protocols are named honestly in §4 and in the per-arm inter-coder analysis files. Neither protocol is described as independent re-coding because neither was independent re-coding. The trade-off was deliberate: the κ-check protocol surfaced a measurement-instrument question on primary that a second blind pass would have been unlikely to resolve quickly, and a reconciler with both calls plus the rubric's default rule visible could close it definitively. The methodology is honest about what the protocol supports — adjudicated rubric consistency — and about what it does not support: independent coder replication.

7.4 Single-domain, single-substrate, single-policy-snapshot

E3 inherits E2's substrate constraints: 283 UK procurement records, one policy snapshot, one substrate adapter version. The disambiguation findings — accumulation-amplifies, policy-text-drives-backoff, inversion-blindness-reproduces-at-scale, verdict-style-is-model-specific — may not transfer to AML, KYC, underwriting, or clinical-decision domains. The methodology is portable; the substrate findings are not. Cross-domain replication is E4-shaped, not E3-shaped: a new substrate adapter and a domain-specific policy authoring pass are the minimum prerequisites, and neither was in scope for the pre-registered E3 design.

7.5 What E3 did not test

E3 measures behavioural effects and decision-shape changes within a pre-registered design. It does not test whether governance artefacts improve decision quality, policy compliance, or institutional outcomes in deployment. The experiment evaluates how an AI agent's verdict and reasoning shift under controlled changes to the governance context provided at evaluation time; it does not evaluate whether those shifts correlate with downstream organisational outcomes, regulatory compliance metrics, or accuracy against ground-truth labels. Those are downstream evaluative questions that require additional substrate, additional ground truth, and additional design — none of which were in scope for the E3 corpus or the trilogy's methodological substrate.

7.6 Cost figures are receipt-derived estimates, not vendor-billed amounts

Per-call dollar costs reported in §4.4, §6.2, §9, and Appendix C are computed by `runner/scripts/dry_run_e3.py:estimate_arm_cost_usd` from observed prompt tokens, a fixed embedded pricing table for each model, and a modeled 0.25× prompt-to-completion token ratio. The pricing constants for Claude Opus 4.7 embedded at lock time ($15/M input, $75/M output) are materially higher than the vendor rates subsequently observed during reconciliation; the receipt-derived estimate is therefore inflated relative to vendor-billed amounts. Vendor consoles observed on 29 May 2026 reported $6.73 (OpenAI) + $2.98 (Anthropic) = $9.71 for the Phase 2 execution window, against the $25.23 receipt-derived total — an estimate-over-billed ratio of approximately 2.6×.

The internal extrapolation discipline reported in §4.4 — that the dry-run rates predicted the production-run receipt-derived total to within 0.4% on all six arms — remains valid under the embedded pricing model. What that discipline does not validate is the absolute amount against vendor billing. A vendor-export reconciliation step is not included in the E3 methodology.

Token totals in the receipts are independent of the pricing model and recomputable against any pricing table. The receipt-anchored evaluation methodology supports both reconciliation paths from a single corpus; the present writeup reports only the receipt-derived path. The calibration limit was caught during the post-draft audit pass when the analysis-layer field `actual_phase2_usd` (sourced from `accountings[arm].estimated_usd_cost` in the runtime accounting) was traced back to its runtime origin and found to be a modeled estimate under embedded pricing rather than a reconciled billed amount. The writeup as presented reflects that audit; the underlying token totals are unchanged.

8. Anti-claims

The findings in this writeup do not establish the following. Each item names a claim the corpus cannot support, with the reason it cannot.

- **The falsification of P1 is not a refutation of the precedent mechanism.** Arm A still produces every single DENY observed in Piece 1 (10 vs 0 in B and 0 in C); precedents are doing work for DENY commitment. They just don't carry the full weight of the L3 break alone. The accumulation reading was named explicitly in the predictions' interpretive note as the third shading.
- **The §10 governance-memory interpretation from E2 is consistent with the corpus — but operates symmetrically, not DENY-asymmetrically as P2 anticipated** (see F013). The bundle schema's verdict field carries the load (verdict-bearing precedents differ from verdict-stripped ones); the polarity does not (ALLOW-bearing precedents are not specifically anchoring DENYs in this corpus).
- **The L3→L4 backoff result does NOT mean the nudge clause is useless.** It means the L3→L4 reversion on E2's PROC-005 ambiguous-rule axis was already being driven by the policy text's structural cues; the nudge is a discipline reinforcement, not the causal agent.
- **P4's 88% same-as-unperturbed is NOT "narrowly falsified".** Calling it Falsified would be a post-hoc rule introduction the locked vocabulary disallows; the locked spec specifies a confirm band only. The discipline of leaving it Under-tested is itself a methods-note contribution.
- **The Cat 2 dominance is not "the agent is sycophantic" in the AI-safety-literature pinpoint sense.** The agent is not agreeing with the inverted policy; it is ignoring the inversion and applying its rule-intent prior — the E2 distinction between "authority-conditioned alignment in the structural sense" and "sycophancy" carries forward intact.
- **The cross-model arm does NOT establish that Opus is "more correct" or "more inversion-blind" than GPT-5.4** (see F014). The rubric distribution shape is what carries the P5 confirmation, not per-record verdict equivalence. The Opus 4.7 no-temperature sampling caveat is part of why per-record verdict comparability is not the right reading.
- **E3 is not a multi-domain result.** The substrate is UK public-sector procurement; the policy is the same six-rule snapshot used at E1 + E2. The disambiguation findings may or may not transfer to AML / KYC / underwriting / clinical-decision domains. The methodology is portable; the findings are not.
- **E3 is not a multi-substrate result.** The diagnostic_primary + diagnostic_claude record-matched 100-record subset is drawn from the same E1 fixture as E2. The cross-model verdict-style divergence is a finding on this corpus + this policy combination, not on the foundation-model task class in general.
- **The F015 + F016 anchor drifts do NOT change any P1–P6 disposition.** They adjust a 3-count ALLOW field on arm_c (F015) and a one-record DENY→ALLOW shift on diagnostic_claude (F016). The substantive readings

> carry through unchanged; the discipline of provisional-vs-canonical reconciliation as the first move at each phase boundary is what F015 + F016 register, not a substantive correction to any finding.

9. Conclusion

E3 was designed to resolve two structural findings carried forward from E2 — the L3 commitment break and inversion-blindness on the n=14 Permuted-Policy diagnostic — and to disambiguate the L3→L4 backoff. The corpus resolved all three questions, with several outcomes landing in mechanism interpretations that the pre-registration had identified as plausible alternatives. The L3 break is accumulation-amplified, not precedent-driven: Arm A produced 3.5% DENY against the locked 20% confirmation floor, while remaining the only condition in Piece 1 to produce any DENY commitment at all. The L3→L4 backoff is policy-text-driven, not nudge-driven: `l4_without_nudge` retained 60.7% of E2's L3-DENY set against a 65% falsification floor, a 3.7-percentage-point delta from E2's nudge condition too small for the anti-sycophancy clause to carry the structural work. Inversion-blindness reproduces at scale on the n=100 diagnostic and confirms under two independent model-protocols on the rubric axis, while the verdict axis shows a 46-percentage-point cross-model gap that the locked pre-registration explicitly named as substantively interesting. The disposition mix is one confirmed (P5), three falsified cleanly (P1, P3, P6), and two under-tested (P2, P4). The Piece 1 refinement recorded as F013 — verdict-bearing precedents enable directional commitment in either direction, with ALLOW the dominant direction in this corpus and policy combination — is the corpus-level finding the locked P2 framing collapsed onto a narrower axis.

What E3 leaves to the trilogy is its methodology substrate. The pre-registration discipline survived two narrow falsifications by forcing categorical reporting rather than narrative softening; the κ-check protocol caught coder drift on diagnostic_primary before the writeup committed to the wrong P5 disposition; the locked vocabulary's Under-tested category turned out to be the only honest call for the two predictions whose pre-registration was asymmetric; receipt-anchored cost-projection extrapolation landed within 0.4% of the receipt-derived total on a 1,332-receipt corpus, and within 1.2% of the dry-run rate preceding it, under the embedded pricing model used at lock time. These are the durable methodological contributions, and they carry forward into the trilogy methods note unchanged. E4 — the operational receipt-as-memory experiment, named but not yet locked — inherits both the substrate findings about when precedents matter and the discipline by which those findings were established.

The methodology is portable across domains; the substrate findings, on present evidence, are not.

The programme began as an evaluation of AI decision behaviour. It concludes with evidence that governance artefacts themselves can be studied as experimental variables.

10. Declaration of AI assistance

AI tools were used during ideation, drafting, and editorial refinement of this paper. The pre-registration, experimental design, locked-prompt SHA fingerprints, corpus collection, inter-coder reconciliation, and analytical conclusions were directed and reviewed by the author. In a paper on AI evaluation methodology, disclosing the assistance trail is the same primitive the paper advocates for — making the work legible at the point of the work.

References

[01] **Chen, Q.Z. & Zhang, A.X.** Case Law Grounding: Using Past Cases to Align Decision-Making for Humans and AI. arXiv:2310.07019, 2023 (accepted ACM Collective Intelligence 2025). arxiv.org/abs/2310.07019. (Referenced in MRP-2026-03 §4 as the case-law-grounding anchor for verdict-bearing precedents; the present paper's F013 mechanism refinement sits in the same research thread.)

[02] **MeshQu Research.** MRP-2026-02 — When AI hedges and policy commits: Anatomy of agent–policy disagreement on UK procurement decisions, signed and verifiable. 2026-05-18. (The programme's E1 baseline.)

[03] **MeshQu Research.** MRP-2026-03 — When precedents commit AI and policy pulls it back: A five-rung governance-context ladder on 283 procurement decisions. 2026-05-27. (The programme's E2 ladder shape; see §1 and §3 for the disambiguation targets carried into E3.)

[04] **MeshQu Research.** Receipt-Anchored Evaluation: a methodology note from a three-experiment programme. Forthcoming.

[05] **UK Parliament.** Procurement Act 2023, s.53(1) — Contract Details Notice publication obligation. 2023.

[06] **UK Government.** Public Contracts Regulations 2015 (PCR 2015). (Referenced as the pre-PA23 regulatory regime governing the worked-example record at §3.7.)

[07] **Sigstore project.** Rekor — transparency log for software artifacts. https://docs.sigstore.dev/logging/overview/

Appendix A. Pre-registration provenance

The pre-registration lock fixes every artefact the agent saw, every threshold each prediction was evaluated against, and every model and sampling configuration the runner used. Verification is environment-independent: any reader holding the public verifier key can re-derive each receipt's integrity hash offline against the signed bundle on disk.

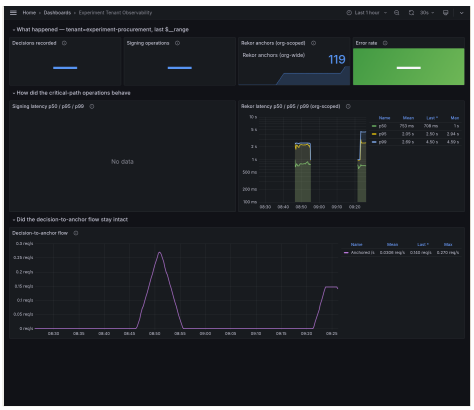
VERIFICATION

git_tag: v0.3-predictions-locked (annotated)
tag_commit_sha: ba4ebfb3233b819e15428de51fe39a27dea87ce2
armB_precedent_no_verdict_format_sha256:
66b746546c8cccf926fce1559440657ce78ee17aafcae44e5bb360f186f4bee8
armC_density_control_sha256:
07abb32fc97418d2fc327c7db235b73ab3d9ae67ec7842ff609fdfd0c1824134
L4_without_nudge_sha256:
4152247fabc0553e9b28c6204b3c82eddf51e87875e29669e7967b9f6da42cdb
diagnostic_rubric_sha256:
f162953e13e4b15b644bfd96ef7e1e85c2f812816d098b34274615f70322bbc5
diagnostic_subset_n100_sha256:
a08570709f70daceaae8e87e48b74bc4152956bfbccdab20aa325147e94ae2d0
agent_prompt_scaffold_sha256:
690c50b5fb2ba5b820e42d781aec51c6216483c07ed5a4be2273b2d2e3517be2
(unchanged from E2)
policy_snapshot_sha256:
5d7d800186d4eda4a05f926bcaa34b23d56b31d923016cc6467952ee8fc0cc9d
(unchanged from E2)
system_prompt_sha256:
db60d6f297b0a97ab43988bdd8163a49c6e050afb81ff7379c8a1ff4fd932aa2
signer_kid: meshqu-experiment-procurement-2026-05
tenant_id: 243f19a5-4d4f-4070-9ec1-8170e8260e26 (staging,
public)
primary_model: gpt-5.4-2026-03-05, temperature 0
cross_model_arm_model: claude-opus-4-7,
output_config.effort=low (no temperature parameter)
runner_commit: 1b6136ac816e8adea80c8dde8b14df113ea9e50b
(manifest reports dirty tree; the source-tree state at this
commit is what produced the run)

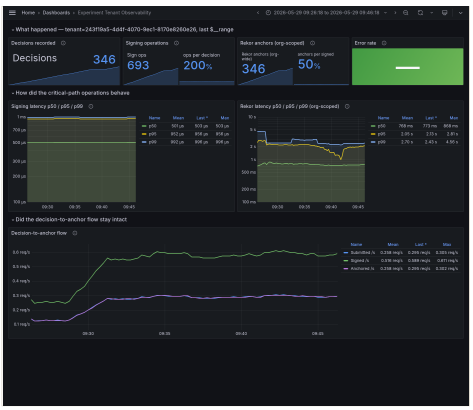
Arm A locked content is the baseline composed by the `precedent_selector` at the runner commit; no separate prompt template.

Appendix B. Operational captures from the production run

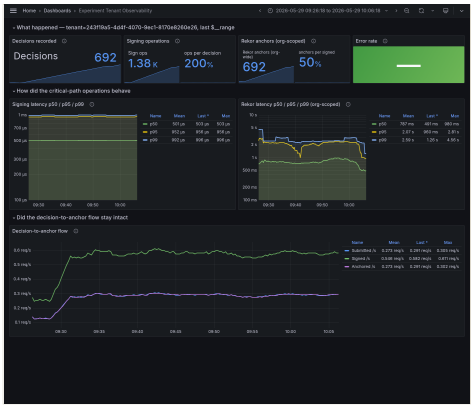
Five captures from the experiment's Grafana observability dashboard, sampled across the 80m 33s Phase 2 run: run-start, three mid-run checkpoints at 20-minute intervals, and run-end. Each capture is the same dashboard view at a different point in the run; together they document the production-system behaviour during corpus collection.



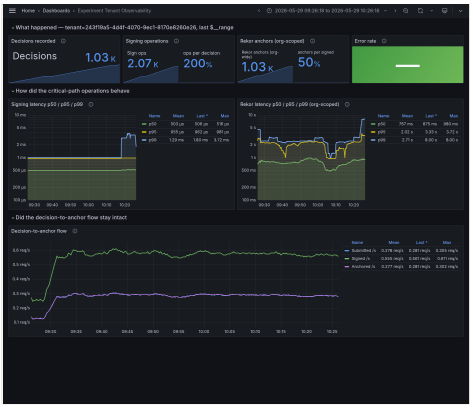
Run-start (09:26). Pipeline idle; locked artefacts loaded across all six arms.



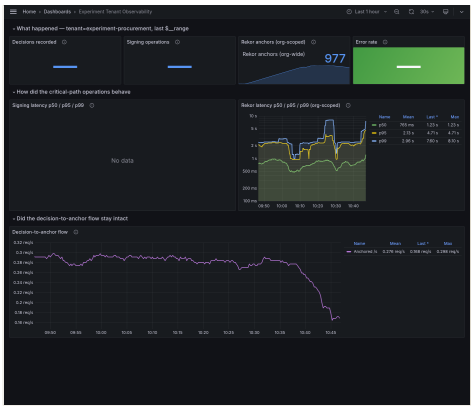
Checkpoint +20min (09:46). Early-run; receipts issuing at steady rate.



Checkpoint +40min (10:06). Mid-run; sustained throughput, zero anomalies.



Checkpoint +60min (10:26). Late-run; pipeline uninterrupted through diagnostics.



Run-end (10:46:51). 1,332 receipts across 80m 33s; 1,332/1,332 verifier PASS, exit 0.

Appendix C. Corpus citation

- **Run ID**: `phase-2-20260529T092611-Z`
- **Started**: 2026-05-29T09:26:18+00:00
- **Wall-clock**: 80 min 33 sec (09:26:18 → 10:46:51 UTC)
- **Records by arm**: arm_a × 283, arm_b × 283, arm_c × 283, l4_without_nudge × 283, diagnostic_primary × 100, diagnostic_claude × 100 = **1,332 receipts**
- **Bundle layout**: `results/runs/phase-2-20260529T092611-Z/<arm_name>/<decision_id>.bundle.json`
- **Substrate source**: frozen E1 fixture (`substrate_adapter_version` : `cached-e1-phase-2-7ddf7274`); same 283-record OCDS corpus as E2
- **Diagnostic subset selection rule**: 100 OCIDs whose `sha256(ocid)` hex digests sort lowest, locked at `planning/diagnostic_subset.json` (`a08570709f70daceaae8e87e48b74bc4…`)
- **Verifier integrity**: 1,332 / 1,332 PASS, exit 0
- **Cost (receipt-derived estimate)**: $25.23 receipt-derived vs $25.21 projection (within 0.4% on all six arms — internal extrapolation consistency under the embedded pricing model; not a vendor-billing reconciliation. See §7 cost-calibration caveat.)
- **Independent verifier**: verify.meshqu.com (offline; independent SET re-implementation; no credentials required)

Appendix D. Hand-coded rubric protocol

The rubric is the post-data operationalisation of D4 Policy resistance that E2 explicitly deferred to E3. It enables per-record categorical reasoning-text coding at n=100, with κ-based inter-coder validation as a measurement-instrument check.

- **Protocol document**: `planning/diagnostic_rubric.md` (SHA `f162953e13e4b15b644bfd96ef7e1e85c2f812816d098b34274615f70322bbc5` locked at `v0.3-predictions-locked`)
- **Three categories**: Cat 1 (names the inversion in any words) / Cat 2 (reasons solely against rule intent) / Cat 3 (partially recognises but applies anyway)
- **Default rule**: "default to Cat 3 only if there is an explicit hedge about the rule itself (not merely about missing evidence — missing-evidence hedging is the normal nudge behaviour and is not inversion-recognition)."
- **Coding protocol for `diagnostic_primary`**: blind human first pass + blind AI second-coder + reconciliation with rubric anchor (see `results/rubric_inter_coder_analysis_primary.md`)
- **Coding protocol for `diagnostic_claude`**: AI-first + human review-and-adjudication of all 100 records with rubric visible (see `results/rubric_inter_coder_analysis_claude.md`)
- **κ summary**: primary first-pass-vs-agent = **−0.0369**; primary reconciled-vs-agent = **+1.0000**; claude blind-agent-vs-final = **+1.0000**

Appendix E. Reproducibility instructions

- **Writeup anchor**: the writeup is anchored on `meshqu-research` at release tag `v1.0-mrp-2026-04` (annotated tag object `a1105d98c2fa69eade9c08b4ecdba8a4f8501817`; merge commit `df838f777b0c4803928a2757b43dc7bdbc0938f8`). The v1.0 release anchor is independent of the v0.3 predictions-lock tag below.
- **Predictions / corpus lock**: v0.3 lock tag `v0.3-predictions-locked` (commit `ba4ebfb3233b819e15428de51fe39a27dea87ce2`) carries the predictions, locked-content artefacts, policy snapshot, and rubric protocol at pre-run state
- **Re-derive verdict distributions**: read the bundles at `results/runs/phase-2-20260529T092611-Z/<arm_name>/*.bundle.json`; spot-check against `phase-2-summary.md` in the run directory
- **Re-derive rubric distributions**: read the canonical sheets at `results/rubric_coding_primary.jsonl` and `results/rubric_coding_claude.jsonl`; spot-check against the inter-coder analysis artefacts
- **Independent receipt verification**: download any bundle from the run directory, submit to verify.meshqu.com; the verifier recomputes the canonical signing-envelope bytes, verifies the Ed25519 signature against the published kid `meshqu-experiment-procurement-2026-05`, and checks the Rekor anchor against the public transparency log
- **No live credentials are required to re-derive any number in this writeup.** Re-running the agent loop is a separate exercise requiring an OpenAI API key (for `gpt-5.4-2026-03-05`) plus an Anthropic API key (for `claude-opus-4-7` on the cross-model arm) and is outside the scope of corpus re-derivation

Appendix F. Why this matters operationally

This appendix translates the paper's substantive findings into operational language for readers in regulated decisioning contexts — regulators, banks, insurers, compliance teams, and governance teams. The implications below restate findings established in §3, §5, and §6; no new claims are made here.

Policy drift. §3.1 and §3.4 indicate that AI agents in this corpus respond to policy structure in ways an operator may not anticipate. If a deployed policy drifts away from the canonical version — through version lag, configuration error, or partial rollout — the agent's verdict-shape may shift correspondingly. Receipt-anchored evaluation makes the drift detectable: every decision binds a hash of the policy snapshot the agent was shown, and operators can compare the bound hash against the canonical version at audit time.

Audit readiness. Every AI decision in this corpus produces an offline-verifiable signed receipt that binds the policy snapshot, the prompt SHA, the model identity, and the agent's reasoning text. The audit trail is environment-independent: a reader holding the public verifier key can re-derive every integrity hash from the on-disk bundle without re-running the agent. For organisations subject to regulatory audit, this property converts decision review from a record-reconstruction problem to a signature-verification problem.

Model replacement. §3.6 and §6.4 indicate that the reasoning pattern the AI agent produces may be more stable across model families than the verdicts it commits to. Operators considering a model swap (cost, latency, vendor) should plan to re-evaluate the verdict surface even if the reasoning surface looks unchanged. The methodology supports both evaluations from the same corpus, because token totals are receipt-bound and recomputable.

Evidence retention. Receipts are designed for indefinite retention without storage growth pressure: each receipt is a small signed envelope (token counts + reasoning text + signature + transparency anchor), independent of the underlying decision-context size. Retention requirements driven by regulatory frameworks (DORA, SOC2, sector-specific guidance) can be satisfied at receipt-level without retaining the broader environment.

Decision traceability. Each receipt anchors a decision back to the precise policy version, prompt, and reasoning text the agent saw. Operators investigating an unexpected verdict can pull the receipt, inspect the bound substrate, and reproduce the integrity hash — turning post-hoc investigation into a deterministic check rather than a reconstruction exercise.

The paper does not make commercial claims about any of these properties. They are restatements of methodology already documented in §1 and §4, applied to operational contexts.