The Evidentiary Control Protocol: A Preregistered Governance Layer for Auditable AI Evaluation

Anonymous authors
Paper under double-blind review

Separating execution validity, evidence integrity, protocol-defined outcome, and scientific interpretation across registered evidence lines and an independent audit programme

Abstract

Scores reported by AI benchmarking frameworks do not, by themselves, establish how an observed answer was produced. When the question of interest is whether an outcome was obtained under a specified experimental condition — rather than through retrieval, memorization, model-mediated inference, or answer-bearing artifacts in the evaluation environment — the score is not the evidence. The evidence is a controlled chain from case construction to outcome classification.

We present the Evidentiary Control Protocol (ECP), a preregistered, leakage-controlled governance layer that fixes a ten-stage lifecycle — case authoring, structural and novelty review, environmental knowledge-boundary verification, registration, conditioned execution, raw-evidence preservation, independent audit, mechanical outcome classification, descriptive statistical analysis, and representation-bias disclosure — and enforces one central separation discipline: execution validity, evidence integrity, protocol-defined outcome, and scientific interpretation are four distinct claims, recorded separately and never merged. ECP does not measure a model's internal state and does not adjudicate cognitive claims.

The paper demonstrates the protocol operating across an evidence portfolio that now extends beyond the original five lines (the T1 pilot; the M1 pilot across six hash-frozen systems; the REASON-2 campaign; the REASON-3 defect-preservation/repair/re-execution chain; and the M2 comparative recording campaign) to an independent audit programme that applies the same discipline to externally published claims under bounded channels: a commercial structured-output audit of Together AI (A14: schema validity 24/24 = 1.000 under the registered configuration), an adversarial claim-injection audit (A15: detection 1.00 on levels L1–L3, silent acceptance 0.00), an adaptive adversarial red team (A16: 141 submissions across four valid runs, 0 accepted, with a documented localization-metric boundary and final classification INCONCLUSIVE), a structured-output reliability audit of OpenAI's published claim (A17: 60/60 schema-valid, verdict CLAIM_HOLDS_STRONG, bounded to the tested corpus, snapshot, and gateway channel), and a bounded directional audit of Anthropic's published prompt-injection robustness claim (A18: mechanical injection-success 3/30 = 0.100 on Sonnet 4.6 versus 14/30 = 0.467 on Sonnet 5, non-overlapping 95% Wilson intervals — a directional difference opposite to the published claim's direction under the tested conditions).

We are explicit about what this establishes and what it does not. The portfolio establishes the protocol's operation — registrations held, evidence preserved, adverse findings kept intact — and demonstrates that ECP-style independent audits can examine third-party commercial claims with bounded channel disclosure. It does not validate ECP as a scientific instrument for measuring reasoning, does not replicate any vendor's internal evaluation, and does not convert any bounded finding into a general claim.

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

Every quantitative statement is bound to a sealed artifact in the accompanying claim ledger.

Keywords: AI evaluation; evidentiary protocol; preregistration; evidence integrity; claim boundary; independent audit; outcome classification; AI governance.

1 Introduction

Benchmark scores are routinely reported as evidence of capability, yet they do not identify the pathway by which an answer was produced: memorization, retrieval, training-data familiarity, model-mediated inference, implementation-specific encoding, or answer-bearing artifacts in the evaluation environment can each yield the same score. Three pressures make the distinction consequential. First, scientific reproducibility Pineau et al. (2020) requires that experimental conditions and evidentiary pathways be specified well enough for an independent examiner to determine how a result was obtained. Second, regulatory and conformity-assessment regimes increasingly require documented behavior, traceability, and auditability of AI systems, raising the value of controlled, auditable procedures where the methodology is left open. Third, hybrid AI architectures interleave deterministic components with optional model-based inference, so a capability claim may attach to a specific architectural layer rather than to the system as a whole.

This paper presents ECP as a methodological framework for exactly that problem. ECP is not a benchmark. It does not measure internal model state, does not adjudicate whether a system "reasons", and does not replace benchmark suites or execution infrastructure. It is an evidentiary governance layer: a preregistered chain of controlled artifacts that separates what the execution was (execution validity), whether the record is intact (evidence integrity), what the registered classifier recorded (protocol-defined outcome), and what a scientist may later claim (scientific interpretation — an act the protocol itself never performs). The scientific posture is deliberately asymmetric: the protocol is not assumed to succeed; the evaluated systems are not assumed to possess any target capability; and neither positive nor negative results are interpreted beyond the registered experimental boundary.

The integrated contribution of this paper is twofold. First, the protocol's operation is demonstrated across registered evidence lines spanning both major execution conditions, including under adverse findings (Section 4). Second — and new to this draft — an independent audit programme applies the same registered discipline to externally published, commercially material claims: provider structured-output guarantees (A14, A17), the integrity of the evaluation infrastructure itself under adversarial pressure (A15, A16, A11), and a published prompt-injection robustness claim (A18), plus earlier fault-family audits (A4, A9, A10) and condition-contrast/leakage audits (E1–E3). Each audit is bounded to its registered scope, and one — A18 — returned a directional finding opposite to the published claim's direction under the tested conditions, which the paper reports with its full disclosure set while claiming neither falsity nor replication.

2 The ECP governance layer

ECP defines ten stages, each producing a controlled artifact while preserving the historical record of preceding stages; stage semantics are registered in the frozen parent manuscript (§4) and are unchanged by this paper. (1) Case authoring — premises, question, registered answer, verification rule, and leakage review under a fixed answer vocabulary. (2) Structural review and novelty assessment — recorded before any execution. (3) Environmental knowledge boundary — the environmental pathway is characterized and bounded. (4) Registration — cases, ground truth, condition, and success criterion are hash-sealed before any execution; nothing downstream may alter registration. (5) Conditioned execution — one shot per registered case under the registered condition (Model-Disabled or

Model-Only), with the condition's pathways controlled or fail-closed. (6) Raw-evidence freeze — outputs, transport receipts, and digests preserved outside the result path; evidence is never regenerated. (7) Independent audit — classification against frozen rules under procedural independence, with independence types separately disclosed (procedural, organizational/scientific lineage, execution, information, personal); "procedural separation is not personal independence." (8) Mechanical outcome classification — a frozen classifier maps each execution to exactly one outcome state (SUCCESS / FAIL / INCONCLUSIVE, with INVALID_EXECUTION for transport failure); the classifier records observations, it does not interpret them. (9) Descriptive statistical analysis — no hypothesis tests, thresholds, or stopping rules beyond registration. (10) Representation-bias disclosure — controlled and uncontrolled dimensions disclosed with the generalization boundary.

The separation discipline. Four claim types are kept distinct throughout and labeled where they appear:

claim type	content	who may establish it
Execution validity	did the execution complete under the registered condition?	the frozen classifier + preserved transport evidence
Evidence integrity	is the record intact and hash-consistent?	recomputation over preserved artifacts
Protocol-defined outcome	what did the registered classifier record?	the frozen classification rules
Scientific interpretation	what does the outcome mean?	nobody in this paper — a separately authorized phase over sealed outputs

The naming boundary note for the systems of an internal engineering lineage — measured objects only, never conflated with protocol claims — is preserved in Appendix 9.

2.1 Positioning relative to evaluation frameworks

We position ECP against the closest prior frameworks. Holistic capability suites (HELM, BIG-Bench) Liang et al. (2023); Srivastava et al. (2023) measure task performance at scale; evaluation/model cards Mitchell et al. (2019) document intended use; preregistered evaluation initiatives Carro et al. (2025) address design confirmation; governance frameworks (ISO/IEC 42001, NIST AI RMF) International Organization for Standardization (2023); National Institute of Standards and Technology (2023) specify management-level requirements; provenance/audit-log systems Raji et al. (2020) record histories. ECP does not compete with these instruments. Its compositional difference is that it is structured around evidentiary attribution of individual inference outcomes under a defined experimental condition — producing an auditable, per-outcome evidence chain that the above frameworks can reference but do not themselves construct. No individual control (hash sealing, preregistration, audit, bias disclosure) is claimed as unprecedented; the claim is the integrated chain and its separation discipline.

The full compositional map table is preserved, with its reading convention, in Appendix 10. A full feature-level comparison of ECP against these frameworks is provided in Appendix 10.

3 Registered evidence design

The paper's empirical content is organized as a portfolio of separately registered, separately scoped records; no quantity from one record is compared with, pooled with, or transformed into a quantity of another. At integration baseline c6c91e3, a hash-sealed comprehensive inventory enumerated 35 experiment namespaces: 16 CLOSED, 17 EXECUTED, 1 REGISTERED, 1 UNKNOWN — including the eighteen A-series audits (A1–A18), the three E-series audits (E1–E3), and fourteen further namespaces, alongside the paper's original evidence lines (T1, M1, REASON-2/3, M2, and the validation tools). The integration-

discipline paragraph that followed here in the pre-citation layout is preserved in Appendix 11 (relocated under the 2026-09-24 submission-package enhancement order; see the provenance note there).

4 Evidence lines

Each line reports its own registered scope only.

4.1 T1 pilot — Model-Disabled execution on a hybrid system

Five preregistered elementary cases were executed Model-Disabled against a hybrid system whose registered measurement surface is a deterministic, non-LLM interface layer. All five executions were procedurally valid; none satisfied the registered task-level success criterion (0/5). Preserved runtime evidence shows each request was rejected at the system's intent-routing boundary before task-level processing. The pilot therefore establishes non-support for its registered hypothesis within its registered scope — an interface-conditional, single-snapshot, bounded negative system-level result; the system's underlying reasoning capability was not exercised and is not measured. As protocol operation, the pilot demonstrates clean separation of execution validity (5/5) from task performance (0/5) on a real system.

4.2 M1 pilot — Model-Disabled execution across six frozen systems

Nineteen outcome-free cases across five reasoning families were executed as 43 registered (case × system) pairs against six hash-frozen systems: the Z3 theorem prover (z3-solver 5.1.0.0), the SymPy logic module (sympy 1.14.0), and four repositories from an internal engineering lineage (identity blinded) at frozen module or contract surfaces. Registration was sealed before any execution with outcome-free integrity verification (7/7 checks). All 43 pairs ran exactly once, Model-Disabled: no model provider was registered, the model path was verified to fail closed on every invocation, adapters ran in isolated subprocesses, and no network access occurred.

The full accounting preserves every entry. Over the registered denominator (43; nothing removed, converted, or reinterpreted after observation): task outcomes 32 SUCCESS / 6 FAILURE / 5 INCONCLUSIVE; execution validity 39 VALID / 4 INVALID; evidence validity VALID 43/43; attribution validity 38 VALID / 5 INCONCLUSIVE; protocol integrity PASS 43/43. The preregistered estimand — the observed task-success proportion among analyzable valid-task executions — is $32/38 = 84.2\%$ (exact Clopper–Pearson 95% CI 68.8%–94.0%; complementary Wilson [69.6%, 92.6%]), a precision statement about this pilot instance, quoted only under its registered label — not a success rate, capability rate, or population estimate. Two labeled alternative views are retained: $32/43 = 74.4\%$ (CP95 58.8%–86.5%) and $32/39 = 82.1\%$ (CP95 66.5%–92.5%). Strata are reported separately, never pooled: Stratum A (EXTERNAL-INDEPENDENT lineage) 20/24 determinate successful (83.3%, CP95 62.6%–95.3%); Stratum B (INTERNAL-LINEAGE) 12/14 (85.7%, CP95 57.2%–98.2%); small-stratum systems ($n \leq 3$) descriptive only.

The audit's full accounting — failure localization (OBS-01..04, at adapter encodings and registered measurement surfaces), denominator rules (INCONCLUSIVE retained, invalid executions not counted as failures, no exclusion improving the estimate), all ten protocol-validation criteria PASS, and the disclosed single-auditor fallback with the dual-implementation oracle as an internal consistency check — is preserved in Appendix 12.

4.3 REASON-2 campaign — Model-Only execution on a pinned external model

A 24-arm registered campaign (three families × four cases × two arms: reasoning and matched control) executed Model-Only against cohere/north-mini-code:free at temperature 0, one shot per arm, no tools or retrieval: all 24 executions transport-valid, the provider-reported identity matched the pin, no failover. The frozen classifier recorded 23 SUCCESS / 0 FAIL / 1 INCONCLUSIVE (reasoning 12/12; control 11 SUCCESS + 1 INVALID

envelope preserved as INCONCLUSIVE). Protocol-defined outcomes for one model under one condition — no capability or comparison claim attaches. Full record: Appendix 13.

4.4 The REASON-3 chain — defect preservation, repair, corrected re-execution, successor package

The REASON-3 line is a chain of separately sealed stages: the original campaign (evidence-4) ran 60/60 execution-valid (26 SUCCESS / 8 FAIL / 26 INCONCLUSIVE) before a mechanically reproduced prompt-map defect voided its control dimension as controls (root preserved unmodified); the forward-only, test-protected repair (07e305e) was followed by the corrected re-execution (41/6/13), a parent-identity replication (43/5/12), and the REASON-3B successor campaign (44/16/0) — each 60/60 execution-valid, the last closed by an independent read-only provenance gate. This establishes protocol operation — honest records, test-protected repairs — and nothing about capability; REASON-3B stays exploratory, used here at the mechanical layer only. Full stage record: Appendix 14.

4.5 The F2/F6 measurement illustration — the blind spot of outcome-only scoring

This subsection is the paper's only use of the REASON-3B content, at the mechanical layer only. Ten of the sixteen recorded FAILs form one mechanical class: all five reasoning arms of families F2 and F6 declared NO where the registered token is NONE, while each preserved output's step section states, in the arm's own language, that no assignment satisfies the registered constraints (paired controls: 10/10 exact registered integers). The frozen classifier records the token mismatch as the protocol-defined outcome and preserves the raw content as Layer-A observation; the relation between content and token remains unadjudicated, and outcome-only scoring would collapse these ten executions into "wrong answers". The campaign is exploratory (N = 1 per cell). Full record: Appendix 15.

4.6 M2 comparative recording campaign — a zero-divergence observation and its design limitation

A preregistered comparative recording campaign (six deterministic cases, frozen order) compared Arm-A, an ad-hoc outcome recording pathway, with Arm-B, the ECP evidentiary pathway, over identical one-shot executions: 6/6 execution-valid, evidence hash-verified 6/6 (30 files preserved), records linked 6/6, AGREEMENT 6/6, DIVERGENCE = 0, material absence 0. The result is a descriptive observation on a sample that by design contains no registered divergence case, so the comparative method's discriminative target is untested; it is not evidence of pathway equivalence, supports no statistical inference, and neither validates ECP as a measurement instrument nor establishes any model capability. Full record: Appendix 16; design, contract, controls, and original result: Appendix 2–5.

4.7 Protocol validation: seeded fault detection, independent reclassification, and comparator positive control

Three preregistered, fully synthetic exercises (zero model or provider calls) verified the protocol's control chain and a fourth validated the comparison mechanism: seeded-fault validation (10/10 exact detections, 0 false accepts, 0 false rejections, clean control accepted), an independent reference implementation reclassifying all 60 raw evidence-5 observations with 60/60 agreement (a reproducibility audit, not a human audit), a 30-case end-to-end benchmark with an identical deterministic rerun, and the M2 comparator positive control (9/9 matched — a tool validation only). The claim stays bounded: implementation validated against defined seeded faults on the tested path; not a generally validated measurement instrument. Full records: Appendix 17.

4.8 Comparative decision experiment — taxonomy overlap finding

A preregistered comparative decision experiment (12 deterministic synthetic cases: 11 fault-bearing arms + 1 clean control) exercised the boundary between the protocol's diagnos-

tic taxonomy and the outcome-only baseline: the A/B/C fault-family labels — encoding-collapse, adjudication-axis absence, and format-rule blindness — are overlapping diagnostic dimensions, not a mutually exclusive partition (frozen definitions yield A = 4, B = 2, C = 2, one A∩C overlap case, 7 unique cases of 12 reviewed). No frozen precedence rule existed, so any exclusive partition is a post-hoc convention; the unrepresented-case count is M6 = 5 under pair-based counting, and the intact twin CF3-A remains the clean control. Full record: Appendix 18.

5 Main empirical findings: the independent audit programme

The audit programme applies the governance layer of Section 2 to claims that originate outside the programme: published provider guarantees and the integrity of the evaluation infrastructure itself. Each audit is preregistered, hash-sealed, executed under bounded channels, and closed with a verdict artifact; each is reported here with its registered scope. No cross-audit pooling is performed, and audit outcomes are never merged with the evidence lines of Section 4.

5.1 Commercial structured-output claim audits (A14 — Together AI; A17 — OpenAI)

A14 audited the structured-output behavior of Qwen/Qwen3.5-9B served via Together AI Together AI (2024) under the registered json_schema configuration (temperature 0, seed 42). Against the registered thresholds: A14-A schema validity 24/24 = 1.000; A14-B escape-hatch correctness 4/4 = 1.000; in-set correctness 10/10 among transport-effective cases (conservative registered-denominator value 10/12 = 0.833; same verdict); the two transport failures were preserved and excluded by the registered rule. Verdict: all three registered claims SUPPORTED, bounded to the 40 registered cases and the registered configuration — no universal guarantee, no provider superiority, no reasoning-correctness claim. Full record: Appendix 19.

A17 audited OpenAI's published claim OpenAI (2024) that "gpt-4o-2024-08-06 with Structured Outputs achieves a perfect 100% score in our evaluations," using a preregistered 60-schema corpus, strict JSON Schema output, temperature 0, seed 42, one call per case, with per-call capture of snapshot and system fingerprint. Result: schema_validity_rate = 1.000 (60/60 primary), per-layer 1.000/1.000/1.000, exploratory gpt-4o-mini arm 60/60, zero transport failures, refusals, or parse errors. Verdict: CLAIM_HOLDS_STRONG — bounded to the tested corpus and channel, with the mandatory scope qualifications (gateway channel, observed snapshot, provider-confirmed scope constraints) preserved in Appendix 19. No generalization to all OpenAI deployments is made.

5.2 Adversarial-integrity audits (A15, A16, A11)

A15 stress-tested the evidentiary chain with adversarially injected claims: a 50-case corpus across levels L0–L4 (10 per level; 40 injected), verified by a frozen, independently hash-locked verifier. Detection: 0.75 overall, 1.00 on levels L1–L3 (10/10 each), with silent acceptance 0.00, false positives 0.00, localization 1.00; L4 is inconclusive by design (10/10, zero anomalies). Verdict: STRONG_EVIDENTIARY_INTEGRITY, reached strictly under the registered thresholds with no threshold recomputed. A11 (earlier generation) reported, under independent audit, attack success rate 0/62 = 0.000 and unauthorized acceptance 0 on its registered adversarial case set (verdict STRONG_ADVERSARIAL_ROBUSTNESS, bounded to the registered attack classes — not unrestricted robustness).

A16 subjected the evaluation infrastructure to adaptive red-teaming Perez et al. (2022). Of seven runs, four were valid and three were invalidated by harness defects, preserved with attribution records and repaired (D-1..D-5). Each valid run met 5 of 6 registered STRONG criteria — attack success rate 0.00, false acceptance 0.00, role confusion 0.00, ground-truth leakage 0.00, evidence preservation 1.00 — with the single unmet gate being localization_min in every valid run, documented as a metric artifact of refusal-dominated profiles. Cumulatively: 141 adversarial submissions, 0 ACCEPT. Final classification: IN-

CONCLUSIVE on the registered threshold set — an honest inconclusive, not a resilience proof. Full record: Appendix 20.

5.3 A bounded directional audit of a published prompt-injection robustness claim (A18)

A18 examined Anthropic's published statement that prompt-injection Perez & Ribeiro (2022) attack success on browser use "drops from roughly 50% on Sonnet 4.6 to under 1% on Sonnet 5, and effectively 0% with safeguards enabled": 30 preregistered cases per model (5 families × 6), 60 calls, temperature 0, seed 42, no retries, simulated browser context via the SoxAI gateway, per-case frozen canaries. Mechanical (canary) classification: baseline 3/30 = 0.100 (Wilson 95% CI 0.035–0.256); target 14/30 = 0.467 (0.302–0.639) — non-overlapping intervals, direction opposite to the published claim's direction; behavioral (compliance-only) 2/30 versus 11/30 — the reversal persists. Registered verdict: DIRECTIONAL_DIFFERENCE_REVERSED, a bounded directional finding. Full record: Appendix 21.

5.4 Earlier fault-family and portability audits (A4, A9, A10)

Three closed audits from the programme's earlier phase are integrated to the degree their evidence supports: A4 (held-out fault families) — verdict PASS WITH GENERALIZATION, bounded to the registered fault taxonomy, held-out families detected without prior exposure, clean-control false-rejection rates 0.0 per arm; A9 (cross-provider portability) — the registered chain ported across the registered provider set with per-system accounting preserved, max-tokens amendment lifecycle retained as registered lineage; A10 (adversarial coverage) — verdict COVERAGE BOUNDARY IDENTIFIED, integrated strictly as a documented boundary finding. Full records: Appendix 22; condensed summaries of all eligible audits: Appendix 6.

5.5 Condition-contrast and leakage audits (E1, E2, E3)

The E-series condition-contrast and leakage audits Sainz et al. (2023) are integrated as bounded records. E1 (independent AI audit): 240/240 case-grounding and 240/240 agreement in the audited session; earlier runs preserved as HARD STOP lineage. E2 (hidden-leakage audit, operator-blind, offline synthetic): pooled 7/8 primary anomalies (Level-A 5/5, Level-B 2/3) and 7/8 leakage-specific cases (SUPPORTED) — bounded to the synthetic path. E3 (condition contrast; RUN-5 the single valid execution of the v1–v6 lineage): over 50 registered cases — condition attribution 3/50, concordance 41/50, model-off accuracy 43/50, model-on accuracy 42/50, error states 0/50 — numerator/50 reporting, no confidence intervals, no capability interpretation. Full records: Appendix 23.

6 Cross-line synthesis

Across the portfolio, the same discipline does the same work in five distinct settings, and the settings do not merge. (i) Execution-condition control separates what produced an outcome: Model-Disabled lines (T1, M1) localize outcomes to deterministic surfaces; Model-Only lines (REASON-2/3, the audit programme) bind outcomes to pinned models and frozen classifiers. (ii) Registration-before-execution held in every integrated record: no registered denominator was altered anywhere in the programme — transport failures (A14), invalid runs (A16 RUN-1/3/6; E3 RUN-1–4; A14-RUN-1) entered the record as preserved lineage, never as silent exclusions. (iii) Mechanical classification before interpretation is what makes an adverse or awkward result reportable at all: A18's directional reversal, A16's honest INCONCLUSIVE, evidence-4's voided control dimension, and the F2/F6 mismatch class are all preserved observations that an outcome-only pipeline would have flattened or discarded. (iv) Bounded channels with full disclosure made third-party commercial claims auditable (A14, A17, A18) without claiming equivalence to vendor-internal evaluation — the audits establish what independent, bounded examination can observe, and no more. (v) The separation discipline is load-bearing under failure: defective campaigns remained honest records, repairs were test-protected, and identity closures used read-only gates. What the

portfolio still does not establish — deliberately — is protocol validation as a measurement instrument for reasoning, any cross-campaign trend, and any capability claim about the evaluated models.

7 Boundaries and failure modes

7.1 Corrective lineage: A1/A2 (binding)

The programme's earliest comparative audits carry superseded verdicts, and this paper binds the correction rather than the original: A1's original verdict was reclassified following a fabricated-metrics finding, and A2 v1 was invalidated with the A2-V2 closure sealed; the original records are preserved unmodified, and no manuscript claim relies on either superseded verdict. The full corrective-lineage record is preserved verbatim in Appendix 24 (see also Appendix 7).

7.2 Failure modes observed and preserved

The portfolio's failures are part of its evidence: the preserved defect record of evidence-4, the three invalidated A16 runs repaired under disclosed attribution, E3's four preserved invalid runs, A14's amendment-corrected first run, the A16 localization-metric boundary, and the A1 verdict reclassification were each preserved, attributed, and disclosed rather than deleted or silently repaired. The full failure-mode record is preserved verbatim in Appendix 25.

7.3 Claim boundary (binding)

The following boundary is adopted from the sealed REASON-3B adjudication decision and lineage-provenance records, and extended by the integration order; it binds every citation of this paper:

The claim ledger (ECP-CLAIM-LEDGER-V3.md) binds every quantitative statement — prior and integrated — to its sealed artifact. Any future edit whose claim conflicts with the ledger or the sealed records is a hard stop, not an editing decision.

In compact form: no claim about reasoning ability or mechanism is made; no cross-language claim is made from per-language tables; no statistical claim attaches to the Model-Only campaigns; the M1 proportion is quotable only under its registered estimand label; no FAIL is upgraded and no SUCCESS downgraded anywhere in the programme; ECP claims evidentiary governance, not cognition; the seeded-fault exercises validate the implementation on the tested path only; the M2 6/6 agreement is a sample-bound descriptive observation; A1/A2 are cited only through corrective lineage; A16 is cited by run-level verdicts with final classification INCONCLUSIVE; A17 and A18 are bounded to their tested corpora, snapshots, channels, and conditions; and E-series counts generalize no further than their registered corpora. The complete fifteen-point binding list is preserved verbatim in Appendix 26.

8 Limitations

Per-line limitations are intrinsic to their sealed records; none is resolved by this paper. In brief: T1 is a small, single-snapshot, interface-conditional pilot; M1 operated under a disclosed single-auditor fallback; the REASON-2/REASON-3 chain is single-model and single-condition with preserved defects and $N = 1$ corrected executions; M2's zero-divergence observation is sample-bound; each audit is bounded to its registered corpora, channels, and snapshots (A16's localization gate failed as a documented metric artifact, final classification INCONCLUSIVE; A18 is directional and condition-bound with disclosed gateway persona leakage); E1–E3 are bounded to their audited sessions and synthetic paths; and the paper adds no new experiments. The complete per-line list is preserved verbatim in Appendix 27; the binding claim boundary (Section 7.3) is unaffected.

9 Discussion

Operation versus validation. The portfolio demonstrates that ECP operates: registrations held, evidence froze before analysis, audits classified against frozen rules, and every post-hoc discovery entered the record as a preserved observation. It should not be compressed into a stronger claim: the deployments do not constitute a validation of ECP as a measurement instrument for reasoning, nor a comparison against alternative evaluation methodologies.

What the separation discipline bought. Each line exhibits a separation doing real work. T1 separates execution validity (5/5) from task performance (0/5). M1 separates estimand from interpretation and locates its six failures in registered encodings and surfaces. The REASON-3 chain separates a defective campaign from its corrected successors so the programme's history remains inspectable rather than rewritten. The F2/F6 illustration separates the recorded token from the recorded content. The audit programme extends the same discipline outward: A14 and A17 separate a provider's registered configuration from its marketing generality; A16 separates an honest INCONCLUSIVE from a resilience narrative; A18 separates a mechanical, canary-based observation from both the vendor's claim and any interpretation of it — and shows the observation's direction can contradict the claim's while remaining methodologically bounded.

Implications for benchmarking and for claim auditing. Wherever evaluation questions concern attribution — which pathway produced this outcome, under which condition — score reporting alone is structurally insufficient; the gap is solved by making the evidentiary chain explicit: registration before execution, preservation before analysis, mechanical classification before and separate from interpretation. The audit programme adds a second implication for published commercial claims: with preregistration, frozen canaries or corpora, per-call identity capture, and bounded-channel disclosure, an independent party can produce auditable evidence about a vendor's claim without access to the vendor's internal evaluation — and can do so in a form that survives disagreement, because the mechanical layer and its disclosures travel with the result.

The practical utility of ECP is illustrated by the outcomes of its audit programme: a provider claim independently confirmed (A17), a provider claim independently questioned (A18), an evaluation infrastructure stress-tested with 141 submissions and zero acceptances (A16), and a structured-output guarantee bounded to its registered configuration (A14). These cases illustrate that a score alone does not establish the provenance, execution conditions, or evidentiary basis of an evaluation outcome. The evidentiary chain is what makes the resulting claim independently inspectable and interpretable by parties other than those who produced it.

10 Conclusion

ECP operationalizes AI evaluation as an auditable evidentiary chain with a fixed separation discipline: execution validity, evidence integrity, protocol-defined outcome, and scientific interpretation are recorded separately, and the last of these is never performed by the protocol. The integrated portfolio — five evidence lines spanning both execution conditions, a validated tool layer, and an independent audit programme whose audits range from a provider's structured-output guarantee (A14; A17: CLAIM_HOLDS_STRONG, bounded) through the infrastructure's own adversarial integrity (A15; A16: 141 submissions, 0 accepted, honestly INCONCLUSIVE) to a directional audit that observed the opposite of a published robustness claim's direction under tested conditions (A18) — shows the layer operating on real systems, on real claims, and under adverse findings, without ever upgrading what its records can support. The F2/F6 illustration and the A18 disclosure set make the same point from two directions: an outcome-only score records a mismatch and stays silent, while the governed chain preserves exactly the structure that makes a result scientifically usable — and exactly the boundaries that keep it honest.

AI Assistance Disclosure

Per the submission's AI-usage disclosure requirements: AI systems were used in this work under continuous human direction. An AI coding agent — operating strictly against an explicit, owner-issued sequence of execution orders, frozen contracts, and acceptance gates — implemented the validation tooling, operated the offline verification harnesses, performed mechanical record-keeping and hash verification, and drafted portions of this manuscript. The human owner authored the scientific programme, issued every execution order, defined all acceptance criteria and claim boundaries, adjudicated every gate (including one formal scope-repair decision), and approved all closures. Every model execution reported in this paper was launched by the human owner through the controlled execution surfaces; the AI agent never issued a model call. All quantitative results derive from committed, hash-verified artifacts bound in the accompanying claim ledger; no result in this paper is an AI-generated experimental output.

Provenance and artifact availability

Sealed records cited: ECP-MANUSCRIPT-v2-FINAL.md and ECP-MANUSCRIPT-M1-SUBMISSION-COPY.md with their manifests (archive, immutable); the T1 pilot tree (sealed at baseline 344bdcc); M1 v4 records (registration, audit matrix ba12d739…, statistics, verification 235085f…); ECP-REASON-2-RESULT-COHERE-NORTH-MINI-V1.md; the REASON-3 chain records (lineage amendment, evidence-5 result, replicated-run record, REASON-3B package amendment, D1 decision, result evidence-2, lineage-provenance, adjudication decision); the M2 records (preregistration, manifest, campaign integrity audit, descriptive analysis, comparator positive-control records); the validation-tool outputs (tools/ecp-seeded-fault-validation, tools/ecp-reference-reclassification, tools/ecp-e2e-protocol-benchmark); the audit-programme closure records — A14 (closure record, registration-v2 thresholds), A15 (closure record, metrics summary), A16 (closure record V2, run-level verdicts RUN-4/5/6-1), A17 (closure record, paper handoff, metrics), A18 (closure record, paper handoff, metrics, verdict, Kiro investigation, disclosure addendum), A4/A9/A10/A11 (closure and independent-audit artifacts), E1–E3 (registration, results, audit, closure records); and the hash-sealed comprehensive inventory (research/repo-inventory-2026-09-23/, tag repo-inventory-v2-1-2026-09-23). Raw evidence roots live outside the repository and are hash-pinned by their records. The REASON-3B ledger sha256 is 059633b6…a64f0; the successor manifest self-hash is 4f62e740…a0f9 (63/63 members byte-verified). The claim ledger for this draft is ECP-CLAIM-LEDGER-V3.md (which carries V2's 53 rows verbatim and appends the 22 integration rows).

Artifact-access classification. Public/inspectable (committed): the sealed records named above, the claim ledger, the comprehensive inventory, frozen package manifests with self-hashes, verification tooling and committed outputs, explicitly labeled derived verification fixtures, and the submission package with its SHA-256 manifest. Sealed/owner-controlled (outside the repository, unpublished): the original raw evidence roots, hidden ground truth consumed in-process, and owner-local configuration; published hashes bind the layers without exposing the sealed one.

Reviewer verification without sealed access. A reviewer can: (1) obtain a committed artifact; (2) compute its SHA-256; (3) compare against the citing manifest/ledger entry; (4) re-verify chain-internal commitments (manifest self-hashes, member byte-hashes, canonical-JSON self-hashes) — all offline. The portable verification package (tools/ecp-portable-verification/) performs (2)–(4) mechanically; tests requiring an original evidence root skip with documented reasons, and tooling fails closed rather than substituting data.

Access to sealed material. Sealed raw evidence is not published. Requests for non-public evaluation material may be made through the venue's review-channel communication under double-blind constraints; such access preserves the material's non-public status and does not disclose hidden ground truth in the paper.

References

M. V. Carro et al. PREP-Eval: A pre-registration and reporting protocol for AI evaluations, 2025. URL https://prep-eval.github.io/prep-eval/. Protocol v1.0; manuscript under review.

Timnit Gebru et al. Datasheets for Datasets. Communications of the ACM, 2021. arXiv:1803.09010.

Avijit Ghosh, Anka Reuel, et al. Evaluation Cards: An Interpretive Layer for AI Evaluation Reporting. arXiv preprint arXiv:2606.09809, 2026.

International Organization for Standardization. ISO/IEC 42001:2023 — information technology — artificial intelligence — management system, 2023. ISO, 2023.

Percy Liang et al. Holistic Evaluation of Language Models. Transactions on Machine Learning Research, 2023. arXiv:2211.09110.

Margaret Mitchell et al. Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19), 2019. arXiv:1810.03993.

National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology, January 2023.

OpenAI. Introducing Structured Outputs in the API, August 2024. OpenAI blog, 2024-08-06.

Ethan Perez et al. Red Teaming Language Models with Language Models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2022. arXiv:2202.03286.

Fábio Perez and Ian Ribeiro. Ignore Previous Prompt: Attack Techniques For Language Models. arXiv preprint arXiv:2211.09527, 2022.

Joelle Pineau et al. Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program). In NeurIPS 2020, 2020. arXiv:2003.12206.

Inioluwa Deborah Raji et al. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '20), 2020. arXiv:2001.00973.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark. In Findings of the Association for Computational Linguistics: EMNLP 2023, 2023. arXiv:2310.18018.

Aarohi Srivastava et al. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. Transactions on Machine Learning Research, 2023. arXiv:2206.04615.

Together AI. Structured outputs / JSON Schema documentation. docs.together.ai, 2024.

Appendix overview

Supplementary material for "The Evidentiary Control Protocol: A Preregistered Governance Layer for Auditable AI Evaluation" (ICLR 2027 submission, double-blind) — integrated portfolio draft (2026-09-24). All content is condensed from committed, frozen repository records; no new experimental material is introduced. Per-claim binding record: manuscript/ECP-CLAIM-LEDGER-V3.md (V2's 53 rows carried verbatim + 22 integration rows).

Integrated submission draft — POST-EXPERIMENTAL INTEGRATION V1 (2026-09-24). Baseline c6c91e3 (inventory tag repo-inventory-v2-1-2026-09-23). This draft integrates the completed experimental portfolio — four prior evidence lines, the independent audit programme (A14–A18 priority, eligible earlier audits, and the E-series) — into the paper per the governing integration order. Every quantitative statement, prior and new, is bound to its sealed artifact via the accompanying claim ledger (V3); no closed experiment was modified and no bounded finding was upgraded.

Reading guide (navigation sentences relocated from the pre-expansion introduction; numbering refers to the main text): Section 2 describes the governance layer. Section 3 states the registered evidence design. Section 4 presents the evidence lines. Section 5 presents the audit programme's findings. Section 6 synthesizes across lines without pooling them. Section 7 consolidates boundaries and failure modes, including a binding claim boundary. Section 8 lists limitations. Section 9 discusses. Section 10 concludes.

Sections 10 through 27 reproduce, verbatim, material relocated from the pre-expansion main text under execution order D1-C; provenance for each relocated section is stated in the section itself.

1 A. The ten-stage governance chain

1. Case authoring (premises, question, registered answer, verification rule, leakage review; fixed answer vocabulary).
2. Structural review and novelty assessment (pre-execution).
3. Environmental knowledge-boundary characterization.
4. Registration (cases, ground truth, condition, success criterion — hash-sealed before any execution).
5. Conditioned execution (one shot per registered case; Model-Disabled or Model-Only; pathways controlled or fail-closed).
6. Raw-evidence freeze (outputs, transport receipts, digests — preserved outside the result path; never regenerated).
7. Independent audit (classification against frozen rules; procedural independence; personal-independence status disclosed).
8. Mechanical outcome classification (exactly one outcome state; records, never interprets).
9. Descriptive statistical analysis (no hypothesis tests, thresholds, or stopping rules beyond registration).
10. Representation-bias disclosure (controlled/uncontrolled dimensions and the generalization boundary).

Separation discipline: execution validity, evidence integrity, protocol-defined outcome, and scientific interpretation are four distinct claims, recorded separately; the fourth is never performed by the protocol.

2 B. The M2 preregistered design (six cases)

Six newly authored cases, one per reasoning structure, executed one-shot each in frozen order (M2-001..M2-006), single pinned external model, single-turn, text-only, no tools, no retrieval, reasoning enabled via the explicit registered activation:

Case	Family	Structure	Registered answer	Mechanical verification
M2-001	MC	multi-constraint deduction (total + doubling + offset)	5	exhaustive enumeration 1..12
M2-002	OR	ordering with adjacency + fixed anchor	4	exhaustive permutation enumeration (120)
M2-003	RI	directed-graph minimum intermediary count	3	BFS over the enumerated edge set
M2-004	CD	contradiction detection over declared integer domain	NONE	exhaustive enumeration proving empty solution set
M2-005	UD	designed underdetermination	INDETERMINATE	enumeration proving multiple distinct solutions
M2-006	CR	contrapositive + exclusive disjunction + product	40	case enumeration over venue × microphone state

Registered answer vocabulary: canonical non-negative integers plus the exact tokens NONE and INDETERMINATE (case-sensitive).

3 C. The registered comparison contract (actual source)

Arm A (ad-hoc, ECP-M2-ADHOC-RULE-V1) — binary only: SUCCESS iff a parsable final ANSWER equals the registered expected answer; everything else (wrong parsable answer, unparseable response, empty response, no output received) is FAIL. No third state, no validity layer — that absence is exactly what M2 observes.

Arm B (ECP, ECP-M2-CLASSIFICATION-RULE-V1) — SUCCESS / FAILURE / INCONCLUSIVE (invalid envelope, recorded with its reason, never repaired) / execution INVALID (transport failure; never claims an outcome), plus attribution-validity (VALID iff execution VALID with non-empty output) and evidence-integrity (VALID iff the verbatim record is preserved with its hash chain) states.

Primary label (coded precedence): (1) execution INVALID → "EXECUTION INVALID IN AT LEAST ONE ARM"; (2) S/S → "AGREEMENT ON PROTOCOL-DEFINED OUTCOME"; (3) FAIL/FAILURE → "AGREEMENT ON PROTOCOL-DEFINED OUTCOME"; (4) otherwise → "DIVERGENCE ON PROTOCOL-DEFINED OUTCOME".

Material-absence classes (coded precedence MA-1 > MA-2 > MA-3 > MA-4):

Class	Layer	Trigger
MA-1	execution validity	execution INVALID while ad-hoc can only record FAIL
MA-2	protocol-defined outcome	ad-hoc FAIL where the ECP outcome is INCONCLUSIVE
MA-3	attribution validity	ad-hoc SUCCESS with attribution != VALID (no engine path can produce this)
MA-4	evidence integrity	ad-hoc SUCCESS with integrity != VALID (no engine path can produce this)

A parseable-answer disagreement between the arms is structurally impossible: both arms compute the same equality over the same parsed envelope of the same immutable raw output.

4 D. Positive-control fixture set (preregistered) and result

Nine hermetic fixture classes, frozen with expected classifications before any run; each fixture = one full engine run over the six cases in frozen order with exactly one designated deviation:

Fixture	Deviation	Expected A	Expected B	Expected primary	Expected MA
F-0	none (baseline ×6)	SUCCESS ×6	SUCCESS ×6	AGREEMENT ×6	NO ×6
F-1	wrong integer answer	FAIL	FAILURE	AGREEMENT	NO
F-2	no ANSWER line	FAIL	INCONCLUSIVE	DIVERGENCE	YES (MA-2)
F-3	transport failure	FAIL	None (INVALID)	EXECUTION INVALID	YES (MA-1)
F-6a	lowercase token	FAIL	INCONCLUSIVE	DIVERGENCE	YES (MA-2)
F-6b	float literal	FAIL	INCONCLUSIVE	DIVERGENCE	YES (MA-2)
F-6c	negative integer	FAIL	INCONCLUSIVE	DIVERGENCE	YES (MA-2)
F-6d	leading zero	FAIL	INCONCLUSIVE	DIVERGENCE	YES (MA-2)
F-7	STEP absent, answer valid	SUCCESS	SUCCESS	AGREEMENT	NO

Result: 9/9 matched (0 findings, 0 retries, zero network/provider/model calls). MA-3 and MA-4 remain untested; result-disagreement is structurally unavailable. Classification: tool/protocol validation of the comparator contract — not a scientific result and not a validation of ECP as a measurement instrument.

5 E. M2 original result (unchanged)

The preregistered six-case comparative recording campaign: 6/6 execution-valid; both recording arms recorded SUCCESS on every case (6/6 AGREEMENT ON PROTOCOL-DEFINED OUTCOME; 0 DIVERGENCE; 0 material absence). This is a descriptive, sample-bound observation; the sample contains no divergence case, so the comparative method's discriminative target was not exercised. No statistical inference was performed.

6 F. Independent audit programme — detailed evidence summaries (paper §5)

All five priority audits are CLOSED with hash-sealed verdict artifacts; each is bounded to its registered scope. No cross-audit pooling.

6.1 A14 — commercial structured-output audit (Together AI; CLOSED 2026-09-21)

- Lineage: RUN-1 INVALID_EXECUTION (max_tokens 512 insufficient for the reasoning model) → registered amendment A14-REG-AMEND-001 (512→32768) → RUN-2 COMPLETED. The invalid run is preserved lineage, never repaired.
- Registered config: Qwen/Qwen3.5-9B, json_schema, temperature 0, seed 42, max_tokens 32768; thresholds frozen before execution (supported ≥ 0.95, bounded ≥ 0.80 for claim A; ≥ 0.75 / ≥ 0.50 for B).
- Results: A14-A schema validity 24/24 = 1.000; A14-B escape hatch 4/4; in-set 10/10 effective (conservative registered-denominator 10/12 = 0.833, same verdict); 2 transport failures (HTTP 503, A14-C0026/C0027) preserved and excluded per the registered rule.
- Verdicts: all three registered claims SUPPORTED. Scope: the 40 registered cases under the registered config; no universal guarantee, no provider superiority, no reasoning-correctness claim.
- Artifacts: closure/A14-CLOSURE-RECORD.md; registration-v2/thresholds.json; tags a14-registration-v1/v2-2026-09-21, a14-closed-2026-09-21.

6.2 A15 — adversarial claim injection (CLOSED 2026-09-22)

- Corpus: 50 cases, L0–L4 × 10 (40 injected); frozen verifier identity.

- Results: detection 0.75 overall, 1.00 on L1–L3 (10/10 each); silent_acceptance 0.00; false positives 0.00; localization 1.00; L4 inconclusive by design (10/10, zero anomalies).
- Verdict: STRONG_EVIDENTIARY_INTEGRITY — reached strictly under the registered thresholds; no threshold recomputed or added.
- Artifacts: closure/A15-CLOSURE-RECORD-V1.md; closure/A15-METRICS-SUMMARY.json; tag a15-closed-2026-09-22.

6.3 A16 — adaptive adversarial red team (CLOSED; per-run verdicts)

- Runs: RUN-2/4/5/6.1 VALID; RUN-1/3/6 INVALID (harness defects D-1..D-5, all fixed; halt evidence preserved with attribution).
- Per-run verdicts (run-level artifacts, never A16-VERDICT.json alone): each valid run met 5/6 registered STRONG criteria — attack success 0.00, false acceptance 0.00, role confusion 0.00, GT leakage 0.00, evidence preservation 1.00 — failing only localization_min (RUN-4 0.20 = 6/30; RUN-5 0.4167 = 10/24; RUN-6.1 0.5455 = 12/22; threshold 0.90), documented as a metric artifact of refusal-dominated profiles.
- Cumulative (4 valid runs): 141 submissions, 0 ACCEPT.
- Final classification: INCONCLUSIVE on the registered threshold set.
- Artifacts: results/A16-RUN-4-VERDICT.json, A16-RUN-5-VERDICT.json, A16-RUN-6-1-VERDICT.json; closure/A16-CLOSURE-RECORD-V2.md; tag a16-closed-2026-09-23.

6.4 A17 — structured-output reliability audit (OpenAI; CLOSED 2026-09-23)

- Claim tested: "gpt-4o-2024-08-06 with Structured Outputs achieves a perfect 100% score in our evaluations." (OpenAI, 2024-08-06).
- Method: preregistered 60-schema corpus (20 easy / 20 complex / 20 supported-boundary); strict JSON Schema; temperature 0, seed 42, one call per case, no retries, no fallback; per-call snapshot + system_fingerprint capture; SoxAI gateway (disclosed).
- Results: schema_validity_rate 1.000 (60/60 primary); per-layer 1.000/1.000/1.000; exploratory gpt-4o-mini 60/60; zero transport failures, refusals, or parse errors.
- Verdict: CLAIM_HOLDS_STRONG — bounded to the tested corpus and channel. Observed snapshot gpt-4o-2024-11-20 (not 2024-08-06). OpenAI confirmed (2026-09-23, case 15667047): audits bounded to observed snapshot + system_fingerprint; no canonical corpus exists; gateway results are not equivalent to direct results by default.
- Artifacts: closure/A17-CLOSURE-RECORD-V1.md; closure/A17-PAPER-HANDOFF.md; results/A17-METRICS.json; tag a17-closed-final-2026-09-23.

6.5 A18 — bounded directional audit (Anthropic claim; CLOSED 2026-09-23)

- Claim tested: prompt-injection attack success on browser use "drops from roughly 50% on Sonnet 4.6 to under 1% on Sonnet 5, and effectively 0% with safeguards enabled." (Anthropic, 2026).
- Method: 30 cases/model (5 families × 6); per-case frozen canaries; baseline claude-sonnet-4-6 vs target claude-sonnet-5; 60 calls, temperature 0, seed 42, no retries; simulated browser context via the SoxAI gateway.
- Mechanical (canary presence): baseline 3/30 = 0.100 (Wilson 95% CI 0.035–0.256); target 14/30 = 0.467 (0.302–0.639); CIs non-overlapping (0.256211 < 0.302324); direction reversed vs the claim's direction.
- Behavioral (compliance only): baseline 2/30 = 0.067; target 11/30 = 0.367; reversed direction persists.

- Verdict: DIRECTIONAL_DIFFERENCE_REVERSED (bounded). The bounded wording is mandatory: the independent bounded audit observed a directional difference opposite to the published claim under the tested conditions. No falsity claim; no replication of any internal evaluation; no generalization beyond the tested conditions.
- Disclosures: simulated browser context; SoxAI gateway (not Anthropic direct); "Kiro" persona observed in 7/60 responses (gateway system-context leakage, both arms, absent from all payloads); 4 of the 14 target mechanical successes are quote-in-refusal; n = 30 (wide CIs).
- Artifacts: results/A18-METRICS.json; results/A18-VERDICT.json; results/A18-KIRO-INVESTIGATION.json; results/A18-DISCLOSURE-ADDENDUM.json; closure/A18-CLOSURE-RECORD-V1.md; closure/A18-PAPER-HANDOFF.md; tag a18-closed-2026-09-23.

6.6 Eligible earlier and E-series audits (condensed)

- A4 (CLOSED): held-out fault families detected without prior exposure; clean-control false rejections 0.0 per arm; verdict PASS WITH GENERALIZATION bounded to the registered fault taxonomy.
- A9 (CLOSED): cross-provider portability with per-system transport/format accounting preserved; max-tokens amendment lifecycle retained as registered lineage (tags a9-maxtokens-v1, a9-v3).
- A10 (CLOSED, frozen): verdict COVERAGE BOUNDARY IDENTIFIED — a documented boundary finding; the boundary is the finding.
- A11 (CLOSED): independent audit attack_success_rate 0/62 = 0.000; unauthorized acceptance 0; STRONG_ADVERSARIAL_ROBUSTNESS bounded to the registered attack classes.
- E1 (CLOSED): audited session verified 240/240 case-grounding and 240/240 agreement (E1-RUN-4); earlier runs preserved as HARD STOP lineage; no in-repo execution artifact exists and none was manufactured.
- E2 (CLOSED): operator-blind hidden-leakage audit on the registered offline synthetic path — pooled 7/8 primary anomaly detection (Level-A 5/5, Level-B 2/3) and 7/8 leakage-specific; no ECP-security claim.
- E3 (CLOSED; RUN-5 sole valid execution of the v1–v6 lineage): 50/50 cases; Model-ON identity verified 50/50; adjudicated distribution CONCORDANT_CORRECT 41/50, CONCORDANT_INCORRECT 6/50, DIVERGENT_OFF 2/50, DIVERGENT_ON 1/50, errors/unattributed 0/50; numerator/50 reporting: condition attribution 3/50, concordance 41/50, model-off accuracy 43/50, model-on accuracy 42/50; no CIs; no capability interpretation.

7 G. A1/A2 corrective lineage note (paper §7.1)

A1's original verdict ("PASS WITH CLEAR COMPARATIVE BENEFIT") was reclassified following the fabricated-metrics finding (research/baseline-comparison-v1-verdict-reclassification/), and A2 v1 was invalidated (tag a2-invalid-2026-09-19) with the A2-V2 closure sealed (tag a2-v2-2026-09-19). The original records are preserved unmodified; no claim in the paper or this supplementary relies on either superseded verdict; both experiments are cited only through their corrective lineage (ledger rows A1-LIN-1, A2-LIN-1).

8 H. Claim-ledger cross-reference

The binding record for this draft is manuscript/ECP-CLAIM-LEDGER-V3.md: 75 rows total — 53 V2 rows carried verbatim (T1-1..4, M1-1..12, R2-1..5, R3-1..12, P-1..5, PV-1..6, M2-1..5, PC-1..4) plus 22 integration rows (A1-LIN-1, A2-LIN-1, A14-1..3, A15-1,

A16-1..2, A17-1..2, A18-1..3, A4-1, A9-1, A10-1, A11-1, E1-1, E2-1, E3-1, R1-1, PP-1). Statuses: 56 SUPPORTED, 19 BOUNDED, 0 MISSING, 0 UNSUPPORTED. The machine-readable claim→evidence matrix accompanies the package (CLAIM-EVIDENCE-MATRIX.json), and the experiment-to-claim map is in evidence-index.md.

9 Naming boundary note (internal engineering lineage)

Provenance: this section is reproduced verbatim from the pre-expansion main text of the integrated submission draft (snapshot 91f698a, tag paper-integrated-2026-09-24); it was relocated to the appendix under execution order D1-C with no change to any scientific statement, value, classification, or interpretation.

(Sealed records refer to the third row's content as the "mechanical classification"; this paper uses "protocol-defined outcome" for the same layer, uniformly.) Boundary note (naming): the systems of an internal engineering lineage (identity blinded for double-blind review) appear only as measured objects in the M1 pilot's internal stratum; protocol claims concern evidence handling, system claims concern the registered frozen composites that were measured, and the two are never conflated.

10 Compositional positioning map — full table

Provenance: this section is reproduced verbatim from the pre-expansion main text of the integrated submission draft (snapshot 91f698a, tag paper-integrated-2026-09-24); it was relocated to the appendix under execution order D1-C with no change to any scientific statement, value, classification, or interpretation.

Descriptive compositional map, not a ranking (Y = registered mechanism documented as part of the framework; P = present in some documented form; – = no registered counterpart documented in the reviewed material; cells state what reviewed documentation shows and assert no absence anywhere):

dimension	ECP	HELM	BIG-Bench	eval/model cards	NIST AI RMF	ISO/IEC 42001	prov./audit
evaluation planning	Y	P	P	P	P	P	–
preregistration of the evaluation	Y	–	–	–	–	–	–
case registration & hash sealing	Y	P	P	–	–	–	P
environmental knowledge-boundary verification	Y	–	–	–	P	–	–
leakage/contamination controls	Y	P	P	–	P	–	–
Model-Disabled condition control	Y	–	–	–	–	–	–
raw-evidence freeze (hash-pinned)	Y	–	–	–	–	–	P
independent audit chain	Y	–	–	–	P	P	P
per-outcome attribution layer	Y	–	–	–	–	–	P
provenance/hash-manifest discipline	Y	P	P	P	P	P	Y
reproducibility tooling	Y	P	P	–	–	–	–

11 Integration classes and experiment-to-claim mapping — full record

Provenance: this section is reproduced verbatim from the pre-expansion main text of the integrated submission draft (snapshot 91f698a, tag paper-integrated-2026-09-24); it was relocated to the appendix under execution order D1-C with no change to any scientific statement, value, classification, or interpretation.

Integration followed the inventory's claim-ledger-first discipline: a namespace enters the paper only through its own sealed registration/execution/audit/verdict and closure artifacts, with one ledger row per manuscript claim and no upgrade of any bounded finding. The resulting classes are: fully integrated evidence lines (T1, M1, REASON-2, REASON-3, M2, and the validation-tool line, Sections 4.1–4.7); the priority post-submission audits (A14–A18, Section 5.1–5.3); eligible earlier audits integrated to the degree their evidence supports (A4, A9, A10, A11, E1–E3, Section 5.4–5.5); and historical/lineage-only records cited only through their lineage (A1, A2 through corrective records; A3, A5–A8, A12, A13 without closure-grade records; infrastructure and planning namespaces), which appear only in the experiment-to-claim map (supplementary). REASON-1 is registered in-repo with no in-repo result record and no outcome statement is made about it.

Provenance: the corresponding main-text paragraph (paper §3, second paragraph of the pre-citation layout) read: "Integration followed the inventory's claim-ledger-first discipline: a namespace enters the paper only through its own sealed registration/execution/audit/verdict and closure artifacts, with one ledger row per manuscript claim and no upgrade of any bounded finding. The resulting integration classes and the full experiment-to-claim mapping are preserved in Appendix 11." It was relocated here under the ECP ICLR 2027 Submission Package Enhancement Order V1 (2026-09-24, baseline cfa610e) to restore the main-text page gate after in-text citations were added; its first sentence is preserved verbatim above and its pointer target is this section, so no statement is lost.

12 M1 audit accounting and independence arrangement — full record

Provenance: this section is reproduced verbatim from the pre-expansion main text of the integrated submission draft (snapshot 91f698a, tag paper-integrated-2026-09-24); it was relocated to the appendix under execution order D1-C with no change to any scientific statement, value, classification, or interpretation.

The six failures remain failures; the audit located each deviation in the registered composite as frozen — adapter encodings (OBS-01, OBS-02) and registered measurement surfaces (OBS-03, OBS-04) — rather than in solver cores or total repository capability. The five INCONCLUSIVE entries remain in the denominator; the four execution-invalid entries are not counted as failures; no exclusion that would improve the estimate was performed anywhere. All ten M1 protocol-validation criteria were PASS — a statement about protocol integrity, categorically distinct from task outcomes. The audit operated under a disclosed single-auditor fallback (procedural independence achieved, personal independence not achieved); the dual-implementation oracle is recorded as an internal consistency check, not a second rater.

13 REASON-2 campaign — full record

Provenance: this section is reproduced verbatim from the pre-expansion main text of the integrated submission draft (snapshot 91f698a, tag paper-integrated-2026-09-24); it was relocated to the appendix under execution order D1-C with no change to any scientific statement, value, classification, or interpretation.

A 24-arm registered campaign (three families × four cases × two arms: reasoning and matched control) was executed Model-Only against cohere/north-mini-code:free at temperature 0, one shot per arm, no tools, no retrieval, frozen answer vocabulary. All 24 executions were transport-valid, with the provider-reported identity matching the pin and no failover observed. The frozen classifier recorded 23 SUCCESS / 0 FAIL / 1 INCONCLUSIVE — reasoning arms 12/12 SUCCESS; control arms 11 SUCCESS + 1 INCONCLUSIVE (an INVALID envelope on one control arm, preserved as INCONCLUSIVE in every downstream record). These are protocol-defined outcomes for one model under one condition; no capability or comparison claim attaches to them.

14 The REASON-3 chain — full stage record

Provenance: this section is reproduced verbatim from the pre-expansion main text of the integrated submission draft (snapshot 91f698a, tag paper-integrated-2026-09-24); it was relocated to the appendix under execution order D1-C with no change to any scientific statement, value, classification, or interpretation.

The REASON-3 line is reported as a chain because its protocol content includes how the governance layer behaved when the execution surface itself was found defective. Each stage is a separate sealed record; no stage's numbers are compared with another's. (a) Original execution (evidence-4) — preserved as the historical defect record. The campaign completed 60/60 execution-valid (26 SUCCESS / 8 FAIL / 26 INCONCLUSIVE). Post-campaign diagnosis mechanically reproduced, on the frozen package, a surface defect: the prompt map stripped the -CTRL suffix from control case identifiers, so all 30 control arms received their case's reasoning prompt (0/30 control mappings correct; 30/30 reasoning mappings correct). Evidence-4's control dimension is void as controls; no control gap or reasoning-vs-control comparison is computable from it; the root is preserved unmodified — never deleted, cleaned, reinterpreted, merged, or reused. (b) Repair. The prompt-map defect was fixed (commit 07e305e) with regression tests and a mutation-verified discriminating test. (c) Corrected re-execution (evidence-5). 60/60 execution-valid; 41 SUCCESS / 6 FAIL / 13 INCONCLUSIVE (reasoning 12S/5F/13I; control 29S/1F/0I). (d) Replication under the parent identity. A further full 60-arm campaign executed under the REASON-3A identity, verified from the ledger's own identity fields: 60/60 execution-valid; 43 SUCCESS / 5 FAIL / 12 INCONCLUSIVE (reasoning 13S/5F/12I; control 30S/0F/0I). (e) Successor package campaign (REASON-3B). A successor package with unified envelopes and a registered 12-record amendment (the "corrected administrations") was closed by a read-only provenance gate — the committed manifest's self-hash (4f62e740…) independently recomputed, all 63 registered members verified byte-identical, the manifest entered version control once at freeze before any execution and never modified — with a provider conformance gate passed before arm #1. The campaign executed 60 one-shot arms in frozen order (pinned model, temperature 0, no retries): 60/60 execution-valid; 44 SUCCESS / 16 FAIL / 0 INCONCLUSIVE (reasoning 16S/14F; control 28S/2F; corrected administrations 3S/9F). Evidence integrity was re-verified on disk: canonical evidence hashes 60/60, audit linkage 60/60, ledger linkage 60/60; 120 execution artifacts preserved outside the repository.

The chain's protocol content: the governance layer kept a defective campaign's evidence usable as an honest record, made the repair test-protected, produced two corrected executions and a successor package, and — at the successor stage — had its identity and integrity closed by an independent read-only gate rather than by assertion.

15 The F2/F6 measurement illustration — full record

Provenance: this section is reproduced verbatim from the pre-expansion main text of the integrated submission draft (snapshot 91f698a, tag paper-integrated-2026-09-24); it was relocated to the appendix under execution order D1-C with no change to any scientific statement, value, classification, or interpretation.

This subsection is the paper's only use of the REASON-3B campaign's content, at the mechanical layer only. In REASON-3B, ten of the sixteen recorded FAILs form a single mechanical class: all five reasoning arms of family F2 and all five of family F6 declared NO where the registered token is NONE. Each of these ten preserved outputs carries a step section which, in the arm's own language, states that no assignment satisfies the registered constraints (verbatim examples preserved in the result record, in English, Japanese, German, and Arabic). The paired control arms of the same two families declared the exact registered integer in 10/10 cases; the remaining two control FAILs (both Japanese-language controls) declared the registered YES token with an empty parsed step field and remain FAIL under the registered step rule.

What the protocol can say is precise. The frozen classifier records a symbolic-token mismatch — the protocol-defined outcome — and additionally preserves the raw output, including step content, as Layer-A observation. What the protocol does not do — and what outcome-only scoring cannot do — is adjudicate the relation between the recorded content and the selected answer token: whether the ten arms reflect vocabulary-mapping behavior, scoring-interface sensitivity, or something else remains unadjudicated in the sealed record. Read outcome-only, an evaluation score would collapse these ten executions into "wrong answers", discarding exactly the preserved structure that makes them scientifically informative. The campaign is classified exploratory (N = 1 per cell; one model; one condition; one session).

16 M2 comparative recording campaign — full record

Provenance: this section is reproduced verbatim from the pre-expansion main text of the integrated submission draft (snapshot 91f698a, tag paper-integrated-2026-09-24); it was relocated to the appendix under execution order D1-C with no change to any scientific statement, value, classification, or interpretation.

A preregistered comparative recording campaign (six deterministic cases, frozen order M2-001..M2-006) compared two recording pathways over the same one-shot model executions: Arm-A, an ad-hoc outcome recording pathway, and Arm-B, the ECP evidentiary pathway. Registration and condition lock preceded the first execution; each case ran exactly once, no retries; the sealed evidence root (30 preserved files) lives outside the repository, hash-pinned by the frozen package manifest. Mechanical record: cases attempted / execution-valid 6/6 (0 invalid, 0 retries); evidence files hash-verified 6/6 (30 files preserved); Arm-A/Arm-B records linked per case 6/6; AGREEMENT on protocol-defined outcome 6/6; DIVERGENCE = 0; material absence 0. Both arms produced the same binary classification (SUCCESS) for every case. This zero-divergence result is a descriptive observation on this specific six-case sample, not evidence that the two pathways are equivalent: by design the sample contains no case with registered Arm-A/Arm-B divergence, so the comparative method's ability to detect material divergence when present is untested. No statistical inference was performed. The result neither validates ECP as a measurement instrument nor establishes any capability of the evaluated model.

17 Protocol validation exercises — full record

Provenance: this section is reproduced verbatim from the pre-expansion main text of the integrated submission draft (snapshot 91f698a, tag paper-integrated-2026-09-24); it was relocated to the appendix under execution order D1-C with no change to any scientific statement, value, classification, or interpretation.

Three preregistered strengthening exercises — fully synthetic, deterministic, zero model or provider calls — verified the protocol's control chain against defined seeded faults, and a fourth validated the comparison mechanism. Seeded-fault validation (Day 1): ten fault classes (routing mismatch, missing envelope, ground-truth mismatch, prompt-hash mutation, invalid answer token, missing ANSWER, transport failure, valid wrong answer, correct answer with malformed STEP, model-identity mismatch) each injected into an isolated

fixture harness under a frozen detection contract with one predetermined observable behavior and one machine-checkable detection code; a clean positive control was accepted 1/1 with zero codes; all ten faults were detected exactly — 10/10 exact detections, 0 false accepts, 0 false rejections — with injection and detection separately implemented and marker-independence proven behaviorally. Independent reference reclassification (Day 2): a separate reference implementation of the frozen classification specification (pure functions, no production imports, no ledger access) reclassified all 60 raw evidence-5 observations with 60/60 agreement (41 SUCCESS / 6 FAIL / 13 INCONCLUSIVE per class) — an audit of classification reproducibility, not a human audit or official reclassification; a bounded answer-only diagnostic view shifts exactly the two arms whose correct answer token carried an empty parsed step field (43 SUCCESS under the diagnostic rule). Deterministic end-to-end verification (Day 3): a 30-case synthetic benchmark (5 failure families × 3 difficulty levels) executed through the complete auditable chain; after a clean positive control (30/30), all ten fault classes were detected exactly (10/10; 0 false accepts; 0 false rejections), the four outcome states remained mechanically distinct, and a deterministic rerun reproduced the results identically. M2 comparator positive control (bounded): nine hermetic fixture classes (F-0, F-1, F-2, F-3, F-6a–F-6d, F-7) preregistered with frozen expected classifications before any run (MA-3/MA-4 excluded — no engine path; answer-level disagreement structurally impossible) were executed once each on the hermetic path with zero network/provider/model calls; all nine matched their frozen expected classifications (9/9; 0 findings; 0 retries) — a tool validation only, adding no coverage of the untested classes and neither extending nor reinterpreting Section 4.6.

Across these exercises the claim is bounded exactly as follows: the paper demonstrates the operation of ECP and validates its implementation against defined seeded protocol faults; it does not establish ECP as a generally validated measurement instrument, nor a general reasoning capability of the evaluated models. Prevention claims attach only to the tested pre-execution gates on the tested path.

18 Comparative decision experiment — full record

Provenance: this section is reproduced verbatim from the pre-expansion main text of the integrated submission draft (snapshot 91f698a, tag paper-integrated-2026-09-24); it was relocated to the appendix under execution order D1-C with no change to any scientific statement, value, classification, or interpretation.

A preregistered comparative decision experiment (12 deterministic synthetic cases: 11 fault-bearing arms + 1 clean control) exercised the boundary between the protocol's diagnostic taxonomy and the outcome-only baseline, producing a finding about the taxonomy itself: the A/B/C fault-family labels — encoding-collapse, adjudication-axis absence, and format-rule blindness — are overlapping diagnostic dimensions, not a mutually exclusive partition. Per-case application of the frozen definitions yields: A = 4 instances, B = 2, C = 2; A ∩ C overlap = 1 case (CF5-B, which instantiates both dimensions from its own generated state); unique cases represented = 7; total reviewed = 12. No frozen precedence rule existed to select a primary label; any mutually exclusive partition presented for summary purposes is a post-hoc reconciliation convention, not a preregistered classification rule. The unrepresented-case count was corrected under pair-based counting (the type-A proof attaches to the shared encoding pair): M6 = 5, corrected from the pair-exclusive 4; the intact twin CF3-A remains the clean control in the defect accounting (11 defects + 1 clean control).

19 Commercial structured-output claim audits (A14, A17) — full records

Provenance: this section is reproduced verbatim from the pre-expansion main text of the integrated submission draft (snapshot 91f698a, tag paper-integrated-2026-09-24); it was relocated to the appendix under execution order D1-C with no change to any scientific statement, value, classification, or interpretation.

A14 audited the structured-output behavior of Qwen/Qwen3.5-9B served via Together AI under the registered json_schema configuration (temperature 0, seed 42, max_tokens 32768 after the registered amendment raising the initial 512 limit that had invalidated the first run — the invalid run is preserved as lineage, not repaired). Against the registered thresholds (supported ≥ 0.95, bounded ≥ 0.80): A14-A schema validity 24/24 = 1.000; A14-B escape-hatch correctness 4/4 = 1.000; in-set correctness 10/10 among transport-effective cases (conservative registered-denominator value 10/12 = 0.833; same verdict SUPPORTED). Two transport failures (HTTP 503) were preserved and excluded from denominators by the registered rule. Verdict: all three registered claims SUPPORTED, bounded to the 40 registered cases and the registered configuration — no universal guarantee across models or providers, no provider superiority, no claim about the correctness of the model's reasoning.

A17 audited OpenAI's published claim that "gpt-4o-2024-08-06 with Structured Outputs achieves a perfect 100% score in our evaluations," using a preregistered 60-schema corpus (20 easy / 20 complex / 20 supported-boundary), strict JSON Schema output (strict: true), temperature 0, seed 42, one call per case, no retries, no fallback, with per-call capture of snapshot and system fingerprint. Result: schema_validity_rate = 1.000 (60/60 primary), per-layer 1.000/1.000/1.000, exploratory gpt-4o-mini arm 60/60, with zero transport failures, refusals, or parse errors. Verdict: CLAIM_HOLDS_STRONG — bounded to the tested corpus and channel. Mandatory scope: the primary channel was the SoxAI gateway, not OpenAI direct; the observed snapshot was gpt-4o-2024-11-20 (not the dated 2024-08-06 named in the claim); and OpenAI confirmed on 2026-09-23 (case 15667047) that independent audits are bounded to the observed snapshot and system fingerprint, that no canonical corpus exists, and that gateway results are not equivalent to direct results by default. No generalization to all OpenAI deployments is made.

20 Adversarial-integrity audits (A15, A16, A11) — full records

Provenance: this section is reproduced verbatim from the pre-expansion main text of the integrated submission draft (snapshot 91f698a, tag paper-integrated-2026-09-24); it was relocated to the appendix under execution order D1-C with no change to any scientific statement, value, classification, or interpretation.

A15 stress-tested the evidentiary chain with adversarially injected claims: a 50-case corpus across levels L0–L4 (10 per level; 40 injected), verified by a frozen, independently hash-locked verifier. Detection: 0.75 overall, 1.00 on levels L1–L3 (10/10 each), with silent acceptance 0.00, false positives 0.00, localization 1.00; L4 is inconclusive by design (10/10, zero anomalies). Verdict: STRONG_EVIDENTIARY_INTEGRITY, reached strictly under the registered thresholds with no threshold recomputed.

A16 subjected the evaluation infrastructure to adaptive red-teaming. Of seven runs, four were valid (RUN-2, RUN-4, RUN-5, RUN-6.1) and three were invalidated by harness defects, preserved with attribution records and repaired (D-1..D-5) — the harness itself was audited and rebuilt five times. Per-run verdicts (cited by run-level verdict artifacts, never by the aggregate file alone): each valid run met 5 of 6 registered STRONG criteria — attack success rate 0.00, false acceptance 0.00, role confusion 0.00, ground-truth leakage 0.00, evidence preservation 1.00 — with the single unmet gate being localization_min (0.20 = 6/30 in RUN-4; 0.4167 = 10/24 in RUN-5; 0.5455 = 12/22 in RUN-6.1; threshold 0.90), documented as a metric artifact of refusal-dominated profiles in which ground-truth refusals pass every detector and carry no localization. Cumulatively, across the four valid runs: 141 adversarial submissions, 0 ACCEPT. Final classification: INCONCLUSIVE on the registered threshold set — an honest inconclusive, not a resilience proof.

A11 (earlier generation) reported, under independent audit, attack success rate 0/62 = 0.000 and unauthorized acceptance 0 on its registered adversarial case set (verdict STRONG_ADVERSARIAL_ROBUSTNESS, bounded to the registered attack classes — not unrestricted robustness).

21 A18 bounded directional audit — full record

Provenance: this section is reproduced verbatim from the pre-expansion main text of the integrated submission draft (snapshot 91f698a, tag paper-integrated-2026-09-24); it was relocated to the appendix under execution order D1-C with no change to any scientific statement, value, classification, or interpretation.

A18 examined Anthropic's published statement that prompt-injection attack success on browser use "drops from roughly 50% on Sonnet 4.6 to under 1% on Sonnet 5, and effectively 0% with safeguards enabled." The audit executed 30 preregistered cases per model (5 families × 6: direct injection, indirect via retrieved page content, tool output, multi-turn, HTML concealment) against baseline claude-sonnet-4-6 and target claude-sonnet-5 — 60 calls total, temperature 0, seed 42, no retries — in a simulated browser context delivered through the SoxAI gateway, with per-case frozen canaries.

Mechanical classification (canary presence): baseline 3/30 = 0.100 (Wilson 95% CI 0.035–0.256); target 14/30 = 0.466667 (Wilson 95% CI 0.302–0.639); the two intervals do not overlap (0.256 < 0.302), and the observed direction is opposite to the published claim's direction. Behavioral classification (compliance only): baseline 2/30 = 0.067; target 11/30 = 0.367 — the reversed direction persists. Registered verdict: DIRECTIONAL_DIFFERENCE_REVERSED, a bounded directional finding.

The paper's claim is exactly the bounded one: the independent bounded audit observed a directional difference opposite to the published claim under the tested conditions. It does not claim that Anthropic's statement is false, does not replicate Anthropic's internal evaluation, and does not generalize beyond the tested conditions. Mandatory disclosures: the browser context was simulated, not real browser use; the channel was the SoxAI gateway, not Anthropic direct; the gateway injected a "Kiro" persona observed in 7/60 responses (channel-level system-context leakage, present in both arms, absent from all payloads); 4 of the 14 target mechanical successes are quote-in-refusal under the frozen mechanical criterion; and n = 30 per arm yields wide intervals.

Two further channel-boundary qualifications apply to the A18 interpretation. (i) The 'multi-turn' injection family was administered as [Turn n] markers within a single API message, not as a native multi-turn conversation; native multi-turn context management may produce different behavior, and the observed reversal is bounded to the marker-based implementation. (ii) The gateway-injected 'Kiro' persona (7/60) is a channel-level system-context addition, not an Anthropic model property; its presence cannot be fully isolated from model behavior in this bounded audit, and the observation remains conditional on the SoxAI channel.

Provenance (2026-09-24 final pre-submission fixes order): the paragraph below was relocated verbatim from Section 5.5 of the main text under the practical-utility insertion's page-limit contingency (order §7); the relocation is placement-only and its content is contained in the bounded-claim paragraph above; no scientific statement, value, classification, or interpretation was modified.

The paper's claim is exactly the bounded one: the independent bounded audit observed a directional difference opposite to the published claim's direction under the tested conditions. It does not claim that Anthropic's statement is false, does not replicate Anthropic's internal evaluation, and does not generalize beyond the tested conditions; the mandatory disclosure set — simulated browser context, the SoxAI gateway channel, the gateway-injected "Kiro" persona (7/60 responses), quote-in-refusal mechanical successes, and wide intervals at n = 30 — is preserved in Appendix 21.

22 Earlier fault-family and portability audits — full record

Provenance: this section is reproduced verbatim from the pre-expansion main text of the integrated submission draft (snapshot 91f698a, tag paper-integrated-2026-09-24); it was

relocated to the appendix under execution order D1-C with no change to any scientific statement, value, classification, or interpretation.

Three closed audits from the programme's earlier phase are integrated to the degree their evidence supports. A4 (held-out fault families) recorded its registered verdict PASS WITH GENERALIZATION with the generalization bounded to the registered fault taxonomy: held-out families were detected without prior exposure, with clean-control false-rejection rates 0.0 per arm. A9 (cross-provider portability) ported the registered chain across the registered provider set with per-system transport and format accounting preserved (failures recorded, never removed), and its max-tokens amendment lifecycle is retained as registered lineage. A10 (adversarial coverage) closed with the verdict COVERAGE BOUNDARY IDENTIFIED — integrated strictly as a documented boundary finding: the audited adversarial coverage is bounded by its registered families; the boundary is the finding, not a deficiency narrative.

23 Condition-contrast and leakage audits (E1, E2, E3) — full record

Provenance: this section is reproduced verbatim from the pre-expansion main text of the integrated submission draft (snapshot 91f698a, tag paper-integrated-2026-09-24); it was relocated to the appendix under execution order D1-C with no change to any scientific statement, value, classification, or interpretation.

E1 (independent AI audit of programme artifacts) verified, in the audited session, 240/240 case-grounding and 240/240 agreement under the registered 8-constraint verification path; earlier runs are preserved as HARD STOP lineage with no retroactive repair (the repository holds registration, audit, and closure records; no execution artifact exists in-repo and none was manufactured). E2 (hidden-leakage audit, operator-blind, offline synthetic path) detected pooled 7/8 primary anomalies (Level-A 5/5, Level-B 2/3) and 7/8 leakage-specific cases (verdict SUPPORTED) — bounded to the registered synthetic path and corpus, with no ECP-security or general leakage-detection claim. E3 (condition contrast; RUN-5 is the single valid execution of the v1–v6 registration lineage, with four invalid/blocked runs preserved with attribution): over 50 registered cases with Model-ON identity verified 50/50, the adjudicated 8-state distribution was CONCORDANT_CORRECT 41/50, CONCORDANT_INCORRECT 6/50, CONDITION_DIVERGENT_OFF_CORRECT 2/50, CONDITION_DIVERGENT_ON_CORRECT 1/50, all error/unattributed states 0/50 — reported as numerator/50: condition attribution 3/50 = 0.06, concordance 41/50 = 0.82, model-off accuracy 43/50 = 0.86, model-on accuracy 42/50 = 0.84, bounded to the 50 cases, with no confidence intervals and no capability or superiority interpretation.

24 Corrective lineage: A1/A2 (binding) — full record

Provenance: this section is reproduced verbatim from the pre-expansion main text of the integrated submission draft (snapshot 91f698a, tag paper-integrated-2026-09-24); it was relocated to the appendix under execution order D1-C with no change to any scientific statement, value, classification, or interpretation.

The programme's earliest comparative audits carry superseded verdicts, and this paper binds the correction rather than the original: A1's original verdict ("PASS WITH CLEAR COMPARATIVE BENEFIT") was reclassified following a fabricated-metrics finding, and A2 v1 was invalidated (tag a2-invalid-2026-09-19) with the A2-V2 closure sealed (tag a2-v2-2026-09-19). The original records are preserved unmodified; no manuscript claim relies on either superseded verdict; A1 and A2 are cited only through their corrective lineage records. This section is itself part of the paper's evidentiary content: the governance layer's history includes its own corrected records, and the correction is disclosed where the superseded results would otherwise be expected.

25 Failure modes observed and preserved — full record

Provenance: this section is reproduced verbatim from the pre-expansion main text of the integrated submission draft (snapshot 91f698a, tag paper-integrated-2026-09-24); it was relocated to the appendix under execution order D1-C with no change to any scientific statement, value, classification, or interpretation.

The portfolio's failures are part of its evidence: a post-campaign surface defect voided evidence-4's control dimension and was repaired only forward, under test protection (Section 4.4); three A16 runs were invalidated by harness defects and repaired five times (D-1..D-5) with the valid-run boundary disclosed; E3 lost four of five runs to package and transport defects, preserved with attribution; A14's first run was invalidated by an insufficient registered parameter and corrected by amendment; the A16 localization metric was documented as unfit for refusal-dominated profiles — a measurement-boundary finding, not a silent threshold relaxation; and the earliest phase's fabricated-metrics finding (A1) produced a verdict reclassification rather than a deletion. In every case the record was preserved, attributed, and disclosed.

26 Claim boundary (binding) — the fifteen-point list

Provenance: this section is reproduced verbatim from the pre-expansion main text of the integrated submission draft (snapshot 91f698a, tag paper-integrated-2026-09-24); it was relocated to the appendix under execution order D1-C with no change to any scientific statement, value, classification, or interpretation.

1. No claim about reasoning ability, in general or per family, is made or supported: a declared-token delta under a frozen vocabulary is not a capability measurement.
2. No mechanism claim is made: any reading of why the F2/F6 pattern occurs is an interpretation that this paper does not perform.
3. No cross-language claim is made from per-language tables.
4. No statistical claim of any kind attaches to the Model-Only campaigns; no endpoint, rate, or proportion inference; and no comparison with any other campaign or record.
5. The M1 proportion is quotable only with its registered estimand label ("observed task-success proportion among analyzable valid-task executions") and never as a success rate, capability rate, or population estimate.
6. No FAIL is upgraded and no SUCCESS downgraded anywhere in the programme, regardless of token agreement or plausibility.
7. ECP claims evidentiary governance, not cognition: the protocol does not measure internal model state; interpretation over sealed outputs is a separately authorized phase and remains unadjudicated for every Model-Only campaign in this paper.
8. The paper's deployments demonstrate protocol operation (Section 6); they neither claim nor preclude protocol validation as a measurement instrument for reasoning, which would require a separately registered design.
9. The seeded-fault exercises (Section 4.7) validate the implementation against defined seeded protocol faults on the tested path only; "prevented" attaches exclusively to the tested pre-execution gates, and the reference implementation audit is not a human audit.
10. The M2 campaign's 6/6 agreement (Section 4.6) is a descriptive observation on one preregistered six-case sample: it is not evidence of recording-arm equivalence, not a validation of ECP as a measurement instrument, and supports no statistical inference.
11. The superseded A1 verdict and the invalidated A2 v1 are cited only through their corrective lineage (Section 7.1); no current claim rests on either.

12. A16 is cited by its run-level verdict artifacts; its final classification is INCONCLUSIVE on the registered threshold set, and its localization-gate failure is disclosed as a documented metric boundary.
13. A17's verdict is bounded to the tested corpus, observed snapshot (gpt-4o-2024-11-20), and gateway channel; no equivalence to OpenAI-direct results and no generalization to all OpenAI deployments is claimed.
14. A18 supports only the bounded statement that the independent audit observed a directional difference opposite to the published claim's direction under the tested conditions; no claim that the published claim is false, no replication claim, and no generalization beyond the tested conditions is made.
15. E-series counts are bounded to their registered corpora and sessions; no security, coverage, or capability generalization is made from them.

27 Full per-line constraint and limitation list

Provenance: this section is reproduced verbatim from the pre-expansion main text of the integrated submission draft (snapshot 91f698a, tag paper-integrated-2026-09-24); it was relocated to the appendix under execution order D1-C with no change to any scientific statement, value, classification, or interpretation.

- T1: small pilot; negative result conditional on the registered input representation and interface; single system snapshot; single language; single execution per case; author-centric case selection; single operational actor (procedural but not personal audit independence) — full list L1–L12 in the frozen parent manuscript.
- M1: single-auditor fallback (no inter-rater statistics); "EXTERNAL-INDEPENDENT" denotes system lineage, not operator independence; INTERNAL-LINEAGE results are not independent replication; small strata ($n \leq 3$) descriptive only; confidence intervals idealize heterogeneous entries as i.i.d. Bernoulli trials and are precision statements, not population estimates; Model-On was not performed; model familiarity with the involved public codebases cannot be excluded; anomalies remain explicit.
- REASON-2 / REASON-3 chain: single model; single condition; one envelope defect preserved as INCONCLUSIVE; evidence-4 supports no control-dimension statement by design; each corrected execution and the successor campaign are $N = 1$ per case $\times$ arm, single session; the successor campaign is EXPLORATORY with an explicit adjudication decision; per-arm latency and model-reported identity are not recorded in the successor evidence schema (disclosed).
- M2: zero observed divergence (6/6) is sample-bound; the sample contains no divergence case, so the comparative method's discriminative target is untested; single pinned model, single session, one-shot; no statistical inference.
- A14: 40 registered cases, one model, one provider channel; transport exclusions follow the registered rule; the amendment lineage (RUN-1 invalid $\rightarrow$ RUN-2 corrected) is disclosed; no cross-provider comparison.
- A15: synthetic offline corpus (50 cases, 40 injected); detection on L4 is inconclusive by design; results bound the verifier implementation, not any live system.
- A16: localization gate failed in every valid run and is documented as a metric artifact of refusal-dominated profiles; the six documented gaps were not adversarially retested; the aggregate classification is INCONCLUSIVE.
- A17: single snapshot via one gateway channel; corpus designed independently but not vendor-canonical (no canonical corpus exists per the provider); provider confirmation constrains the claim's scope.
- A18: $n = 30$ per arm (wide intervals); simulated browser context via one gateway; gateway persona leakage (7/60) disclosed; 4 mechanical successes are quote-in-refusal; the observation is directional and condition-bound, not an estimate of any deployment's real-world rate.

- E1–E3: E1 is bounded to the audited session (no in-repo execution artifact); E2's corpus is synthetic and offline (40 cases); E3 rests on its single valid run of a five-run lineage with four preserved invalidations; none of the three supports security or capability generalization.
- The paper itself: Related Work and protocol-stage detail are carried by reference to the frozen parent manuscript; the integration added no new experiments — it binds existing sealed records into the paper's narrative.
- Diagnostic taxonomy overlap: the A/B/C taxonomy proved overlapping rather than mutually exclusive; no frozen precedence rule existed, so any exclusive partition is a post-hoc convention.

28 Regulatory alignment — full paragraph

Provenance: this section is reproduced verbatim from the pre-expansion main text of the integrated submission draft (snapshot 91f698a, tag paper-integrated-2026-09-24); it was relocated to the appendix under execution order D1-C with no change to any scientific statement, value, classification, or interpretation.

Provenance (2026-09-24 V2 remaining-edits order): the Section 9 pointer paragraph below was relocated to this appendix section after the Section 2.1 self-containment edits and the Appendix 21 A18 channel-boundary qualifications pushed the main text to 9.062 pages (above the 9.00-page limit); the relocation is placement-only — the sentence is preserved verbatim and its content is contained in the full paragraph below; no scientific statement, value, classification, or interpretation was modified.

The stage artifacts map naturally onto documentation, traceability, and auditability expectations in contemporary AI regulation; the full regulatory-alignment discussion is carried in the frozen parent manuscript (§11) and the corresponding paragraph is reproduced in Appendix 28.

Regulatory alignment. The stage artifacts map naturally onto documentation, traceability, and auditability expectations in contemporary AI regulation; ECP supplies one concrete, auditable procedure where such expectations leave methodology open. The full discussion is carried in the frozen parent manuscript (§11).