ClaimWork: A Reproducible Evaluation Platform for Procedural AI Systems

CO-97 Denial Management as the First Benchmark

Saloni Garg

November 2026

Status paper, version 1.0, 2026-11-03. Authoring method, stated up front because the paper's own standard requires it: this venture and its benchmark were researched, designed, and built with a large language model as the domain expert of record, under a written protocol (quote verified rules, mechanical citation checks, gold computed by code). No human coding expert has audited the benchmark yet; a costed audit step exists and is triggered before any customer delivery. Realism is unmeasured. Agreement figures in this paper are cross model agreement, never expert agreement. Every number below is either read from a pinned primary source, produced by the project's own code, or labelled as an estimate.

Abstract

Frontier labs and applied AI vendors are spending heavily on reinforcement learning environments and on evidence that procedural AI systems work. The sellable unit in that market is not data; it is a verified task environment: a task distribution, a sandbox, and a grader that cannot be gamed. This paper describes ClaimWork, a reproducible evaluation platform for procedural AI systems, and its first benchmark: reworking claim denials under a single, narrow denial code, CO-97 (bundling), in US medical billing. The benchmark's ground truth derives from a published, versioned federal table (the CMS National Correct Coding Initiative edits), which makes gold computable by a 34 rule decision table rather than authored by judgement. The paper reports the market evidence for the opportunity, the domain selection process, the benchmark design, a completed build against the real 2026 Q4 CMS files (39 tasks across 12 archetypes), a verifier audit in which gold scores 1.0 with zero false passes across a degenerate solution battery, and a first calibration run (234 attempts by three frontier models, total cost $6.62) whose difficulty profile lands most archetypes in or near the target band, exposes a stale knowledge failure in all three model families, and identifies three presentation fixes required before the difficulty of three archetypes can be judged. It closes with the business case, calibrated outcome odds, the constraints on the venture, and the one remaining step between the current state and a sell decision.

1. The opportunity

The economics of RL training have shifted value from trajectories to the apparatus around them. Synthetic trajectories are cheap to produce (roughly $0.28 each in the coding domain; hundreds offered for $100), and

raw data without verification is close to worthless, because a training signal is only as good as the grader that scores it. What buyers pay for is the gym: a task, a sandbox, and an unhackable verifier.

Observed pricing in this market (researched 2026-11-01): task plus grader at $200 to $2,000 for standard domains and $5,000 or more for coding; made to order environments at a median of about $35,000; high fidelity application replicas far above that; exclusivity at four to five times base price. Company reported run rates confirm the scale: Prime Intellect above $100M annualised, Mercor reported between $500M and $2B, Handshake approaching $1B, and Anthropic reported to have discussed spending over $1 billion on RL environments in a year. The founder's own prior work sits in this market: an RL environment for GPU training failure diagnosis with frozen splits and a hosted evaluation, and a published analysis showing that widely used verifiers passed 16.83 percent of no op attempts (1,683 of 10,000), which is the failure mode this venture's product is designed to exclude and to prove absent.

A second buyer class matters as much as the labs. Vendors selling AI agents for revenue cycle management (RCM) increasingly face health system procurement processes that require independent validation evidence for coding assistance. The vendor is the paying conduit; the demand originates with its customers. The pitch writes itself: your customers' procurement will demand independent evidence; we sell it to you before they ask.

2. Domain selection

Domains were ranked on competition, ground truth quality, and whether a solo or small entrant can reach a defensible product.

- Accounting and FP&A: high and closing competition (multiple vendors hiring accounting SMEs for exactly the tasks first proposed).
- Generic enterprise workflows: highest competition; the leading entrant spent two years recreating enterprise applications and was acquired.
- Legal: high on evaluations, medium on environments.
- Mortgage underwriting judgement and freight dispatch: rejected on ground truth. Underwriting gold is contested and carries documented bias risk; freight has no ground truth at all, only an invented cost model.
- A vendor census published in July 2026 found zero RL environments for medical billing, payroll, insurance claims adjusting, freight, or hotel property management. Follow up checks found the gaps narrowing unevenly: insurance claims gained a live competitor (AutoStack, with published pricing used as an anchor below), mortgage gained benchmark incumbents moving toward environments, payroll gained a benchmark owner. Medical billing denial management stayed empty of environment vendors; the RCM agent companies in the space are buyers of evaluation, not sellers of it.

Within medical billing, the pilot slice is one denial code, CO-97 (bundling), for one reason: it is the only common denial whose correct resolution is a published, versioned table. CO-16 is more frequent (about 20 percent of denials) but it is a catch all whose fix is whatever the paired remark code names, which makes it many tasks under one code and, by the counting method of section 5, roughly 200 rules of ground truth. CO-97 coding errors are about 15 percent of denials, and five categories together are about 75 percent. Start where gold is a table; expand only where the table stays small.

3. The product

ClaimWork sells an audit and evidence package: the environment, the task set, the verifier, a written audit report of that verifier's own failure rates, and full version metadata. It deliberately does not sell "certification", a word that implies an accredited certifier and, in the products that legitimately use it, bundles a standard, retesting, and insurance at five figures per year.

Two buyer classes: RCM agent vendors buying independent evaluation (pricing anchors from the one live adjacent competitor: $5,000 for a feasibility week, $20,000 for an evaluation including an oversight report), and frontier labs or data foundries buying pilots (proposed at $25,000 to $50,000) and production tasks. A quarterly refresh subscription is a pricing hypothesis ($10,000 to $25,000 per quarter), untested, and is the line on which the business model's renewal condition depends.

Every release is pinned to a six axis version manifest, because correct coding is a moving target: CPT (annual, Jan 1), ICD-10-CM (annual, Oct 1 fiscal), HCPCS (quarterly), NCCI edits (quarterly), CARC and RARC code sets (three times a year), and the physician fee schedule (annual), plus dated payer bulletins. The current build is pinned as ClaimWork 2026.Q4.

4. Benchmark design

The world model stores encounters as structured facts first (site, session times, practitioner, units, overlap), with any note text rendered from the facts, so documentation sufficiency is a deterministic check. There is no LLM judge anywhere in the scoring path. An episode runs from the denied state through the agent's intervention and deterministic readjudication to a final ledger state: payment posted or correct contractual write off, never merely resubmission.

Three synthetic payers exercise payer variation: M (Medicare model, NCCI exact, accepts modifier 59 or the specific X modifier), A (commercial, NCCI aligned, requires the specific X modifier), and B (stricter, with a small number of bulletin bundled pairs and shorter filing limits: M 365 days, A 180, B 90).

Twelve archetypes cover the space of the denial: A1 indicator 0 hard bundle; A2 distinct site (XS); A3 separate encounter (XE), including an overlapping times flip variant; A4 separate practitioner (XP); A5 non overlapping service (XU), including a thin documentation flip; A6 payer specific modifier policy; A7

documentation that does not support the modifier, the compliance trap; A8 an edit deleted for the date of service; A9 an edit whose indicator changed between quarters, the wrong quarter trap; A10 medically unlikely edit exceedance; A11 a combined PTP and MUE claim; A12 the workflow case where a correction is already in process and the correct action is to hold.

Scoring is per element and deterministic: corrected claim 0.45, ledger 0.30, close reason 0.15, timeliness 0.10, with compliance penalties (unsupported modifier minus 0.50, patient billed for a contractual bundle minus 0.50, duplicate original submission minus 0.25, upcoding minus 1.00 and automatic failure). Degenerate strategies (no op, unchanged resubmission, copied prior claim) score zero by construction.

The design's audit gates: verifier false pass rate below 1 percent and exactly zero on compliance traps; gold false negative rate zero; expert agreement at kappa 0.80 or above on a blind sample (deferred with the human audit; cross model agreement is reported in the interim and labelled as such); archetype retention where the best model scores 30 to 65 percent; contamination canaries in every release; the test split never published. The threat model covers reward hacking, contamination, template learning (countered by minimal pair variants whose gold flips), stale knowledge (A8 and A9 exist to measure it), gold capture, and the shared rule code blind spot: the readjudication engine and the verifier share the decision table, so the table itself is validated against the primary sources by quote verification, and the verifier recomputes expected outcomes independently rather than trusting the environment.

5. The decision table

The central hypothesis of the venture is that CO-97's ground truth is small. It was tested by writing the table and counting. Gold derivation for any CO-97 episode is 34 rules: five separable sub decisions (applicability 5, bypass 10, units and MUE 6, workflow 5, timeliness and close 5), plus 3 interactions, plus 3 validation points. The naive cross product collapses because applicability resolves first and renders most inputs irrelevant. Every rule carries a verbatim quote from the 2026 NCCI Policy Manual Chapter 1, string matched against the source by an automated checker (7 of 7 quotes verified; the checker also rejects fabricated code pairs and verifies MUE values against the pinned table).

Of the three validation points the source text closed two: V2, MAI 1 is a claim line edit where modifiers may report the code on separate lines, each line adjudicated separately; V3, MAI 3 bypass requires evidence the units were actually provided, were correctly coded, and were medically necessary (the reconsideration checklist), and MAI 2 is an absolute date of service edit whose override would be contrary to CMS policy. V1 remains provisional: when two distinctness circumstances hold at once, the verifier accepts any matching X modifier, because the manual supports "most descriptive modifier" for the 59 versus X choice but does not order the X modifiers among themselves.

The table also yields the expansion criterion used for every future domain: count the rules before entering; the bounded band is roughly 20 to 60. CO-97 is the only measured row (34). Payment posting is

hypothesised small and stays second in the expansion order; CO-16 is hypothesised at 200 plus and stays behind it. The table must not be quoted outside the simulated world: XU is a rule only because the fact schema carries an explicit overlap field, and that kind of dependence is recorded wherever it occurs.

6. Build report

The build ran against the real files, downloaded from CMS: the NCCI PTP quarterly change files effective 2026-10-01 (1,119 additions: 667 indicator 1 and 452 indicator 0; 14 deletions; 7 modifier indicator changes), the full MUE Practitioner Services table (15,212 codes), and the 2026 NCCI Policy Manual Chapter 1. The decision table is executable code; the task generator instantiates the archetypes from the pinned files with gold computed by the table for every task; the environment implements the canonical state machine with every agent action logged; the verifier scores per element with an independently recomputed expected outcome.

Two facts from the real data corrected the design. First, MAI 1 (the line edit) exists for only 41 codes; MAI 2 covers 5,493 and MAI 3 covers 9,678, so unit exceedance archetypes are built mostly on MAI 2 and MAI 3, not MAI 1 as first assumed. Second, the full PTP Practitioner table sits behind the AMA licence click through on the CMS site, so this development build's pair universe is the quarterly change files: a biased sample of new and changed pairs only. The full table is a precondition of any customer release and requires one licence acceptance by the founder or her explicit delegation.

The development set is 39 tasks: A1 to A9, A11, and A12 at 3 each, A10 at 6 (both unit variants), generated with a fixed seed, each carrying a contamination canary.

7. Verifier audit

The verifier was audited before it was trusted, in the manner the product promises to audit for customers. Gold false negative rate: zero. All 39 gold solutions score exactly 1.0. False passes: zero across the degenerate battery (no op mean 0.077, whose single 1.0 is the A12 hold archetype where inaction is correct by design; unchanged resubmission mean 0.077; modifier spray mean minus 0.081, negative as specified, because the verifier recomputes the expected outcome itself, so a sprayed modifier that the environment happens to pay earns no ledger credit; duplicate original mean minus 0.173; patient billing on bundle tasks 0.20, which is gold's 1.0 minus the ledger element and the 0.50 penalty; wrong modifier mean 0.28, never passing).

The audit also caught three defects in the project's own machinery before those numbers were produced: a missing verifier branch for deleted edits, bundle close semantics leaking into MUE only tasks, and a battery policy polluting its own statistics. All three were fixed at the source and the audit re run. They are recorded in the audit report because an audit report that lists only successes is marketing.

8. Calibration results

The first calibration ran on 2026-11-01: 39 tasks, three Anthropic models (Sonnet 5.5, Opus 5.5, Fable 5.1), two replicates each, 234 scored attempts, total cost $6.62 against a $50 hard cap. Protocol caveat: all three models are from the family that drafted the benchmark, so these are diagnostic numbers, not headline calibration; headline numbers require a second provider.

Full pass rates: Opus 5.5, 53 percent (mean score 0.621); Sonnet 5.5, 41 percent (mean 0.574); Fable 5.1, 29 percent (mean 0.382), a figure that is not comparable: its attempts failed provider side, through intermittent safety layer refusals of the billing work items (observed directly, with empty content under a "cyber" refusal category) and probable truncation of long thinking traces against the output token cap, while its disposition accuracy on attempts that did score, 52 percent, was in line with the other models.

By archetype, pooled pass rates: A12 hold, 94 percent, and A7 the compliance trap, 89 percent, are saturated; the A7 trap did not bite at all (zero unsupported modifier penalties in 18 attempts; models refused modifiers the facts did not support). A10 (64 percent), A9 (61 percent), A4 (50 percent), A1 (44 percent), and A2 (33 percent) sit in or near the 30 to 65 percent retention band, with A3 (28 percent) just below. A11, the combined claim, is hard at 6 percent, legitimately. Three archetypes recorded zero passes, and each teaches something different. A6, the payer bulletin pair, has a negative pooled mean (minus 0.306): every unsupported modifier penalty in the run, 11 in total, landed here, because models applied the standard NCCI answer and ignored payer specific bundling. The trap works. A8 exposed the stale knowledge failure in all three families: models uniformly removed the bundled line and closed the episode as paid as bundled for an edit that the pinned file had deleted for that date of service. A5 defeated every model on presentation: the XU circumstance arrives as a structured fact (overlap: false), and no model connected it to the modifier.

Three fairness caveats attach to the zero pass archetypes, and they are the v0.2 fix list. The work item does not include the pair's current edit indicator, the lookup a real biller performs, so v0.1 conflates memorised edit table knowledge with rule application; the v0.2 work item carries the pinned edit row. Payer B's bulletin pair is not identified in the work item, though a real biller has the bulletin; v0.2 includes its text. Encounter facts are terse codes rather than prose; A5 should be retested under prose rendering. Until those fixes land, the difficulty of A5, A6, and A8 cannot be cleanly judged.

Cross model agreement, reported as a consistency check and nothing more: across the 43 replicates where all three models produced scored plans, the panel agreed unanimously on the core disposition in 51 percent of cases, was unanimous and matched gold in 35 percent, and its majority matched gold in 51 percent. Element analysis shows models losing the most credit on the claim itself (63 percent of available credit earned), then on posting exactly what readjudication pays (60 percent), with close reason and timeliness at 76 percent each.

9. Business case

Estimates, labelled as such, calibrated against the anchors above. Under the adopted lean protocol the spend before any revenue decision is the $500 envelope, of which $6.62 is spent. (The pre lean estimate of $40,000 to $75,000 and 6 to 9 months to first cash was superseded when the domain expert was replaced by the protocol of section 5.) Conditional on a genuinely good audit report, the estimated likelihood of at least one RCM vendor paying for an evaluation within six months of selling is 40 to 55 percent; of at least one lab or foundry pilot within twelve months, 30 to 45 percent; of either in year one, roughly 60 to 70 percent. Viability, defined as three or more paying customers, one renewal, and $250,000 to $500,000 annualised, is estimated at 15 to 25 percent within 24 months; a profitable small business outcome ($100,000 to $250,000 per year) at roughly 30 to 40 percent; the zero outcome at 40 to 50 percent, most likely as "product built, buyers polite, procurement never closes".

The customer process: a free sample with its audit report; a technical review call, which is the real sales event; a paid pilot with success criteria; procurement (security review, data processing agreement); then production plus quarterly refresh. Channels: lab partnership teams, data foundries, vendor directories for environment suppliers, enterprise environment networks, RCM vendors as evaluation customers, and design partners through industry networks. The five viability conditions, in kill order: the verifier survives external red teaming; calibration lands in band; table validation quality holds up (this is the condition the deferred human audit tests); one reference customer lands; the quarterly refresh actually renews.

10. Constraints, risks, and deferred items

The largest scientific risk is not construction but acceptance: whether buyers accept 34 rules as sufficient for the slice of CO-97 the benchmark claims to model, explicitly not full production billing. The validity domain contract (what the benchmark measures, and what it does not) is therefore part of the product, not a disclaimer.

Named limitations: scores on this benchmark upper bound real world performance, because encounter facts are canonical, payer policies are synthetic models, medical necessity and free text interpretation are excluded by design, the development pair universe is the change files rather than the full edit table, and the calibration is single provider and single shot (one action plan per attempt, not an interactive agent loop). Realism of variants is unmeasured and will not be quoted until a human rates a sample.

Deferred, with triggers and prices, not deleted: the human CPC audit of the table and a task sample ($600 to $1,200; 37 items; the table freezes at five or fewer gold changing corrections with no missing rule, fails above twelve or on a missing rule requiring a new sub table), triggered by a buyer asking how the table was validated or by a signed paid feasibility; the full PTP table download, one licence click; the transfer proof (a small model trained on train variants, evaluated on a frozen test split) on the existing $500 Verda compute credit; headline calibration with a second provider; and company formation and immigration structuring

($3,000 to $5,000 of legal work), required only before paid delivery, because the founder is on H-1B and the line between passive ownership and managing a business is the line between this venture being an asset and it being a liability. No company has been formed.

11. Where the venture stands

The sprint's own success checklist, scored as written, with the LLM only amendment applied: decision table implemented with validation points closed or recorded provisional (yes; V1 provisional); 12 archetypes generated from real data (yes, 39 tasks); citation validation at 100 percent for pair, indicator, and date (yes, mechanically checked); gold at full score in its own verifier (yes, 1.0, zero false negatives); traps and degenerate solutions at their stated expectations (yes, zero false passes); one polished audit report (yes, written, with the authoring method disclosed on its face). One item is partial: cross model re derivation was run as a disposition panel and its agreement measured (section 8), but the founder's adjudication of the disagreement list, working from the quotes, has not happened. That session, estimated at 60 to 90 minutes, is the last step in the protocol before the table can be called frozen, and it is deliberately the founder's, because adjudicating gold is the one judgement the protocol does not delegate.

The pre committed rule was: if the checklist passes, stop building and start selling. It stands at six of seven with the seventh scheduled rather than failed. The recommendation of this paper is therefore sequenced, not open ended: founder adjudication session; v0.2 presentation fixes and a retest of A5, A6, and A8 only (a few dollars at measured prices); then the sell decision on the strength of a frozen table, an audit report, and a calibration profile, with the free sample package as the opening move. Total additional spend to that decision point: under $25 at measured prices, inside the existing envelope, with the CPC audit still correctly deferred until a buyer's question or a signed feasibility triggers it.

Appendix A. Version manifest (as built)

NCCI PTP change files 2026 Q4, effective 2026-10-01 (full table pending the AMA licence click through). NCCI MUE Practitioner Services 2026 Q4, 15,212 codes. NCCI Policy Manual 2026, Chapter 1. CPT, ICD-10-CM, HCPCS, CARC and RARC axes not yet pinned in the development build. Fee schedule amounts are synthetic, deterministic per code, and are not Medicare amounts.

Appendix B. File inventory (workspace/claimwork)

BENCHMARK_SPEC_v0.1.md; DECISION_TABLE_v0.1.md; GOLD_AUTHORING_GUIDE_v0.1.md; LEAN_PLAN_v0.1.md; AUDITOR_VALIDATION_PACK.md; AUDIT_REPORT_v0.1.md; CALIBRATION_v0.1.md; this paper. Code in src/ (parser, decision table, checker, environment, generator, gold executor, verifier, audit battery, calibration harness and runner). Data and build outputs in data/ and

build/, including the pinned source files, the 39 task set, the stage 4 audit results, and all 234 calibration records.

Appendix C. Spend ledger

Sprint spend to date: $6.62 (calibration run, plus one diagnostic call), of the $500 stage gated envelope. Verda compute credit: $500, untouched, earmarked for the transfer proof. Deferred outside the envelope: CPC audit $600 to $1,200 on trigger; legal and immigration structuring $3,000 to $5,000 before paid delivery only.