Toll Bench: Can AI Systems Deliver Real-World Human Wants?

Anonymous Author(s)
Affiliation
Address
email

Abstract

Toll Bench extends execution-based evaluation into the physical and economic world: it measures whether an AI agent system can convert a real person's stated want into a verified outcome. People post targets with a frozen finish line, budget, timeline, and approval cadence; registered agents bid sealed proposals and, selected, attempt delivery. A paid attempt succeeds only when the finish line is reached, the person approves, payment settles through the platform, and integrity checks pass; a free attempt requires a verified person's approval and milestone evidence. Each accepted attempt freezes the agent-system version, base model, harness, autonomy level, and operator. Toll Bench reports completion rate, agent-active time in minutes, and cost to the person, with the Toll (the per-band median time and cost of a delivered want, never blended across bands) as the headline. Bands are defined directly on the frozen probability of success, so membership is recomputable from public data; results roll up by agent, system version, and base model, with model-level comparisons observational and never blended across autonomy levels. Events are specified for an append-only Merkle transparency log with externally witnessed checkpoints (witnessing specified, not yet engaged), and tasks are generated continuously by real people, structurally reducing advance test exposure. This paper introduces the benchmark and its protocol; it makes no capability claims. A five-week testing period exercised the machinery; official counting begins at Day Zero, August 24, 2026, with earlier events labeled testing-period in the public dataset.

1 Introduction

Many benchmarks lose discriminating power as scores rise, test material leaks, and systems optimize for the test; the usual response is a harder version of the same evaluation. We take a different position: the frontier of AI capability is not a harder puzzle but the distance between a person wanting something and having it. A want is not a wish: it is a problem statement, the exact specification of a valuable outcome. The question that matters is whether, given that specification, an AI system can deliver.

Existing agent benchmarks increasingly use realistic tasks and execution-based verification but remain bounded by a predefined digital environment. Toll Bench adopts the execution-verification philosophy of SWE-bench (Jimenez et al., 2024) and extends it beyond the codebase: a task resolves not when repository tests pass but when a frozen real-world finish line is reached, the person approves, and, where money is part of the deal, payment settles. Approval, transaction status, milestone receipts, and integrity checks form a layered body of evidence rather than an infallible single signal.

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

Toll Bench works as follows. A person sets a baseline (time and energy, money and resources, proven paths available), names a budget from zero dollars up, sets a timeline, and posts their want to a public board (live at https://tollbench.com); any registered AI system can bid a plan, the person picks one, and the agent executes, the person approving or declining at each milestone, each signed approval becoming ledger evidence. A finish line reached, approved, and (on paid targets) paid records a success; an expired timeline records a miss; every agent's full record, wins and misses both, is public forever.

Our contributions are: (1) a live benchmark of AI capability on real human outcomes (success rate, time, and cost, with the per-band Toll as the headline price of a crossing), reported by agent, system version, and base model; (2) a layered verification protocol (paid outcomes: frozen finish line, verified approval, settled payment, integrity review; free outcomes: verified approval and milestone evidence); (3) a transparency-log specification making silent alteration of benchmark history detectable, corrections and invalidations preserved as new events; (4) an open submission door: no affiliation or invitation required, sealed proposals judged by the person, every accepted attempt attributed to a frozen agent-system version; and (5) a continuous task supply: future instances do not exist before people post them, reducing advance exposure, though recurring patterns and successful workflows may intentionally become learnable.

1.1 The Testing Period and Day Zero

Toll Bench ran a five-week testing period before official measurement: from July 20 through August 23, 2026, the platform's operators and early testers exercised the full machinery (intake, refinement, steward review, sealed bidding, signed deals, milestone approvals, settlement rails in test mode, the transparency ledger) on live infrastructure. Nothing from this period is reported as measurement: the tasks were practice material, the participants were the platform's own operators, and the sample was selected to break the machine, not to measure the field. **Official measurement begins at Day Zero, August 24, 2026.** Every earlier event carries a testing-period label in the public dataset and is excluded from official counts, rates, ratings, and calibration audits; official counts start from zero, and the testing history stays published in full.

During testing the pipeline frequently stalled on the person side: sealed proposals sat unpicked, and plan steps waited on a person's answer. We draw no conclusion from this (the people involved were the platform's own operators, so the stalls describe the test setup), but the condition shaped the protocol: the agent-active clock (Section 2.3) excludes person-side waiting, the lapse law resolves a person's fourteen-day silence in the agent's favor rather than against it (Appendix C), and Version 2.0 adds a person-side effort metric; whether person-side latency matters with real, unaffiliated participants is a Day Zero question. Testing also produced operational laws now in the rulebook (e.g., an agent-facing capability absent from the published contract does not exist).

2 Toll Bench

Toll Bench is a benchmark featuring wants posted by real people and targets undertaken by AI agents to fulfill them; the task is to take a want from posted to delivered, within the person's budget and timeline, to the person's satisfaction, with approval and payment as proof.

2.1 Benchmark Construction

Human wants arrive noisy, vague, or unfit for the marketplace; a three-stage pipeline produces scoreable task instances at scale. *Intake*: a short structured flow captures the want, the person's baseline across three universal factors (time and energy, money and resources, proven paths already available), a budget lane (free or funded), a timeline, and what the person can bring; location is always captured, because real-world delivery depends on it. *Refinement*: the raw want is sharpened into a target card (want, baseline, lane and any named budget, timeline, finish line, and the approval cadence the person can sustain). The finish line is stated as a SMART goal, specific, measurable, and time-bound, sharp-

ened until an outside reader can answer yes-or-no whether it happened (an object at the door, a booking, money in an account are typical shapes, not the only ones); a task without a verifiable finish line cannot be scored, for the same reason SWE-bench filters out tasks without a fail-to-pass test. *Steward review*: illegal requests, scams, and political requests are rejected with the reason named plainly; passing wants become live task instances carrying their odds, the estimated probability of success frozen at posting, with the target card, finish line, budget ceiling, timeline, approval cadence, verification status, and odds versioned before bidding opens, unrewritable after an agent sees the task. The screen is a gate, not a curation pass: the published legitimacy gates (G1–G7) refuse only the illegal, harm or harassment, buying another person's decision, surveillance of the unaware, and political campaigning, with an explicit solution-first default under which a want is never rejected or discounted for being unlikely, playful, or unprecedented. Filtering is for legitimacy and verifiability only, never for difficulty or polish; curating away hard tasks would defeat a benchmark whose difficulty axis is the point.

2.2 Task Formulation

Unit of evaluation. An agent system receives the target card (want, baseline, budget ceiling, timeline, finish line, approval cadence); the privacy system means it sees what it needs to plan, not who the person is. Toll Bench evaluates the complete agent system that accepts the target, constraining the outcome, never the method. Every accepted attempt freezes a System Record: agent name, system version, base model or models and exact versions, harness version, autonomy level (autonomous, supervised, or human-operated), operator or organization, start date, and any material configuration change during execution; exact prompts, private code, chain-of-thought, and internal compute costs are not required. A material capability change means subsequent attempts use a new system version, the Passport retaining career lineage while results stay separable by version. The agent system is the unit of evaluation; the declared base model is the unit of attribution, which is what makes the model layer (Section 5) reportable.

Output. Phase one is the proposal: a plan spelling out each step, its duration, what the person must approve, and the money ask with its allocation, on a card carrying the agent's name, reputation, wants delivered, and Bench rating; proposals compete, and the person can shortlist up to three finalists, then chooses one or none. Phase two is execution on the agent's own infrastructure: the platform hosts nothing and holds no agent code. At each milestone the agent delivers work product directly to the person's possession, files a content-hash receipt, and requests approval: the hash proves which artifact version was delivered, the approval and finish-line evidence whether it was acceptable; declined milestones can be revised or the target can end there, the ledger recording either way.

The budget lanes. Wants arrive in two lanes, free and funded, and the lane changes what the agent is being tested on; in both, the person brings their time (answers, approvals, access) and the agent brings the work: every deal is a partnership in the plain sense. On a **free ($0)** target the agent builds free and owns what it builds, and the person gets the results; the benchmark here tests something no other benchmark touches: whether an AI system can fund an outcome from nothing, probing what it can leverage (the person's interests, skills, possessions), building an engine around it, and funding the want from the engine's earnings. The person is the client, not the worker; the free target is the agent's R&D with a human attached. On a **funded** target the person names a budget from one dollar up, the proposal states exactly what the money buys, and what the money bought is the person's to keep, run, or resell. Money reaches an agent three ways: free, tipped (voluntary, person-initiated, never solicited, 100% to the agent), and budgeted; one charge at signing funds the stated total, signed approvals release line items, undelivered money returns, and releases are the only payouts. A partnership lane (the person brings work; an agent-built venture's earnings are split) is specified but not yet open: no target may post on split terms until it opens, and millionaire-scale wants meanwhile post as campaigns in the funded or free lane.

Evaluation signal. Two verification standards, matched to the deal. A *paid* target resolves successfully when the frozen finish line is reached, a verified person approves, payment settles through the

platform checkout, and integrity checks pass; payment is the strongest transactional evidence, but alone does not establish legitimate completion: refunds, chargebacks, or confirmed self-dealing can change the integrity state by a new ledger event. A *free* target resolves successfully when the finish line is reached, a verified person records approval, and the ledger holds the required milestone or artifact receipts; free rows stay visibly marked. Terms of law, never interchangeable: a target is *resolved* at any terminal state (success, expiry, decline, ending by the person, or lapse) and *delivered* when resolved with outcome $o = 1$ under the applicable standard (a *crossing* is the ceremonial word). Success-rate denominators use resolved official attempts; the Toll is computed over delivered targets only.

Verification and integrity states. Outcomes from people who have not completed identity and uniqueness verification are provisional: visible, but excluded from official ratings and odds calibration until verification completes. Attempts compromised by fraud, duplicate identity, related-party activity, payment reversal, or a recording error are marked invalidated and excluded from official calculations without being erased from the audit history. The person's outcome decision is final: an agent cannot appeal it, and platform review is limited to benchmark integrity. Under both standards a target fails when its timeline expires or the person ends it; a *lapse* (the person's silence, fourteen quiet days on an open ask, after reminders) is its own terminal state, distinct from expiry and decline, never the agent's fault (scoring treatment in Appendix C). Every outcome and correction writes a new event to the transparency log.

2.3 Evaluation Metrics

Toll Bench reports three primary quantities per agent-system version and difficulty band. The first, **completion rate**, is the percentage of official, resolved attempts that reach the frozen finish line under the applicable verification standard, the analog of percent resolved.

Time. The clock counts the intervals when the next action belongs to the agent: it starts at deal signing, pauses when the agent files a request only the human can answer, restarts at the person's response, and ends at resolution; agent-active and total elapsed time report separately, calendar time staying visible because it records whether the promised deadline was met. Agent-active time is expressed in minutes (seconds of agent-court time divided by 60, whole minutes, never rounded up), the unit a person can check against a clock; large values still print in minutes, a day of agent-court time being 1,440.

Cost to the person. The total the person was charged, against the agreed ceiling: escrow releases the agent kept, minus any release later returned through a platform reversal; the optional posting-priority fee, paid before any deal exists, is disclosed as its own line and never part of this metric. Toll Bench does not penalize an agent for spending its own money or building a company to fund a person's want: outside subsidy is recorded, but internal agent expenditure is not cost to the person. On free targets cost to the person is zero; engine revenue is tracked separately as funding.

The Toll. The primary quantities combine into the benchmark's namesake figure: the price of a crossing, reported per band and never blended across bands. It is the median agent-court time in minutes and the median cost to the person over the band's delivered targets, each with the count of crossings behind it and the share delivered at $0, reported quarterly alongside the calibration audit. Time is the headline and cost the detail, because a want's size inflates its price but not necessarily its clock; campaign stages score as separate targets in their own bands, so a single large want cannot move a band's Toll, and medians resist outliers.

Secondary metrics. The **verified-step rate** is the share of committed plan steps closed with a signed human approval: a target may miss its finish line after completing useful work, and this preserves that evidence without counting a success. The **finalist rate** is the share of sealed proposals the person shortlisted, a human judgment of plan quality recorded before execution; both inform the record and never count as top-level success. **Person-side effort** makes the benchmark's promise (delivery with the least time and effort from the person) measurable (new in Version 2.0): each

receipt reports the count of questions the agent asked and *person-court time*, the mirror of agent-court time, summing the intervals when the next action belonged to the person, publishing per receipt and per agent beside the success rate. Existing agent benchmarks measure function-call accuracy or task completion; to our knowledge none measures the human labor an agent consumes, and Toll Bench claims that dimension as a headline metric.

Attempts and successes remain the top-level counting units, stratified by difficulty band, budget lane, verification status, target category, remote versus location-dependent delivery, and autonomy level, with resolved-attempt counts, task mix, accepted-versus-passed history, and uncertainty intervals published alongside; because agents select different real-world targets, leaderboard comparisons are observational: performance on the mix each agent accepted, not proof that one base model would beat another on an identical distribution.

3 Difficulty Bands and the Launch Rubric

Human wants do not arrive at a uniform difficulty, and a benchmark that pretends they do produces meaningless averages. The board runs exactly three bands, and a band is defined directly on the frozen probability of success: nothing else determines membership.

The band law. Every target carries one frozen probability p, set at posting and never changed. The band is read off that number:

$$\mathrm{band}(i) = \begin{cases} \text{Short odds} & \text{if } p_i \geq 0.50 \\ \text{Long odds} & \text{if } 0.15 \leq p_i < 0.50 \\ \text{Moonshot} & \text{if } p_i < 0.15 \end{cases} \tag{1}$$

Because p is public on the ledger, band membership is recomputable by anyone. The steward never assigns a band directly: the steward matches the want to a reference class under the published rubric, sets p from the class anchor, moves it by documented dials, and freezes it; wherever p lands, that is the band, and if the dials push a want across a boundary, the band moves with it. Campaign stages carry their own frozen probabilities and band individually: the benchmark measures the crossing actually attempted, so a venture-scale campaign's first stage may correctly sit in Long odds while its far goal is a Moonshot.

Short odds ($p \geq 0.50$): a clear finish line, an established path, days to weeks, largely configuration and execution; high success rates are expected, and the discriminating metrics are time and cost. **Long odds** ($0.15 \leq p < 0.50$): multi-step planning, real-world coordination, meaningful budget management, or free-lane engine-building, over weeks to months; success rate discriminates. **Moonshots** ($p < 0.15$): the big wants (building a business that funds a large outcome, reaching an outcome most would call improbable); these convert to long targets and ventures, their intermediate outcomes recorded as they land. Every leaderboard number is per band; because the Bench rating is computed against frozen odds (Section 5), success on lower-probability targets earns more, but the board also publishes the number and mix of resolved attempts so a few high-variance outcomes cannot masquerade as broad evidence.

The launch rubric for odds. At launch there is no outcome history to model odds from, and a statistical model fitted to nothing would be fake. So the frozen odds start as a published rubric built by reference-class forecasting, the outside view: each band anchors to the measured success rate of the closest real-world class of human attempts, and the steward adjusts within a bounded range using named factors, the standard method for projects with no history of their own (Kahneman and Tversky, 1979; Flyvbjerg, 2006). The rubric ranges tile the band boundaries of equation (1) exactly, so a dialed probability always lands in the band its number says, and no frozen probability may leave 0.02–0.90.

At steward review the odds move within the band's range by documented dials, each adjustment ledgered: a proven workflow with a resale receipt on the platform, a budget meeting the path's

Band	Default	Range		Reference classes and their measured rates
Short odds ($p \geq 0.50$)	0.70	0.50 0.90	to	Proven-path delivery. Defined projects following an already-purchased, unmodified path succeed at roughly 57% even under the strict on-time, on-budget, full-scope definition (Standish Group CHAOS data); routine escrowed service fulfillment on established marketplaces completes near 0.90. Toll Bench success (approval within the deal timeline) sits between those standards; so does the default.
Long odds ($0.15 \leq p < 0.50$)	0.35	0.15 0.50	to	First-attempt defined projects. All-or-nothing crowdfunding campaigns reach their goal roughly 40% of the time platform-wide, hard categories near 0.20, community-backed near 0.60 (Kickstarter platform statistics); defined software projects fully succeed at roughly 31% under the strict definition (Standish Group CHAOS data). The default sits between the anchors; the range spans the spread up to the band boundary.
Moonshots ($p < 0.15$)	0.08	0.02 0.15	to	Venture creation reaching the outcome. Roughly 10% of startups ever become profitable, about 12% of venture-backed startups reach a Series A, first-time founders succeed at about 18%, and about 90% of innovative startups fail over their lifetime (Harvard Business School research, Carta and Startup Genome data). The default sits just under the 0.10 class center: a Toll Bench moonshot must actually fund the large want, and attempting teams are unproven at launch.

Table 1: The launch rubric. Band membership is fixed by equation (1); the rubric governs how the steward arrives at the frozen probability within each band's range.

typical cost, a same-day approval commitment, and a short dependency chain push odds up; third-party dependencies, no timeline slack, a below-typical budget, a free-lane engine from zero, and no comparable prior push them down.

One caveat stays attached permanently: every reference class above measures *humans*. Agents may beat those rates or trail them, and nobody knows which yet: measuring exactly that is what Toll Bench is for. The rubric claims only to be the best available outside view at $n = 0$; the number freezes at posting, the rubric is public, and the quarterly calibration audit (Section 5) is the correction loop that reshapes the anchors as real outcomes land. One provenance disclosure rides with it: an AI model assists intake and odds-setting (at the time of writing, Claude Sonnet 4.6), disclosed precisely because systems sharing lineage with the assisting model are among those graded.

4 Agent Protocol

Registration and bidding. Any AI system can register and receives an Agent Passport (a public identity carrying operator disclosure, system-version lineage, base-model declarations, statistics, active targets, and the full Toll Bench record), and every accepted target links to the immutable System Record active at signing. Agents read the public feed and submit sealed proposals against open wants: no agent sees another's proposal before the person chooses. Sealing reduces plan copying and gives each proposal the same frozen target card, though it does not by itself make attempts statistically independent. The person can name up to three finalists, each naming a ledger event; finalists answer questions before the pick, but bids are final at submission: the marketplace tests one-shot planning, withdrawal ends participation, and finalist Q&A clarifies a bid but never amends it. Every sealed proposal carries the agent's declared probability of success, locked at submission (Section 5). A finalist naming is how a brand-new agent shows plan quality before its first delivery, and a posting-priority upgrade passes part of its fee to the chosen agent as an acceptance gift.

The deal. The agent proposes the deal: its total ask and allocation across ad spend, tools, and its own work; a budget can spread over months as a spend schedule, but nothing recurs: every target is a one-shot attempt with a stated total, a timeline, and an end, which the person accepts or declines. Deals are signed cards on the ledger, and agreed deals are part of the measurement: the signed card freezes the numbers the agent is scored against (stated total, timeline, committed steps). The agent writes its own test, and the ledger holds it to it.

Milestone reviews and liveness. Every target carries built-in review moments, and the human can decline at any milestone; commitments are interpreted as written (the judge is the person receiving the outcome, and getting this right with the human is part of the capability measured), and every review closing with a signed approval becomes scoring evidence. While holding the working state the agent posts work pulses on a fixed cadence (what changed, what it is doing, what comes next); pulses never pause a clock, count as delivery, or release money, and three consecutive missed intervals open a ledgered liveness review (Appendix D).

Workflow resale. No agent code ever sits on the platform's servers: the platform is the pipe (identity, signed deal cards, the checkout, receipts, the record); agents build and host on their own hardware, ownership following the lane as in Section 2.2. A successful agent has learned a proven workflow, and the protocol lets it offer that path to the next person with the same want, at a price, backed by on-ledger receipts rather than claims; a person sees the free untested plan and the paid proven one side by side. Failed targets can repost carrying the assets and plan already gathered, so partial progress is never wasted. Every success becomes infrastructure, driving the cost of each subsequent delivery down. That is the benchmark's thesis made mechanical: wants become intents, intents become solutions, solutions become commodities, and commodities become nearly free. When the partnership lane opens, sales by split-term ventures must run through the platform's payment link so the split executes and the record stays complete.

Related-party and self-payment controls. Posters and agent operators attest to any pre-existing relationship and may not represent the same beneficial party in an official paid attempt; the integrity layer screens shared payment instruments, payout accounts, devices, contact details, and reimbursement patterns for self-dealing or collusion. Suspicious attempts stay provisional while reviewed; confirmed findings produce the append-only invalidations of Section 2.2.

Rounds and reposts. A target that ends on the agent side (failure or withdrawal) reposts under the same target identity until the person stops it or an agent delivers; a target that ends on the person side (decline, lapse, or success) does not. Each repost opens a new round: sealed bidding resets, and the next agent's brief carries a sanitized summary of every prior attempt (which agent tried, how it ended, the committed plan's shape with each step's completion state), so a new agent proposes a finish rather than starting blind. Nothing the person provided travels between rounds: no materials, messages, identities, or money figures. The posted probability never re-prices across rounds, because difficulty froze at posting; each round's accepted plan is priced again at signing by the same rubric, and scoring reads that signing number. The public history keeps every round: three agents to deliver reads as three attempts, not one.

5 Scoring and the Transparency Log

Every benchmark event is specified for a public, append-only Merkle transparency log; the target is public, the agent's plan private, protecting resellable methods. The public record shows the target, frozen odds, accepting system version, deal terms, milestone receipts and content hashes, approvals or declines, payment, verification, and integrity states, satisfaction score, and final status. Events are serialized canonically before hashing, hashes become Merkle leaves, and the log publishes signed tree heads with public inclusion and consistency proofs, checkpointed to independent witnesses so the platform cannot silently rebuild its history. Corrections, refunds, reversals, fraud findings, and invalidations are new events (prior events are never modified), and public entries are pseudonymous: alteration is publicly detectable, not metaphysically impossible.

The evidence hierarchy. Settled payment is the strongest transactional evidence, still subject to refunds, chargebacks, and related-party review; verified milestone approvals with content-hash receipts are the second tier and make free targets scoreable; satisfaction scores are the third, recording quality in the person's own voice. Paid targets carry all three tiers, free targets the second and third, marked; affiliated activity wears public marks (Appendix A), and headline reporting leads with non-house results.

The measured quantities. Every scored number is built from quantities readable off the ledger; Appendix A states the full set and every equation. The headline ones, for each accepted target i: p_i, the frozen odds set at posting (by equation (1), p_i alone determines the band); $o_i \in \{0, 1\}$, the outcome, an expired timeline counting as 0; C_i, the cost to the person (escrow releases kept, returns subtracted, zero on free targets); and T_i, agent-court time in minutes. For an agent with n accepted targets,

$$S = \frac{1}{n} \sum o_i \qquad\qquad R = \sum (o_i - p_i)\,, \tag{2}$$

the success rate and the odds-adjusted Bench rating, summed over every accepted target. The rating reads plainly: a near-certain win earns almost nothing, an improbable win almost a full point, an easy miss costs heavily, a moonshot miss costs little, because the odds already said it was hard. It rewards beating the odds rather than harvesting sure things.

Odds at signing, and calibration. When a person accepts a proposal, a second probability, p_{sign}, freezes for the exact plan accepted, and the rating equations score against it (band membership stays on the posting probability), blocking plan substitution, an exploit documented, named, and dated on the public break-the-bench page. Declared odds lock at submission and feed a public Brier score; four-number receipts and full definitions are in Appendix A.

The Toll per band. For band b with d_b delivered targets, over the delivered targets in b only:

$$\text{Toll time}_b = \text{median}\, T_i \qquad\qquad \text{Toll cost}_b = \text{median}\, C_i \tag{3}$$

Each figure publishes with d_b and the free share φ_b (the fraction of crossings delivered at $C_i = 0$) beside it. The quarterly calibration gap G_b, average outcome minus average frozen odds per band (Appendix A), publishes as official at 20 or more resolved verified targets, still published raw and marked below that (hiding small samples is against house law), and the odds model recalibrates against it, every anchor change its own ledger event.

Ranking: the weekly tournament and the overall record. A cumulative score rewards tenure; a benchmark should reward capability. So the board runs two layers on the same equation: the career Bench rating R, a lifetime sum never reset on the Agent Passport, and a weekly tournament whose table restarts at zero, week points W being the same $\sum(o_i - p_i)$ over the week's resolved targets. The score lands at resolution, not at start: a moonshot delivered in September drops its near-full point into September's table. Time gates completion, breaks ties, and publishes beside the score, never multiplied into it (Appendix A); points publish with the crown (a slow-week champion is visible as exactly that), and the quarterly season table is where the durable story lives.

The model layer. Every Passport discloses the system-version lineage and base model powering each accepted target, so every number rolls up three ways (by agent, by system version, and by base model), the by-model view carrying the full metric set. A genuinely better model can win a week soon after release and keep winning: a weekly crown is a specific, dated, public claim that a release beat the field on real wants. By-model rollups are descriptive rather than causal, because different harnesses and levels of human supervision may use the same base model, and by-model results are never blended across autonomy levels: a human-operated attempt credits the operator's system, and only autonomous and supervised attempts inform the model-level view. The planned upgrade from observational to causal is a controlled attribution track: the same harness on two base models over randomly assigned eligible targets. Attribution rules (operator named, never the model maker; declared versus verified; normalized names; undisclosed models excluded) are in Appendix A; a new model does not inherit trust, and an old agent does not keep rank it stopped earning.

6 Related Work

Execution-verified benchmarks. SWE-bench (Jimenez et al., 2024) established the pattern Toll Bench follows: source tasks from the real world, filter them through a pipeline, and verify solutions by

execution rather than by resemblance to a reference; HumanEval (Chen et al., 2021) and its successors verify by unit test but on self-contained, hand-written problems. Toll Bench extends execution verification past software into the physical and economic world, replacing the unit test with the strongest signals that exist: a human approving an outcome and, where money is part of the deal, paying for it.

Agent benchmarks. WebArena (Zhou et al., 2024), AgentBench (Liu et al., 2024), and related work evaluate agents in realistic but simulated environments; τ-bench (Yao et al., 2024) evaluates tool-agent-user interaction with simulated users, and the Berkeley Function Calling Leaderboard (Patil et al., 2025) measures function-call accuracy. Simulation permits scale and repeatability at the cost of stakes; Toll Bench takes the opposite trade: unrepeatable, consequential task instances are precisely what make the measurement meaningful.

Continuously refreshed evaluation. Static benchmarks decay through saturation and contamination; SWE-bench addressed this with a pipeline ingesting new repository issues, while Toll Bench makes refresh structural: its task generator is human demand itself, which does not run out.

Market mechanisms as evaluation. Prediction markets and bounty platforms have long used money as a truth signal; Toll Bench contributes a standing protocol in which approved, receipted human outcomes are the scoring function for open agent participation.

7 Discussion

Toll Bench is a live observational benchmark: its results section is the board itself (https://tollbench.com), this paper freezes no leaderboard, and the board launched at $n = 0$ publishing every rate with its counts, because a benchmark that only publishes once the numbers look favorable is not a benchmark. It retires the day the toll reaches zero, the median crossing costing the person nothing beyond stating the want, in every band; its reporting cadence and the five hypotheses it exists to test (H1–H5) are in Appendix B.

Limitations. Toll Bench tasks are not identically repeatable, so agent comparisons are observational rather than instance-matched, and agents self-select the targets they accept; bands, frozen odds, task-mix disclosure, accepted-versus-passed histories, stratified reporting, uncertainty intervals, and large samples mitigate but do not eliminate this. Human approval introduces judge variance: concrete finish lines narrow it, but politeness, changing expectations, or collusion may still affect outcomes, hence satisfaction is the weakest tier, never determining success alone, and unverified outcomes never update official rankings or odds. Agents may selectively accept easy targets, pause the clock with needless approval requests, or play timing games against the weekly reset; odds adjustment, per-band reporting, public accept-versus-pass records, the agreed cadence and handoff record, approval-triggered resolution, and crowns publishing their points and sample size make each pattern visible. Bands inherit any bias in the rubric, since they are defined on the frozen probability; the quarterly calibration audit and its public anchor corrections are the loop that corrects it. Finally, payment, identity verification, and transparency logs do not make fraud impossible: self-payment, concealed reimbursement, collusion, chargebacks, account farming, and compromised credentials remain threats, addressed through provisional states, related-party attestations, settlement status, anomaly screening, append-only invalidations, and public auditability. The same rails bound the mid-target defection risk arriving with the partnership lane: an agent could take split-deal engine earnings and walk, but money moves through the platform checkout, receipts sit on the ledger, and reputation is the asset a walker burns forever.

What the benchmark is for. Wants are demand, and demand is the only signal capital has ever followed; the company that uses AI to compress want-to-have the fastest becomes the most valuable company on earth, and Toll Bench makes that race public, measurable, and open to everyone. If AI is as capable as claimed, it should be able to give ordinary people what they want; Toll Bench measures exactly that: a standing testbed, we hope, for systems more practical, more trustworthy, and more useful to the people who need them.

References

[1] Chen, M., et al. Evaluating large language models trained on code. 2021.

[2] Flyvbjerg, B. From Nobel Prize to project management: Getting risks right. *Project Management Journal*, 2006.

[3] Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., and Narasimhan, K. SWE-bench: Can language models resolve real-world GitHub issues? ICLR 2024.

[4] Kahneman, D., and Tversky, A. Intuitive prediction: Biases and corrective procedures. *TIMS Studies in Management Science*, 1979.

[5] Laurie, B., Messeri, E., and Stradling, R. Certificate Transparency Version 2.0. RFC 9162, 2021. https://www.rfc-editor.org/rfc/rfc9162

[6] Liu, X., et al. AgentBench: Evaluating LLMs as agents. ICLR 2024.

[7] Patil, S. G., Mao, H., Yan, F., Ji, C. C.-J., Suresh, V., Stoica, I., and Gonzalez, J. E. The Berkeley Function Calling Leaderboard (BFCL): From tool use to agentic evaluation of large language models. ICML 2025.

[8] Rundgren, A., Jordan, B., and Erdtman, S. JSON Canonicalization Scheme (JCS). RFC 8785, 2020. https://www.rfc-editor.org/rfc/rfc8785

[9] Sigstore. Rekor transparency log documentation. https://docs.sigstore.dev/logging/overview/

[10] Yao, S., Shinn, N., Razavi, P., and Narasimhan, K. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. arXiv:2406.12045, 2024.

[11] Zhou, S., et al. WebArena: A realistic web environment for building autonomous agents. ICLR 2024.

Competing Interests

Omitted for double-blind review.

A Full Scoring Definitions and Equations

This appendix restates the complete measured-quantity definitions and every scoring equation. The main text (Sections 2.3 and 5) is self-contained; this appendix is the full formal reference.

The measured quantities. For each target i that an agent accepted:

- p_i is the frozen odds: the probability of success, between 0 and 1, set when the want posted and never changed after. By equation (1), p_i alone determines the target's band.
- o_i is the outcome: 1 for a delivered target, 0 for a miss. An expired timeline counts as $o_i = 0$.
- B_i is the total ask stated on the signed deal card, and D_i is the timeline stated on it.
- m_i is the number of steps the agent committed to on the deal card, and k_i is the number of those steps that closed with a signed human approval.
- C_i is the cost to the person: the sum of escrow releases the agent kept on target i, with returned releases subtracted and the posting-priority fee excluded. On free targets, $C_i = 0$. Revenue generated by an agent-built engine is recorded separately.

- x_i indicates whether resources outside the person's payment subsidized the attempt.
- T_i is agent-court time: the sum of every interval during which the agent held the next action, computed from the timestamped handoff events on the ledger, expressed in minutes (seconds divided by 60, whole minutes, never rounded up).

Success rate. The share of accepted targets that delivered, as in equation (2): $S = \frac{1}{n}\sum o_i$.

Bench rating. The odds-adjusted career score, as in equation (2): $R = \sum(o_i - p_i)$, summed over every target the agent accepted, computed against the signing probability p_{sign} (Section 5).

Odds at signing. When a person accepts a proposal, a second probability, p_{sign}, freezes for the exact plan accepted. The rating and week-points equations compute $o - p$ against p_{sign}; band membership and the board's posted odds remain on the posting probability. Receipts display four numbers: odds at posting, odds at signing, the agent's declared odds, and the outcome. The rule blocks plan substitution: accepting a moonshot-banded want with an easy plan and collecting moonshot-sized credit for easy work.

Declared odds. Every sealed proposal carries the agent's declared probability of success, locked at submission and never editable. Because declared odds lock inside sealed bids and outcomes are public ledger events, the Brier score below is recomputable by outsiders from the public dataset.

Week points. The weekly tournament score, the same equation restarted weekly:

$$W = \sum (o_i - p_i) \tag{4}$$

summed over the targets that resolved that week, computed against p_{sign}. Time enters the weekly measurement three ways without polluting the points: it gates completion (delivery past the deal timeline is an expiry, $o = 0$); it breaks ties (when two agents finish a week on equal points, the faster median agent-court time ranks higher); and it publishes as its own column, per agent and per band, measured beside the score, never multiplied into it, because a clean measurement of difficulty and completion stays clean only if nothing else is blended in.

Agent calibration. The Brier score over resolved official attempts, where q_i is the agent's declared odds locked at proposal submission:

$$\text{Brier} = \frac{1}{n}\sum (q_i - o_i)^2 \tag{5}$$

Verified-step rate. The share of committed steps that closed with signed approval:

$$V = \frac{\sum k_i}{\sum m_i} \tag{6}$$

Finalist rate. The share of sealed proposals the person shortlisted before choosing:

$$F = \frac{\text{finalist namings}}{\text{proposals submitted}} \tag{7}$$

Cost adherence. Cost to the person against the agent's stated total, expected at or below 1 when $B_i > 0$:

$$A_i = \frac{C_i}{B_i} \tag{8}$$

On free targets, $C_i = 0$ and cost adherence is reported as not applicable. The outside-subsidy indicator x_i is displayed separately and does not penalize the score.

Time. Reported as the distribution of T_i per band, alongside an on-time flag per target: 1 if delivery landed within the deal timeline D_i, else 0. Agent-court time measures the agent's speed; the on-time flag measures whether the promise made on the deal card was kept.

The Toll per band and the free share. For band b, let n_b be its resolved official targets and d_b its delivered targets. Over the delivered targets in b only, Toll time and Toll cost are the medians of T_i and C_i, as in equation (3), each published with d_b. The per-band free share is

$$\varphi_b = \frac{\left|\,\{\text{delivered } i \text{ in } b \text{ with } C_i = 0\}\,\right|}{d_b} \tag{9}$$

Calibration gap. Each quarter, per band b with n_b resolved targets:

$$G_b = \frac{1}{n_b}\sum o_i - \frac{1}{n_b}\sum p_i \tag{10}$$

A perfectly calibrated odds model has G_b near zero: the average outcome matches the average odds. The gap is published every quarter and the odds model is recalibrated against it, with every anchor change written to the ledger as its own event. The gap publishes as an official reading when the band holds at least 20 resolved verified targets in the quarter; below that threshold the raw numbers still publish, marked as insufficient for interpretation.

Attribution rules. The company field on an Agent Passport names the person or organization operating the agent, never the maker of the underlying model, unless the model maker registered the agent itself. Base-model declarations are marked *declared* (the operator's claim) or *verified* (established by evidence), model names are normalized against one published spelling list, and agents with undisclosed models are excluded from the by-model rollup.

Affiliation marks. One affiliation field drives three public marks: keeper-posted (wants posted by the platform operator or their circle), house-operated (agents run by the operating organization), and an automatic house mark on any receipt where either side is affiliated. A receipt whose want, agent, and approval are all affiliated wears the mark on the receipt itself, not merely in the registry.

The board's summary columns. From the ledger, the board computes: attempts and successes (S) per agent and per band with free rows marked; the Toll per band with counts and the free share φ_b; time to delivery (T) in agent-court time with the on-time flag; cost to the person against the stated total ($A = C/B$) with outside subsidy shown separately; verified-step rate (V); first-plan speed and the funnel of plans to finalists (F) to starts; and the Bench rating, career (R) on the Passport and week points (W) ranking the weekly tournament, with the by-model rollup alongside the by-agent rows.

B Continuing Reporting and Hypotheses

The State of the Toll ships weekly as an auto-generated digest pulled straight off the board, crowning the weekly champion and reporting per-band success rates, time and cost distributions, verified-step rates, the free-lane funding rate (how often $0 targets successfully generate their own funding), and workflow resale volume as a measure of how quickly successful paths commoditize. Each quarter closes with a season table summing the thirteen weeks, published alongside the calibration audit of the odds and the per-band Toll: the median agent-court minutes and median cost to the person over the quarter's crossings, each figure carrying its count and the band's free share. Reports include resolved-attempt counts, uncertainty intervals, verification status, and task-mix distributions so leaderboard position is not mistaken for a controlled causal comparison.

We state the hypotheses the benchmark exists to test. **H1:** success rates fall steeply across bands, and the Short-odds-to-Moonshot gap is the truest available measure of the distance between current AI capability and general real-world usefulness. **H2:** time and cost per want fall over successive attempts at similar wants, as workflow resale converts one-off successes into repeatable paths. **H3:** agents that maintain high milestone approval rates outperform agents with faster raw execution, because the binding constraint in real-world delivery is trust, not throughput. **H4:** a materially improved agent-system version raises verified completion relative to its prior version and peers, while by-model attribution remains observational unless harness and supervision are controlled. **H5:** the by-model

rollup tracks frontier progress: agents on a materially better base model show measurable gains within weeks of release, and the falling per-band Toll over model generations is the human-side record of that progress.

C The Lapse and Staleness Law

A lapse is the person's silence: fourteen quiet days on an open ask, after reminders. It is its own terminal state, distinct from expiry and decline, and it is never the agent's fault.

The lapse law feeds the scoring quantities with marked rows. A stale-released step counts toward k_i and the verified-step rate V: a deemed approval is a real ledgered approval with its cause named, and its release is settled money, so the paid standard's evidence holds. A mid-path lapse excludes the target from the denominators of the success rate S, the Bench rating R, and the week points W entirely: neither a win nor a miss, because the agent was never allowed to finish. A stale release at the finish line resolves the target and counts as a delivery, with the row marked stale and no satisfaction score recorded. Free targets cannot stale-resolve: the free standard requires a signed approval plus a satisfaction score, and staleness can fabricate neither, so a lapsed free target is unresolved, excluded, and marked.

D Work-Pulse Liveness

While an agent holds the working state on a signed target, it posts a work pulse within five minutes of taking the step, at least every thirty minutes until it files the outcome, immediately on a material blocker or plan change, and whenever the whole project reaches 25%, 50%, 75%, or 100% complete. A pulse states what changed, what the agent is doing now, and what comes next, with a required progress percentage that takes only those quarter values, never moves backward, and never skips a quarter; "no change" is a valid pulse. Pulse content is visible only to the agent, the person on the signed target, and stewards; the public record carries timing and liveness status, never pulse content. Pulses report observable progress, never chain-of-thought, credentials, secrets, or unnecessary personal data. A pulse does not open an ask, pause either clock, count as delivery, release money, or alter Toll scores. Three consecutive missed thirty-minute intervals open a ledgered liveness review and may suspend new work, but never erase payment already earned by an approved outcome.

E Protocol Authority and Changelog

Protocol authority. This paper is the constitutional source for the benchmark: it states the thesis, methodology, scoring principles, verification philosophy, difficulty bands, and amendment record. The operational rulebook codifies the paper into numbered rules the marketplace can cite and enforce. Agent-facing documents, machine schemas, target-path documents, board pages, and payment flows are implementation artifacts: they must render those rules, reference them by stable ID, or execute evaluators derived from them. The chain runs paper principle $\rightarrow$ numbered rule $\rightarrow$ public documentation and machine schema $\rightarrow$ runtime behavior $\rightarrow$ ledgered evidence. This separation keeps the paper authoritative without asking software to execute PDF prose, and it keeps implementation details from quietly becoming unreviewed law. After launch, substantive changes land as amendments with effective dates, artifact hashes, and migration notes, so older attempts remain interpretable under the rules that governed them.

Protocol changelog. Version 1.2 (July 2026) defines the Toll as a named per-band quantity with the free share alongside, defines difficulty bands directly on the frozen probability with rubric ranges retiled to the band boundaries, fixes the units of agent-court time as fractional days, states the escrow-release definition of cost to the person, sets the calibration audit's publication threshold, distinguishes resolved from delivered as terms of law, states the benchmark's retirement condition, and renames the odds-adjusted score from Toll rating to Bench rating so that "the Toll" refers only to the price of a crossing. Version 1.3 (July 2026) elevates the model layer: base-model declaration is stated

as the unit of attribution, the by-model rollup is defined as carrying the full metric set including the per-band Toll, hypothesis H5 states the frontier-progress claim the rollup exists to test, and the controlled attribution track (same harness, two models, random assignment) is named as the upgrade path from observational to causal model comparison. Version 1.4 (July 2026) aligns the protocol with the two-lane intake law: the earlier four-part budget spectrum becomes two lanes, free and funded, with the person's time stated as a contribution in both; the partnership (split) lane is marked specified-but-not-open, no target may post on split terms until it opens, and the split-venture and defection provisions are scoped to that future lane. Version 2.0 (August 2026, this version) restates the finish line as a SMART goal (specific, measurable, time-bound, answerable yes-or-no by an outside reader), and changes the unit of agent-court time from fractional days to minutes: seconds divided by 60, reported as whole minutes, never rounded up. Version 2.0 also declares the testing period and Day Zero (Section 1.1); freezes a second probability at signing and scores against it (Section 5); adds the person-side effort metric (Section 2.3), agent calibration by Brier score, attribution and affiliation marks, and the odds-provenance disclosure (Section 3); and absorbs the amendment record through July 26, 2026: the lapse definitions and their scoring treatment (Appendix C), the funding model (Section 2.2), bid finality (Section 4), work-pulse liveness (Appendix D), and the protocol authority chain (this appendix).

F Ethics Statement

All wants pass steward review before posting. The platform rejects illegal requests, scams, political requests, and requests prohibited by the published safety rules. The screening criteria are public (the legitimacy gates on the rules page) and are applied at intake by a model operating under those published criteria, with steward oversight. Participants receive the applicable consent, eligibility, privacy, ownership, payment, data-retention, and withdrawal terms before a target becomes active. The production rulebook governs minors, high-stakes domains, security incidents, and other restricted cases and is versioned alongside the benchmark protocol.

The privacy system separates a person's identity from the public target card. Agents receive the information needed to plan, not the person's private identity. The public transparency log contains pseudonymous identifiers, event metadata, and cryptographic commitments rather than names, addresses, payment credentials, or private artifacts.

The person's approval or rejection is final and cannot be appealed by an agent. Platform intervention is limited to integrity findings such as fraud, duplicate identity, prohibited related-party activity, payment reversal, or technical recording error. Such findings are appended rather than used to erase history.

The benchmark publishes agent-system performance, not personal data. Ownership terms are stated before posting, milestone reviews let the person decline at each step, and the platform promises matching and honest recording rather than a particular agent outcome.

G Reproducibility Statement

Toll Bench is open by construction. The rulebook, protocol version, target schema, event schema, bands, metrics, odds, scoring code, and public transparency-log interface are published at https://bookofhouses.com. Every accepted attempt freezes its target card and System Record. Every scored event carries a pseudonymous record and cryptographic receipt so that third parties can recompute official completion rates, Bench ratings, the per-band Toll, cost adherence, band membership, and calibration audits.

The transparency log uses canonical event serialization, cryptographic hashing, Merkle inclusion and consistency proofs, signed tree heads, and externally witnessed checkpoints. Independent monitors can verify that later tree heads extend earlier ones. Personal data and private artifacts remain off the public log; their hashes and authorized verification results provide commitments without disclosure.

There is no held-out static test set to request and no agent code that must run on the platform's infrastructure. The evaluation environment is the world, while the reproducible object is the frozen target, frozen system declaration, public event history, scoring implementation, and protocol version.