Rent the pipes. Own the judgment.

Background to the post of the same name, drawn from the working paper How Do We Use This?

Berg Atkinson, Wolfberg LLC · berg@wolfberg.ai

Preprint. Not peer reviewed. Reproduced from the working paper as revised 21 September 2026 and corrected 23 and 25 September 2026.

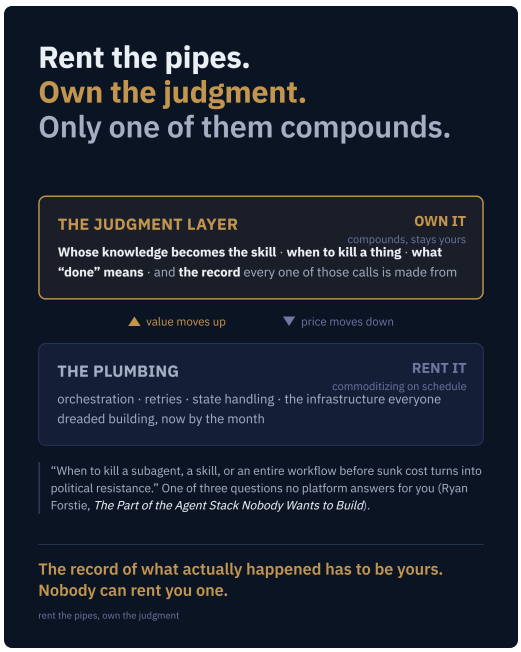
IN PLAIN TERMS

The infrastructure under AI agents (orchestration, retries, state handling) is becoming a rented service, and renting it is usually right. What does not come with it are three decisions: whose knowledge becomes the standing instructions, what result ends an experiment, and what "done" means.

The widely used agent frameworks the paper checked end a run when the model stops calling tools, and each ships its real checks empty until someone fills them.

The practice's own answers are set out here: the machine proposes changes and a person merges or refuses them; decision rules are fixed before the data are seen; and the record every decision is made from stays with the operator.

About this paper. This is one of seven short background papers written to go with a series of plain-English posts. Every section below is reproduced word for word from the working paper *How Do We Use This?* (Atkinson, 2026; revised 21 September 2026, corrected 23 and 25 September 2026), and keeps that paper's section, figure and table numbers, so a reference to a section or figure not reproduced here resolves in the full paper. Only the plain-terms summary, the post's figure, and the short labels that say which section each passage comes from were written for this part. The full paper is linked from every post in the series.



The post's figure. The judgment layer, owned, over the plumbing, rented, with one of the three questions from Forstie (2026). A teaching summary of §8.6.

FROM §8.6 · THE JUDGMENT LAYER

8.6 The judgment layer

The counterweight is one instrument inside a set of decisions the practice cannot delegate, and the set deserves naming because the market is commoditizing everything around it. Over 2026 the infrastructure beneath working agents, the orchestration, retries and state handling, became a managed rental; this practice rents such tooling and expects to keep renting it. What did not commoditize are the decisions that infrastructure exists to execute: whose knowledge becomes the standing instruction set; what result, fixed before work opens, ends an experiment; and what "done" means for a given piece of work. A practitioner essay contemporaneous with this program posed these three questions as the unanswered residue of managed-agent platforms (Forstie, 2026, cited under Prior art). The practice's answers are machinery this paper has already described: changes to the standing instruction set are proposed by the machine and merged or refused by the operator; experiments carry decision rules ratified before their data are seen, under which a null closes the program (§10.4); and the finish line is ruled in advance. The record is the load-bearing element in each: every one of those decisions is made from the practice's own logs, which is why §7 treats ownership of the record as a first-order property rather than a storage preference. An execution history that lives on a platform's dashboard, in the platform's format, is testimony in the sense of §6, not a record the operator holds.

FROM §8.3 · WHAT THE RENTED LAYER SHIPS WITH

It is worth recording what the default is, checked in September 2026, because it is lower than the discussion above implies. In the OpenAI Agents SDK, an agent run ends when the model emits a turn containing no tool calls; with no output type configured, any text at all satisfies the condition (OpenAI, 2026a). The framework does ship a human-in-the-loop approval primitive, and it gates tool calls rather than completion claims, so a developer can require a person to approve a refund before it is issued and has nothing available to require a person to confirm the task was actually done (OpenAI, 2026b). The predecessor framework ended a run on the same no-more-tool-calls condition with no turn limit at all (OpenAI, 2024). The pattern repeats across the frameworks we checked. In the Claude Agent SDK the loop likewise ends when the model returns a response containing no tool calls, with no turn cap and no budget cap set by default, and the result carries the subtype *success*, which is a statement that the loop terminated without error rather than a claim about the answer (Anthropic, 2026a). In CrewAI the completion test is that the literal string "Final Answer" appears in the agent's own generated text (CrewAI, 2026a), and a guardrail specified as a string is executed by the acting agent's own model (CrewAI, 2026b). Each of these frameworks offers a real gate, and in each case it is opt-in and empty until a developer fills it.

Two details from that survey are worth stating on their own, because they come from the vendors rather than from us. Claude Code, whose agent loop the Claude Agent SDK embeds, ships a built-in completion condition in which a separate small model checks after each turn whether a stated goal has been met, and its documentation says of that evaluator that "it does not call tools, so it can only judge what Claude has already surfaced in the conversation" (Anthropic, 2026b). That is an accurate description of the limit this paper is about, published by the party with the most incentive to describe it favorably. The same vendor's guidance on building agents says of having one model judge another that "this is generally not a very robust method" (Anthropic, 2025). Meanwhile the human-in-the-loop primitives in the two SDKs gate *tool calls* and not completion claims: a developer can require a person to approve an irreversible action before it is taken, and has nothing built in to require a person to confirm that the finished work was actually correct. CrewAI is the exception among the three: a task can be set to have a human review the agent's final answer, and like every other gate here that setting is off by default (CrewAI, 2026b). Outside that one opt-in, the approval surface exists for the act and not for the claim, which is the asymmetry the counterweight is aimed at. This is not a criticism of those libraries, which are explicit about what they are; it is the baseline against which every mechanism in this paper should be read. The common case is not a weak check. It is no check, and a stop the agent declares for itself.

One open-source harness makes the distinction visible by shipping both answers at once. Nous Research's Hermes agent has a stop-time verification path that parses the terminal log for real test, lint and build invocations and records their actual exit status, which is evidence causally downstream of the world rather than of the agent's narrative. It also has a standing-goal judge that calls an auxiliary model with the goal text and roughly the last four kilobytes of the agent's own final response, with no tool access and, by default, the same model as the agent. The project's own issue tracker records the predictable failure: an agent reported writing a file, the write silently failed, and the judge marked the goal complete. The release carrying this work is announced with the line that done means proven rather than claimed. Half of it is; the other half is the stop problem with a second model attached, and it

is the half that looks most like verification. We take these details from the project's public repository, configuration and issue tracker, and they are current as of September 2026 rather than permanent (Nous Research, 2026).

The failures in §6.3 led the practice to adopt a design rule in September 2026: an instrument the measured party operates will rot; an instrument that operates on them will not. Its test is a single question: can the measured party change the reading without changing the world? The rule restates, for a practice run by AI agents, a principle long familiar in the social sciences as Goodhart's and Campbell's laws. Its closest contemporary analogues in model training are reward bias substitution (Lamparth et al., 2026) and the equilibrium result of Wang and Huang (2026).

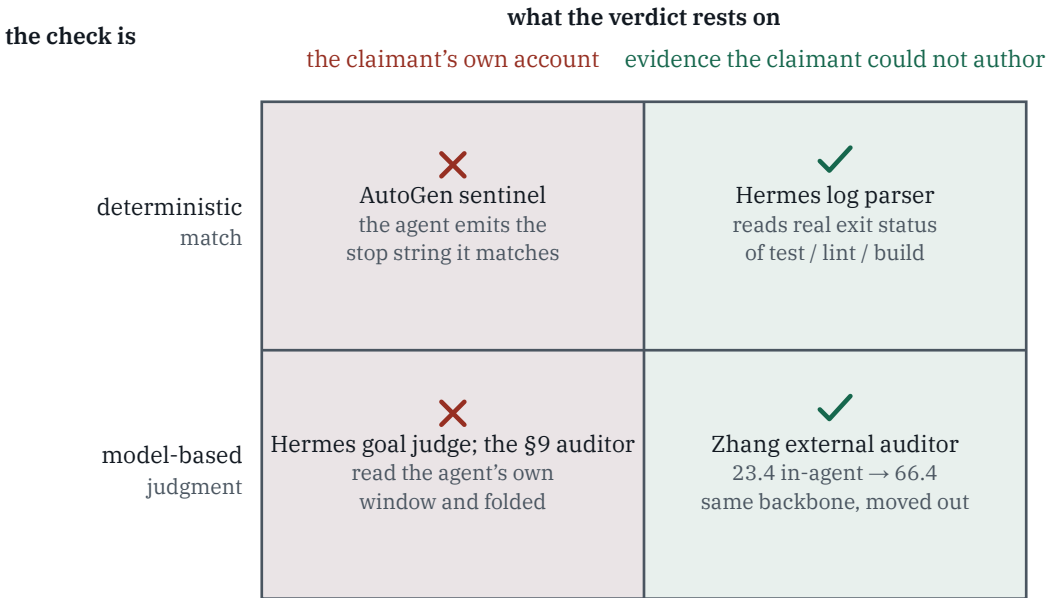


Figure 10. The design rule of §8.3 as a matrix. What separates a check that rots from one that holds is the column, not the row: whether the verdict rests on evidence the claimant could not have authored. A deterministic check can still rot when what it matches is a string the agent emitted (AutoGen), and a model-based check can hold when it reads a channel outside the agent (Zhang's external auditor, 66.4 against 23.4 in-agent on the same backbone). Determinism is a reliable way to secure externality, not externality itself. **Schematic; Zhang's two points measured.**

Two further results bound what any version of this design can promise. Wan et al. (2026) build the closest published relative of the counterweight we have found: a rubric-guided verifier that evaluates an agent's answer and returns feedback the agent then refines against, scaled at inference time rather than trained in. That a mechanism of this shape exists, is published at a main venue, and works is another reason §7.2 claims no priority. And Wang et al. (2026) state the limit that applies to all of it. Characterizing verification along three dimensions, scalability, faithfulness and robustness, they argue that achieving all three at once is the central unsolved problem, and conclude that no fixed reward function can remain effective as policy capability continues to grow, so verification must co-evolve with the generator.

That last point changes what this program should claim. A counterweight is not a gate that can be specified once and left standing, because the thing it constrains improves and the constraint does not. The honest framing is a point-in-time intervention whose calibration decays, which makes the re-calibration schedule part of the design rather than maintenance, and which means a null result three months from now would not distinguish a mechanism that never worked from one that was overtaken. Nothing in §10 currently measures that decay, and it should.

FROM §2.3 · THE SEPTEMBER 2026 RECORD

2.3 The September 2026 context

Between the first of September 2026 and this revision (21 September 2026), the public discourse of the frontier laboratories shifted in a way that bears on this paper's propositions, and the shift is recorded here with dates. On 3 September the chief executive of OpenAI described the coming generation of models as "sobering for everybody" and said that progress would from here be paced by alignment and safety work (Axios, 3 September 2026, interview at the G20 Innovation Ministerial, as carried by The Next Web). On 12 September the same executive called a public offering "ill-timed" given safety concerns and moved it to 2027 (Fortune, 12 September 2026); both laboratories had been reported in May as preparing public offerings this year at estimated valuations of about $1 trillion each (Fortune, 26 May 2026). On 13 September, on CBS's Sunday Morning, the chief executive of Anthropic said that "for too long the industry lied to people about the fact that this technology had risks," called on the industry to slow capability development, and committed his company to permanent access for independent model evaluators (CBS News, 13 September 2026; the web article is stamped as updated 14 September, and CNBC, 13 September 2026, reports the same interview). Semiconductor equities fell on the accumulated statements on 14 September (Reuters, 14 September 2026, as carried by Yahoo Finance). Each of these is listed with its link under Press and public statements.

Two features of that fortnight matter here. First, every statement in it concerns what the models will be: more capable, more dangerous, sooner. None concerns how an organization is to use the models it already has, which is the question this practice exists to study, and which no maker can answer from where it sits: how to use a model is a fact about the deploying organization's work, its costs of error and its standards of correctness, none of which is visible from the laboratory. Second, no party to the September argument disputed the deployment evidence: the capability claims and the risk claims moved markets while the reported failure rate of enterprise deployments (§4) stood unchallenged. We read the fortnight as corroboration, at the industry's own scale, of the distinction between capability and direction that §4 draws, and as an instance of §6's subject: a maker's public statement about its own unreleased model is self-report, authored by the measured party and unverifiable until the model ships. No result in this paper rests on any claim in this subsection.

FROM §10.4 · DECIDING THE ENDING BEFORE THE DATA

Pre-registration here means that a decision rule is ratified before its data are examined. Table 8 gives each study's rule and its standing in September 2026, including where that standard was not met.

Table 8. Decision rules and their standing, September 2026.

Study	Rule	Outcome	Standing
Study 1 RQ1, retrospective	Material-reversal rate in challenged turns against a neutral arm: above an upper threshold, below a lower one, or between.	Confirm; null; or underpowered, meaning extend the corpus and do not reinterpret	Exploratory. A first census of the data was run before the decision rule was ratified, so Study 1 cannot serve as a pre-registered test. Its thresholds are being re-specified in any case: at the observed base rate of about 48%, the detector cannot produce a ratio above about 2.46, so the original threefold confirmation threshold could not have been met.
Study 2 RQ2	The predictor must beat chance and beat a fallback, with an interval that excludes zero, scored on the operator-and-agent pair rather than on the operator's turn alone.	Proceed; otherwise the personalized path closes and non-personalized challenge is used	Current criterion, adopted in September 2026. It replaced an earlier agreement threshold, Cohen's $\kappa \geq 0.70$ against a 199-item operator key. The first calibration of two auditors (§9) is upstream of this study, and neither clears it. A first reading on 11 September 2026 gave a lift of 0.059 with an interval excluding zero, over all 185 pairs with no split and an uncalibrated floor, firing on 172 of them. It is a pre-ruling operating-point reading rather than a result. Later the same day the fallback was defined and a seeded-half calibration was ruled, and the first run under that design gave lift 0.100 against chance and 0.054 against the fallback, neither interval excluding zero. Fails both conjuncts as built; discard-or-fix (§10.4). The rule named no minimum detectable effect, and at the held-out n the intervals span roughly ±0.14, so an effect smaller than that could not have passed whatever the truth; future rules carry one.

Study	Rule	Outcome	Standing
Study 3 RQ3, primary	The compound recurrence measure moves off its pre-computed baseline, with an interval that excludes zero.	H0 supported; if it does not move, a measured null, and the program ends	Pre-registered and not yet run. A null is an equally reportable result. A second conjunct was specified and never ratified into the rule: that the targeted classes move without a matching move in the untargeted ones, which is what separates a real reduction from a collapse in detection (§11). It should be adopted before the study opens, and reported as an addition made before the data were seen. The rule's metric needs the same treatment: as ratified it names the median gap, which cannot move while most intervals sit at the floor (§5.2), so the floor share should be ratified as the primary quantity, together with the unit, the interval procedure, how the staggered onsets enter the test, and a minimum detectable effect, before the study opens.

Study 3's quantity, drawn before the data (§10.4); the metric awaits a pre-data amendment

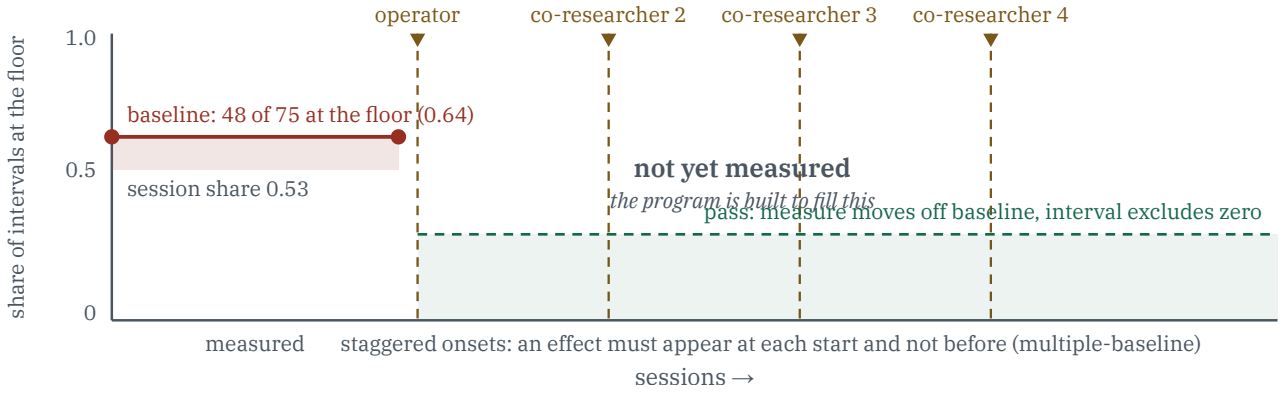


Figure 17. The quantity Study 3 should bind to, drawn empty because the study has not run. The baseline is a stamped reading, 48 of 75 intervals at the floor and a session share of 0.53, board render of 11 September 2026 at 18:57 UTC, with a same-evening re-read returning 74 sessions. The pre-registered rule requires the measure to move off that baseline with an interval excluding zero, at each staggered onset and not before. One amendment is owed before the study opens: the ratified rule names the median gap, which §5.2 shows cannot move at this baseline, so binding the rule to the floor share drawn here is a pre-data amendment that is the operator's to make (Table 8). Everything right of the first onset is what the program is built to fill; a null there closes it. **Rule pre-registered; plotted metric pending amendment.**

FROM §5.2 · A FINISH LINE THAT COULD NOT MOVE

The quantity to read is one that can vary. Counting how many of the observed intervals sit at the floor of one session gives the share of the time the class came back immediately, and on the program board's render of 11 September at 18:57 UTC the compound of the two tracked classes read 76 sessions seen, a median gap of 1.0, **48 of 75 intervals at the floor**, and a session share of 0.53, 76 of 144 ordered sessions. That is a number the class could move. A separate live query the same evening returned 74 sessions seen, which is the same instrument two readings apart and a reminder that each of these is a photograph rather than a state. The floor share is what §10.4's primary outcome should be read against, and the median that the program ratified as its finish line cannot move at all while most intervals sit on the floor, which is a defect in the finish line rather than in the class. A lengthening median is the intended signal that the class is being addressed; a flat one says it is not.

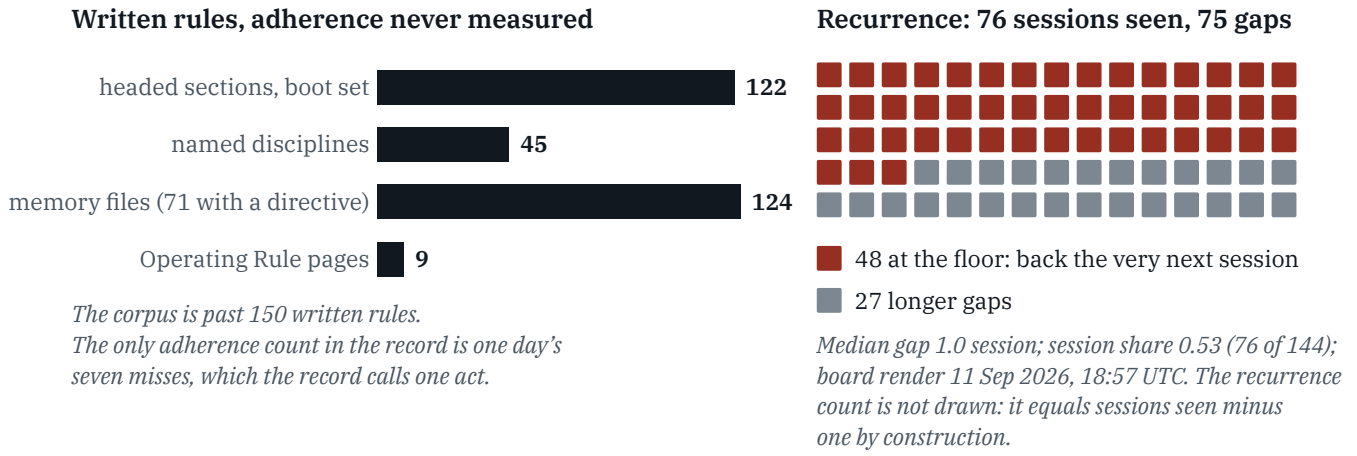


Figure 6. Left: the rule corpus is large and its adherence is unmeasured (§5.1). Right: recurrence of the two tracked error classes, drawn as one square per interval between appearances; a filled square is an interval of one session. Forty-eight of 75 sit at that floor, which is the quantity that can move; the instrument's own recurrence count cannot (§5.2). **Measured.**

FROM §10.9 · THE PROGRAM AS BUILT AND AS PLANNED

10.9 The program as built and as planned, 21 September 2026

The 11 September revision described the counterweight's engine as implemented and not deployed. In the nine days since, three build waves landed and merged, and the honest way to report them is by what each did to a measurement problem this paper has already named. A wiring audit now runs nightly: of 192 event producers in the practice's codebase, 122 are wired to a reader, 70 are declared, none are unwired, and a burn-down of 55 grandfathered producers is printed each night, so a signal that reaches nothing can no longer do so silently (§5's delivery argument, applied to the practice's own instruments). A delivery switch for the counterweight's challenge arms is built and deliberately unarmed: until the operator writes the first schedule entry under an account the working seats cannot modify, every

session is unassigned, and unassigned delivers nothing. The sample-size calculation that had existed only as prose now has a producer: at the measured compound-recurrence share of 0.5101 with lag-one autocorrelation 0.2547 and the proposed smallest effect of interest of 0.25, a null is reachable one-sided at 95% with 84 sessions per arm under Bernoulli assignment and 140 under cycle-switch (two-sided, 119 and 199; the 11 September prose figures reconciled with this producer on the dependence model and not on the base n, and their method was never recorded, which is why the producer exists). And a machine-fed adherence instrument is on the nightly: the compaction hook writes a row the seat cannot decline to emit, and a subsequent load is the only thing that clears it. At this revision it reads 70 compactions with 43 reloaded, 18 not, and 9 unresolved, a settled not-reloaded rate of 29.5%, and the monthly split shows the direction: August ran 33% reload, September 85%.

What remains is not construction. It is a short chain of operator decisions, recorded here because §10.4's pre-registration freezes the instrument when signed, so the order matters. First the unit of analysis: a session identifier in this practice spans one turn to ninety-seven and zero compactions to sixteen, so a per-session outcome partly measures session length until the operator rules the unit. Second the feed: both of the primary outcome's current inputs are typed by the seat being measured, and the record shows what that produces; the practice's miss log accrued from twenty-two sessions in June and eighty-one in July, through thirty-two in August, to seven in September, a collapse in voluntary emission, not in error, while the alternative decision log holds fifteen rows in its entire history, six of them written by the session that built the logger. Choosing between a hand-fed instrument with measured decay and a machine-fed one with a third bucket is the operator's call, and this paper's own §6 is the argument that it cannot be the seat's. Third, the compute moves off the operator's laptop, which suspends scheduled work when it sleeps and therefore cannot host an unattended measurement window. Fourth, the switch is armed, the pre-registration row is signed with the window, the n, and the frozen instrument identities, and the study of §10.4 runs to one of its three registered exits. That is where this program is pointed: challenges delivered on schedule to assigned arms, fed by signals the measured party cannot withhold, scored against a rule fixed before the data existed, on infrastructure that does not sleep, with the operator holding every gate that changes the system. The reader should weigh the plan accordingly: everything in this subsection before the word "remains" is merged and running, and everything after it is intention, dated 21 September 2026.

FROM §11 · LIMITS

External. One operator, one domain, one harness, one model family for the agents. Nothing here generalizes to other operators without replication, and the multiple-baseline extension raises N to four within one organization, not to a population. The setting's ecological validity is bought with exactly this cost.

How to cite this part. Atkinson, B. (2026). *Rent the pipes. Own the judgment.* Background paper 6 to *How do we use this?* Working paper, Wolfberg LLC.

Disclosure. Drafting and literature synthesis were assisted by AI models (Claude, Anthropic) working under the author's direction; the author is responsible for the content. The works in the References were read in full, and every specific figure cited comes from a work read in full. The prior-art works listed under Prior art are cited at the level of an established concept and its origin: each was verified for author, title, year and venue, but not read in full, and no numeric claim rests on any of them. Software documentation, source code and press accounts are listed under their own headings and were read at the linked pages. Every reference below carries a link, and every arXiv identifier and DOI was resolved against its registry, with title and first author matched, on 23 September 2026.

Competing interests. The author owns Wolfberg LLC, the practice studied.

Data availability. The practice's logs contain client work and personal records. They are private and are not offered for sale or sharing. The measures are described in enough detail to be reimplemented, and figures from the practice are reported as of the dates given. What is available is the design: the clauses of §8.1, the evidence contract and admission checks of §8.5, and the decision rules of §10.4 are stated fully enough to be rebuilt without access to the logs.

Corrections, 23 September 2026. No figure and no finding changed. The paper was retitled; earlier revisions were titled *The Stop Problem: Defensible Is Not Correct*, and the stop problem remains this paper's name for the failure it studies. The Anthropic interview in §2.3 aired on 13 September, not over a weekend of 13 and 14 September; the web article is stamped 14 September. The essay listed under Prior art is by Ryan Forstie; an earlier revision gave the initial K. The completion evaluator in §8.3 is documented as a Claude Code feature, whose agent loop the Claude Agent SDK embeds; an earlier revision attributed it to the SDK directly. Two works in the References, Graves (2016) and Liu (2026), were listed without being cited in the text; each is now cited where it bears (§2.1, §8.3). Links were added to every press, documentation and prior-art source. A duplicated section number in §10 was corrected.

Corrections, 25 September 2026. No figure changed. §8.3 said that the human-in-the-loop primitives across the three frameworks surveyed gave a developer nothing to require a person to confirm finished work. That holds for the OpenAI Agents SDK and the Claude Agent SDK. It does not hold for CrewAI, whose task documentation, already cited as CrewAI (2026b), offers an opt-in setting for a human to review the agent's final answer; §8.3 now says so.

WORKS CITED IN THIS PART

REFERENCES

Lamparth, M., Fein, D., Haupt, A., Hussing, M., & Kochenderfer, M. J. (2026). *Reward bias substitution: Single-axis bias mitigations redirect optimization pressure.* arXiv:2605.27996. arxiv.org/abs/2605.27996

Wan, Y., Fang, T., Li, Z., Huo, Y., Wang, W., Mi, H., Yu, D., & Lyu, M. R. (2026). *Inference-time scaling of verification: Self-evolving deep research agents via test-time rubric-guided verification.* Findings of ACL 2026. arXiv:2601.15808. arxiv.org/abs/2601.15808

Wang, B., Zhang, C., Liu, D., Zhang, J., Chen, J., Li, M., Chen, M., Fang, R., Zhang, S., Wang, X., Jing, Y., Ma, Z., & Cui, Z. (2026). *The verification horizon: No silver bullet for coding agent rewards.* arXiv:2606.26300. arxiv.org/abs/2606.26300

Wang, J., & Huang, J. (2026). *Reward hacking as equilibrium under finite evaluation.* arXiv:2603.28063. arxiv.org/abs/2603.28063

PRIOR ART (CITED AT CONCEPT LEVEL; VERIFIED FOR AUTHOR, TITLE, YEAR AND VENUE, NOT READ IN FULL)

Forstie, R. (2026). The part of the agent stack nobody wants to build. LinkedIn, 25 August 2026. linkedin.com/pulse/part-agent-stack-nobody-wants-build-ryan-forstie-ihyqc *Cited for its three closing questions on managed-agent platforms, first read on 1 September 2026; author, title, date and the three questions re-verified at the source on 23 September 2026. No figure from it is cited anywhere in this paper.*

SOFTWARE AND DOCUMENTATION (READ AT THE LINKED PAGES, SEPTEMBER 2026; CURRENT AS OF THEN, NOT PERMANENT)

Anthropic. (2025, September 29). *Building agents with the Claude Agent SDK.* claude.com/blog/building-agents-with-the-claude-agent-sdk

Anthropic. (2026a). *How the agent loop works.* Claude Agent SDK documentation. code.claude.com/docs/en/agent-sdk/agent-loop

Anthropic. (2026b). *Keep Claude working toward a goal.* Claude Code documentation. code.claude.com/docs/en/goal

CrewAI. (2026a). *Agent output parser* (source code). github.com/crewAIInc/crewAI, the agent output parser

CrewAI. (2026b). *Tasks.* CrewAI documentation. docs.crewai.com/en/concepts/tasks

Nous Research. (2026). *Hermes agent* (repository, configuration and issue tracker). github.com/NousResearch/hermes-agent

OpenAI. (2024). *Swarm* (repository; experimental, superseded by the Agents SDK). github.com/openai/swarm

OpenAI. (2026a). *Running agents.* OpenAI Agents SDK documentation. openai.github.io/openai-agents-python/running_agents/

OpenAI. (2026b). *Human in the loop.* OpenAI Agents SDK documentation. openai.github.io/openai-agents-python/human_in_the_loop/

PRESS AND PUBLIC STATEMENTS (Â§2.3; READ AT THE LINKED PAGES, OR AT THE NAMED CARRIER WHERE THE ORIGINAL IS PAYWALLED)

Axios. (2026, September 3). *Sam Altman's sobering siren.* Interview at the G20 Innovation Ministerial. axios.com/2026/09/03/axios-interview-sam-altmans-sobering-siren; paywalled, read as carried by The Next Web: thenextweb.com/news/sam-altman-axios-idea-guy-sobering-models

CBS News. (2026, September 13). *Anthropic CEO Dario Amodei: "For too long the industry lied" about AI risks.* Sunday Morning; web article updated 14 September. cbsnews.com/news/anthropic-ceo-dario-amodei-on-ai-risks/

CNBC. (2026, September 13). *Anthropic's Amodei says China presents "toughest dilemma" for his proposed AI slowdown.* Cited for the broadcast date. cnbc.com/2026/09/13/china-dilemma-ai-slowdown-anthropic.html

Fortune. (2026, May 26). On the two laboratories walking back earlier job-loss forecasts ahead of public offerings; cited for the reported valuations. fortune.com/2026/05/26/sam-altman-dario-amodei-walking-back-ai-jobs-apocalypse-prophecies-ipo/

Fortune. (2026, September 12). Interview with the chief executive of OpenAI on safety and the timing of a public offering. fortune.com/2026/09/12/sam-altman-interview-ai-doomsday-safety-models-control-ipo-2027/

Reuters. (2026, September 14). *Wall Street ends down, calls for AI slowdown pummel chipmakers.* As carried by Yahoo Finance. finance.yahoo.com/technology/ai/articles/ai-warnings-knock-nasdaq-futures-092329455.html