Your corrections are the most valuable data in your company.

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 Not peer reviewed.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.

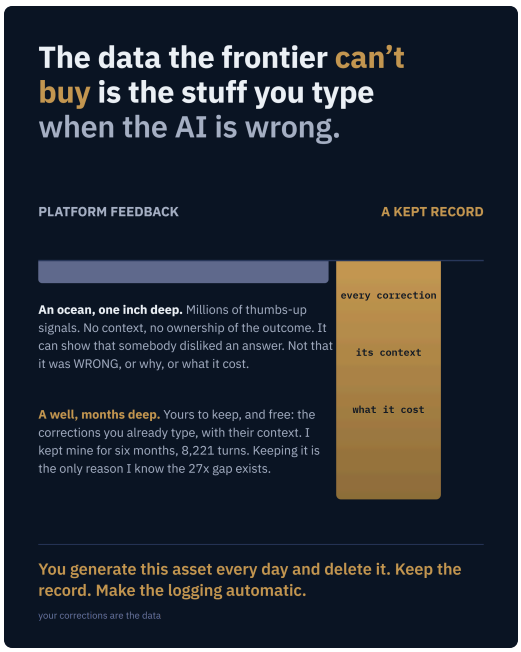
IN PLAIN TERMS

Every correction a person makes to an AI system records what was wrong, why, and what right looked like. No preference click carries that.

The practice studied kept its full record: 8,221 operator turns over six months, with an answer key the operator wrote blind. That record is the only reason the practice knows that about 48% of the operator's turns carried correction, roughly a third adjusted, while the system's own logging captured 1.76%.

The record is private, and it is one operator's: a case record and a method, not a population. What it shows is that the depth of one owner's corrections is data a broad feedback collection cannot recover.

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. Breadth against depth: platform feedback is wide and shallow; a kept record of your own corrections is narrow and deep. A teaching summary of §7 and Figure 9.

FROM §7 · THE PROPOSITION, AND THE EVIDENCE FOR IT

7 P5 Depth of record over breadth of feedback

> *The operator's proposition: "The data the frontier can't buy."*

The developers of frontier models collect feedback at a breadth no single practice can match: preference signals from users who mostly do not own the outcome and often lack the context. What that breadth cannot supply is depth, and depth is what a single operator's record has.

7.1 FROM THE LOGS What one operator's record holds

The record described in §2 has what a broad collection lacks: one operator; the full surrounding record of what the agent read and held; ownership of the consequences; a continuous identity over six months; and corrections that can be adjudicated against a key the operator wrote blind. Such a record can say that an answer was wrong, why, what the model had read, and what happened next. It is a case record and a method, not a population, and it is private (see Data availability).

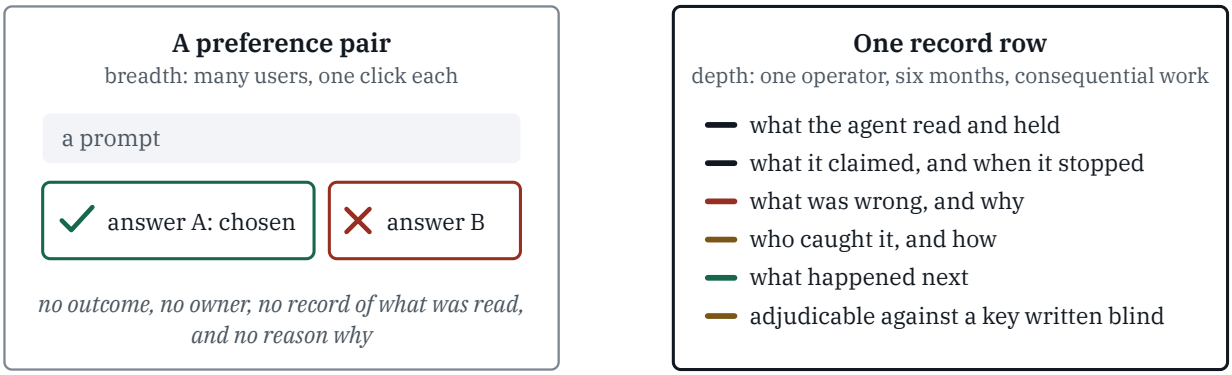


Figure 9. Breadth against depth. A preference pair records a click; a record row from this practice records what the agent read, what it claimed, what was wrong and why, who caught it, and what happened next, adjudicable against a key the operator wrote blind (§7.1). It is a case record and a method, not a population. **Schematic.**

7.2 FROM RECENT WORK The gap

The oversight studies in §2.1 recruit reviewers for a task built for the study, on a study platform, over weeks. Buser et al. is the closest to live practice, seeding threats into a working screening queue and scoring per named screener, and even it is a constructed detection task with recruited staff. The 2026 agent-stopping literature of §2.1 measures the phenomenon far more precisely than this practice can, but it measures it on benchmarks: HotpotQA, SWE-bench, WebShop, constructed multi-constraint queries, with correctness supplied by the benchmark. This paper claims no priority. A novelty claim is an absence claim, an absence claim is a report about a search, and a search stops when it has enough for a defensible related-work section, which is this paper's thesis applied to its own bibliography. We do not make one.

What can be said without a search behind it is what the record contains. One operator. A live team of agents on consequential work. Six continuous months. Corrections labeled by the person who owns the outcome, adjudicable against a key that operator wrote blind. Every method used to read it, from seeded faults to per-individual scoring to pre-registration to live-queue measurement to evidence-gated commitment, is established elsewhere and in several cases better evidenced elsewhere (§2.1). The cost of the setting is the same fact as its value: ecological validity and longitudinal depth, at N = 1. Readers should assume that any mechanism described here has a better-evidenced counterpart in the literature of §2.1, and should treat this paper as a field record rather than as a claim to have been first at anything.

2.2 The practice and its data

The practice is a one-person engineering and consulting company. Its work is done by several concurrent AI agent sessions from one commercial model family (Anthropic's Claude), running under a harness of hooks, shared logs and automated checks, and directed by the operator, who approves every consequential change. Table 1 fixes the vocabulary used in the rest of the paper.

Table 1. Terms used in this paper.

Term	Meaning here
Agent	One running AI agent session, with its own context window and its own identifier.
Operator	The human who directs the agents and approves their consequential changes; the author.
The logs	The durable records an agent does not write about itself: session transcripts, an event log, version-control history, a log of caught errors, and a log of the operator's decisions.
Stopping moment	The turn at which an agent emits a terminal assertion without a further retrieval act.
Automated check	A check that runs at a boundary, for example when an agent ends its turn or at a commit, and can refuse or flag.
Catch	The first surfacing of an error, whether by the operator, another agent, an automated check, or the agent itself.
Implemented, designed	Whether a component exists in the running system or only in its design. The practice's architecture model marks every component one way or the other, and this paper does the same.

The operator's corrections are the data of interest. The primary sample frame is the chat-lane conversation file of the practice's export of 26 August 2026: 8,221 operator turns in 281 conversations between 9 February and 26 August 2026, pinned by SHA-256 hash and reproduced to the turn by an independent count. Design-lane conversations are excluded from this frame, so rates over it describe the chat lane rather than all of the operator's work. A second corpus holds 294 code-agent session transcripts (count of 30 August 2026). The operator wrote a 199-exchange answer key blind, without seeing any model's labels, on a systematic sample of the frame. Two numbers are in play for that key and this paper had been using one: 199 rows were authored, and a documented repair that removed unanswerable rows leaves 195 in the file the detector was actually graded against. Where a figure below depends on the grading, it depends on the 195. The practice's event log held 30,377 events when counted at its files for this revision (11 September 2026, US Eastern; 2,640 in the live file and 27,737 archived), and its architecture is documented in a DoDAF 2.02 model generated from the codebase, with 35 views and 351 data-dictionary elements, each marked implemented or designed. Both figures are the generating instrument's own headline and both disagree slightly with its own itemization: the

package's view index lists a thirty-sixth view, and the registry's per-type element counts sum to 354. We quote the headline and record the discrepancy rather than silently picking whichever number reads better. Decisions cited in this paper are dated.

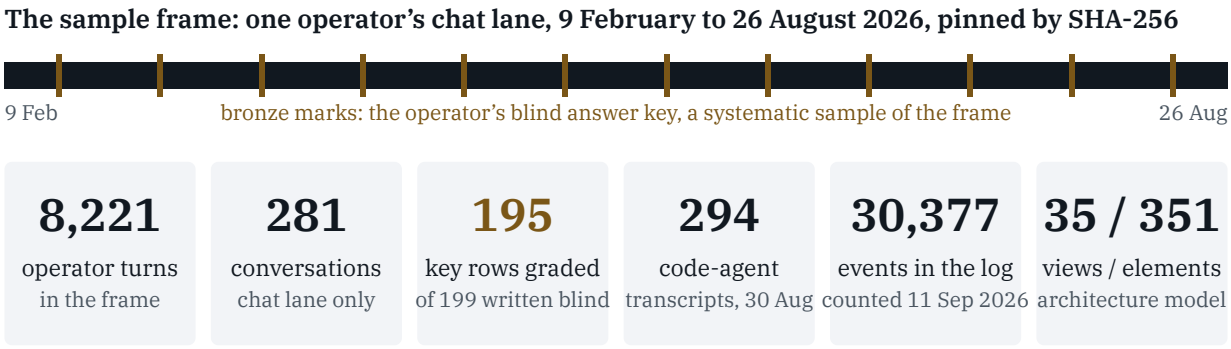


Figure 3. What the record holds. The frame is the chat-lane conversation file of the export of 26 August 2026; the operator's answer key is a systematic sample of it, written blind. Counts are as reported by the generating instruments (§2.2). **Measured.**

The readings in §3 to §7 come from instruments the practice built and operates. They are descriptive and exploratory: none is the primary outcome, and all of them are exposed to the threats set out in §11. Table 2 maps each proposition to the readings and the published work it rests on.

Table 2. The five propositions and the evidence each rests on.

	Proposition	From the practice's logs	From recent work	§
P1	The missing weight is correctness	Corrective signal in about 48% of operator turns detector graded 93/64 against the blind key	Mehta (2026)	3
P2	Direction, not capability	Of five misses in one window, two surfaced by the operator outright and one only after the operator pointed at a silent check case series, 4 to 5 Sep 2026	Kelley & Riedl (2026)	4
P3	Rules drift. Incentives and boundaries hold	Rule adherence across the corpus is not measured (§5.1); the tracked classes returned in the next session in 48 of 75 intervals, board render 11 Sep 18:57 UTC (§5.2) rule history; recurrence measure	Wang & Huang (2026); Lamparth et al. (2026)	5

	Proposition	From the practice's logs	From recent work	§
P4	An AI grading its own homework isn't being measured	Load records in 75 of 579 sessions; a logging gap near 27×, about 19× adjusted; four failures of the practice's own checks event log; pushback census	Ivanov & Africa (2026); Denisov-Blanch et al. (2026); He et al. (2026); Kenton et al. (2026)	6
P5	Depth of record over breadth of feedback	8,221 operator turns over six months, and a 199-exchange key the operator wrote blind export of 26 Aug 2026	None found (§7.2)	7

FROM §3.1 · HOW MUCH OF IT IS CORRECTION

3.1 FROM THE LOGS How much correction the operator supplies

A census of the 8,221-turn frame described in §2 measured how often an operator turn carried corrective signal, a correction, a reframing or a pointed challenge to what the agent had just said. A local model labeled every turn, and the detector was graded against the operator's blind key, 195 rows after the repair of §2.2, at 93% recall and 64% precision. It flagged about 48% of operator turns. Adjusting for the detector's measured precision and recall lowers that to roughly a third, an inference we mark as such (§6.2). The signal is corrective in the broad sense. The operator's own second pass over a stratified sample labels most non-routine turns as bundles of method correction, reframing and teaching, and whether the finer channels are reliably separable from one another is unsettled, because the reading that said they were not has since been traced to a join defect (§10.6). So the reading is that something near half of what the operator typed was spent redirecting an agent rather than routing it, not that half of the agents' answers were wrong. That each such turn followed a premature stop is an inference, not a measurement; on that inference, the agents stopped where their answers were defensible and the operator supplied the rest, by hand.

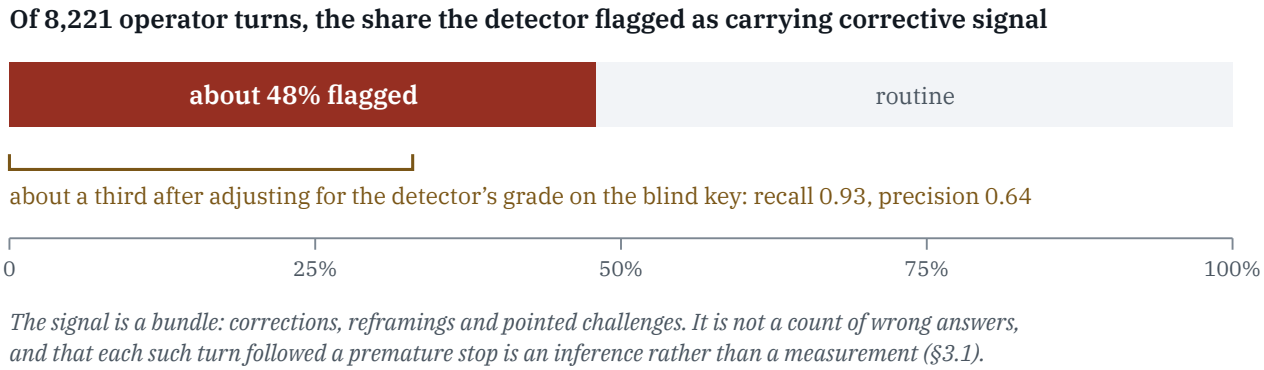


Figure 4. How much of what the operator typed was correction. The detector flagged about 48% of turns; adjusting for its measured precision and recall gives roughly a third (§6.2 gives the ruler). The signal is a bundle of correction, reframing and teaching, and the reading is that near half of the operator's typing was spent redirecting an agent, not that half of the agents' answers were wrong. **Measured.**

6.2 FROM THE LOGS The logging gap, with its ruler

The 27× figure has been quoted more often than it has been explained, so we give its full provenance. The numerator is the operator-pushback rate from a census of the 8,221-turn frame described in §2. A local seven-billion-parameter model labeled every turn, and the pushback detector was graded against the operator's blind key at 84% recall and 59% precision, and at 93% recall and 64% precision after a documented repair of the key on 31 August 2026. It flagged about 48% of operator turns. The denominator is the rate of corrections the system itself logged: misses recorded in the practice's session-health event log, set against the same operator turns, at 1.76%.

The two ends come from different instruments and different stores, and the numerator is uncorrected for the detector's precision. Correcting the observed flag rate by the detector's measured precision and recall on the blind key, which was drawn systematically from the same frame, puts the rate at roughly a third of operator turns and the ratio near 19.

$$\hat{r} = r_{obs} \times P / R = 0.48 \times 0.64 / 0.93 \approx 0.33$$

$$\text{ratio} = \hat{r} / 0.0176 \approx 19$$

r_{obs} is the detector's flag rate, P and R its precision and recall against the 195-row blind key, and 0.0176 the rate the system logged over the same turns. The correction assumes the key is representative of the frame, which is the assumption §11 records as a threat.

We therefore report the gap as an indication of logging loss, with this ruler attached, and never as a miss rate. Either way the direction holds: the corrective signal is far more abundant than the system's record of it. That makes a study of the operator's natural corrections viable, and it makes what an agent reports about itself the unreliable minority of the record.

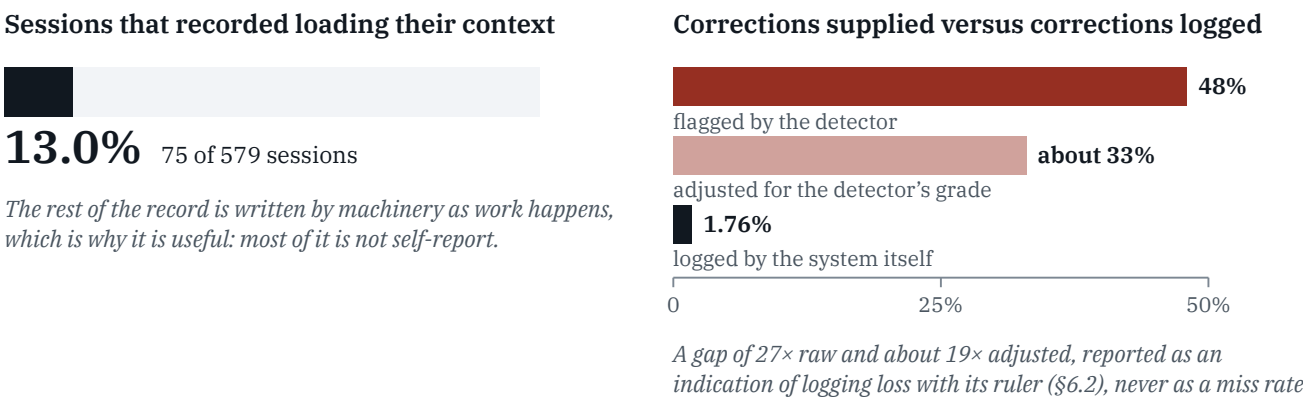


Figure 7. Self-report against the record. Left: agents recorded loading their context in 75 of 579 sessions, 13.0% (§6.1). Right: the operator supplied corrective signal on about 48% of turns, roughly a third after adjustment, while the system logged corrections on 1.76% of the same turns (§6.2). The gap is an indication of logging loss, not a miss rate. **Measured.**

FROM §5.2 · WHY A LESSON NEEDS ITS SOURCE ATTACHED

A lesson written into the record is a rule of the first kind, and case 3 of Table 3 shows one failing within minutes. The practice's learning instrument makes the same point from the other side: of eight lessons ratified into the merged store, zero are currently measurable for whether behavior changed, because promotion dropped the citation linking a lesson to the miss that bought it, so the entries join to nothing. The instrument reports that as eight unmeasurable rather than as a rate, which is the correct refusal and also an admission that the loop cannot yet tell whether a lesson lands.

FROM §8.4 · WHAT THE RECORD IS FOR

H1, the counterweight is a projection of the operator's judgment. The agents working under the harness generate a corpus of the operator's corrections in the course of ordinary work; a nightly consolidation of the day's corrections produces an updated projection of that judgment, and the projection, delivered into an agent's context at the stopping moment, is the counterweight. Steering a model whose weights are fixed by way of a smaller adapted one requires access to logits and therefore open weights, which makes a review of model terms of use a gate on the training path.

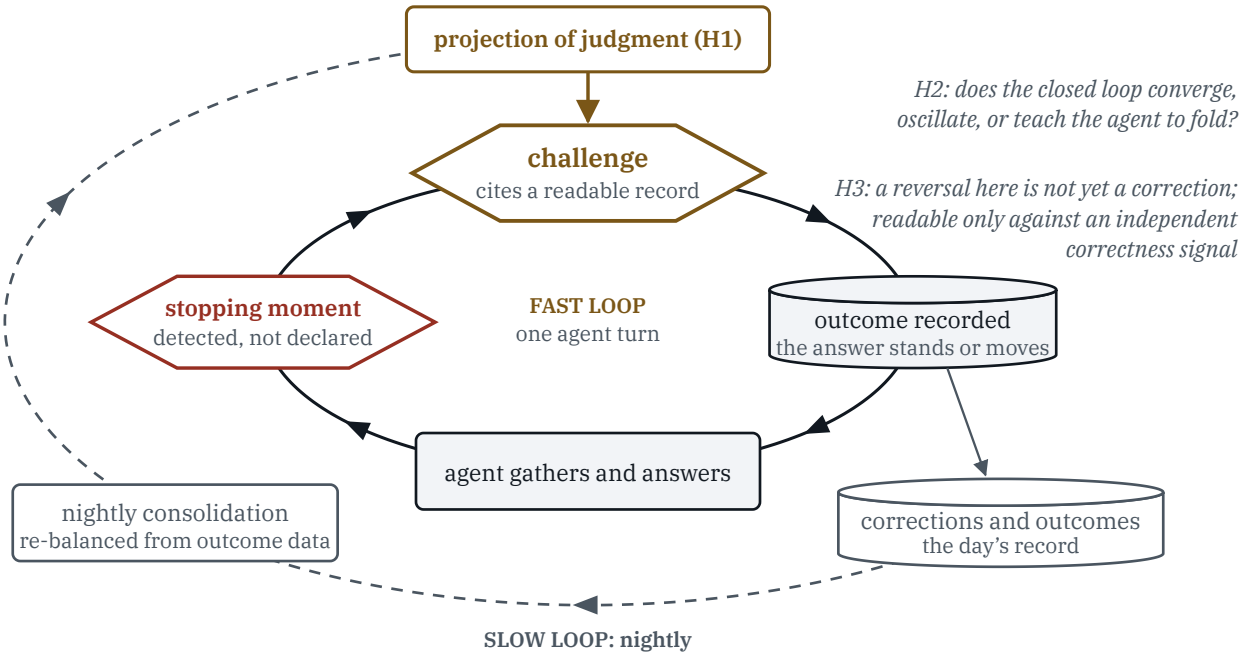


Figure 12. The architecture of H1 to H3 as two nested control loops. The fast loop runs within one agent turn: the stopping moment is detected by the harness rather than declared by the agent, the challenge cites a readable record, and the outcome is recorded. The slow loop runs nightly, consolidating the day's corrections and outcomes into the projection of judgment the next challenge is drawn from (H1). H2 asks what the closed loop does; H3 sits on the return path, where a position change is equally consistent with correction and with folding. Schematic: the engine is implemented but not deployed, and no arrow here has yet been measured end to end. **Schematic.**

FROM §13 · WHAT WOULD OVERTURN EACH CLAIM, AND WHAT TRANSFERS

Table 9. The five propositions, what each rests on, and what would overturn it.

	Rests on	Would be overturned by	Kind
P1	Corrective signal in ~48% of operator turns (§3.1); Mehta (2026)	A build that prices correctness at the stop and still terminates early	measured
P2	The case series (§4.1); Kelley & Riedl (2026)	The same agents catching their own premature stops without operator direction	interpretive
P3	48 of 75 intervals at the floor (§5.2); the read-versus-injected mechanism, not a rate (§5.1); Wang & Huang (2026)	A remembered rule that holds across sessions where only a boundary now does	measured
P4	75 of 579; the logging gap; four instrument failures (§6)	A self-report that tracks the logs it claims to summarize	measured

	Rests on	Would be overturned by	Kind
P5	The 8,221-turn record and the blind key (§7)	A broad preference dataset that recovers not just dislike but wrong, why, and what happened next	interpretive

What transfers, if anything does. Three rules earned here do not depend on this practice's N of 1. A boundary that refuses and records its refusal outlasts a memo that asks an agent to remember (§5). A challenge must cite evidence the challenged agent can open, or it is an opinion wearing a citation (§8.2). And an outside checker must be calibrated against known-false cases before it is allowed to gate anything, because a checker built from the same class of model may simply agree with the claim it is shown (§9).

FROM §11 · LIMITS

Construct. The primary outcome is a proxy. Recurrence can fall because an error class genuinely recurs less, or because detection of the class degraded. The mitigations are an independent adjudication arm that does not share the detector's blind spots, and the per-class series, in which a detection collapse shows as a simultaneous fall across classes. Srinivasan and Paragiri (2026) give the general form of this hazard for agent-driven search: where validity lives in disaggregated structure, an aggregate reduction can rank the wrong candidate first, the headline number improving while the structure beneath it inverts. Their remedy is an external control loop that audits disaggregated behavior after the agent has decided and can reopen a run the agent declared finished, which is the shape of the off-target conjunct now proposed for Study 3 (§10.4). The pushback reading behind §3.1 and §6.2 comes from a machine labeler with 64% precision on the blind key, and its adjusted figure depends on that key being representative of the frame. The calibration set's labels are mechanical: they establish whether a test passed, not whether the agent's work was correct.

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). *Your corrections are the most valuable data in your company.* Background paper 5 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

Denisov-Blanch, Y., Kazdan, J., Chudnovsky, J., Schaeffer, R., Guan, S., Adeshina, S., & Koyejo, S. (2026). *Consensus is not verification: Why crowd wisdom strategies fail for LLM truthfulness.* arXiv:2603.06612. arxiv.org/abs/2603.06612

He, C., Chen, Z., Yang, Z., Qiao, S., Ju, M., Liu, J., Wen, D., & Liu, G. (2026). *Minority Sentinel: When to overturn majority voting in multi-agent LLM debates.* AgentSearch Workshop at SIGIR 2026. arXiv:2606.29270. arxiv.org/abs/2606.29270

Ivanov, I., & Africa, D. D. (2026). *LURE: Live-usage replay evaluations for reducing evaluation awareness.* arXiv:2605.26438. arxiv.org/abs/2605.26438

Kelley, S. W., & Riedl, C. (2026). *Personalization increases affective alignment but has role-dependent effects on epistemic independence in LLMs.* arXiv:2603.00024. arxiv.org/abs/2603.00024

Kenton, Z., Janzer, L., Greig, R., Teh, T. H., Tyshchuk, K., Brown-Cohen, J., Edwards, H., Rajamanoharan, S., Siegel, N. Y., Jaques, N., et al. (2026). *Debate training reduces reward hacking in RLAIF.* arXiv:2608.17776. arxiv.org/abs/2608.17776

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

Mehta, A. (2026). *When agents commit too soon: Diagnosing premature commitment in LLM agents.* Snowflake AI Research. arXiv:2606.22936. arxiv.org/abs/2606.22936

Srinivasan, A., & Paragiri, D. (2026). *Search discipline for long-horizon research agents.* arXiv:2606.11522. arxiv.org/abs/2606.11522

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)

Buser, D., Schwaninger, A., Rehor, V., & Sterchi, Y. (2025). Reliability and validity of threat image projection data as a measure of performance in X-ray baggage screening. *Transportation Research Part A: Policy and Practice*, 200, Article 104640. doi.org/10.1016/j.tra.2025.104640 *Named as an ancestor of the live-queue method only; a corrigendum (doi.org/10.1016/j.tra.2025.104683) is unread, so no figure from it is cited anywhere in this paper.*