In the Loop or Watching the Loop? The Human Gate Is a Rater Too

A micro1 working paper on rater effects, drift and the shelf life of an AI evaluation, the fine print on its five-week number, what an independent recomputation found, and what the paper's three checks imply for the human who approves AI work. Is that human in the loop, or only watching it?

Berg Atkinson, Wolfberg LLC · berg@wolfberg.ai

Preprint. Not peer reviewed. Published 16 October 2026.

Version 3.0, 20 October 2026. Revised 2026-10-20: impersonal register, author block removed, figures redrawn. Numbers, table values and status labels are unchanged from version 2.2 (17 October 2026), which put the October share on attributed rows (43 of 151, 28.5%). Earlier versions: 2.1 (17 October), 2.0 (first person, 16 October) and 1 (third person). Ages and counts are as of the original publication date, 16 October 2026; no source was re-checked for this revision.

Version 3.1, 30 October 2026: set in the house research-note format; the figures are shown on house plates and called exhibits. Text and numbers are otherwise unchanged from Version 3.0.

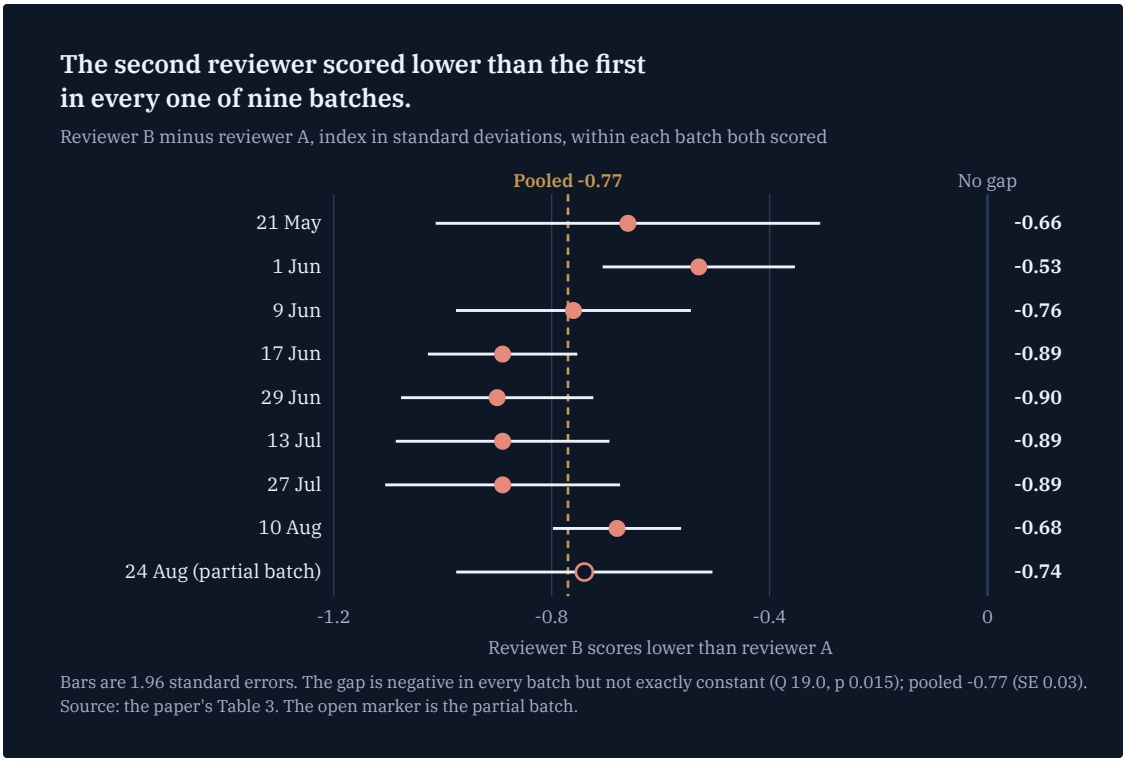


Exhibit 1. The reviewer effect. Within each batch that both reviewers scored, reviewer B scored below reviewer A. The two reviewers scored different interviews from the same batches (Section 4.1), so the gap is a reading of reviewer severity only if the interviews were split without regard to difficulty. **From the paper's Table 3; bars are 1.96 standard errors, computed by the analysis seat (a Claude instance assigned to recomputation; see Section 6). The paper is a working paper by micro1's own team about micro1's own agent (Section 4.3).**

IN PLAIN TERMS

micro1 recruits and vets people for AI-training work. It ran an AI interviewer named Zara on real job interviews, had human reviewers score recordings of them, and posted a working paper on 4 October 2026 and a landing page on 14 October asking a plain question: can a company tell whether its AI agent is getting better? Across 2,611 scored interviews in 14 batches (26 March to 24 August) the authors report three things. A new reviewer who scored harsher made the agent look flat when it was improving: reviewer B scored 0.79 standard deviations below reviewer A, and after adjusting for who scored what, the agent's index rose by 0.53. The agent's measured reliability also moved between logged releases, so an evaluation goes stale: after about five weeks by the authors' point estimate, and anywhere from 1 to 21 across its interval. And interview-related support tickets fell about 10% a month relative to technical-issue tickets.

The analysis seat read the paper and the landing page, checked the design against the tables, and recomputed the arithmetic. The numbers reproduce. The fine print matters more than the arithmetic. No interview was scored by both reviewers, so the headline gap rests on an assumption about how interviews were split. The five-week figure rests entirely on ONE step in late May that the change log does not explain: free that step and the estimated drift for the index is 0.000, though the interval still runs to 0.151, so the data cannot say drift is zero. The agent is micro1's own product, one author is micro1's Chief Economist, and micro1 also sells evaluation products to AI labs. In the paper's favor, the improvement appears in reviewer A's scores alone, with no adjustment at all, and the authors say plainly that their data cannot show the evaluation caused anything.

The paper's three checks were then turned on Wolfberg's own practice, which is where the note has the most at stake. In a full read of the practice's transcripts (8 to 16 October), run by Wolfberg's agents, one model read 36,657 items; 3,163 were exact duplicates and were not read again, and each of the other 33,494 was read twice. The two passes disagreed on 5,126 of them (15.3%), and on 14.5% of the items where both passes came from the main model (a fallback model read some items after refusals). The model and a human hand check agreed on "did the seat fold?" at a kappa of 0.157 on 23 episodes, while the model's own two passes agreed at 84.7%. In one window of about an hour and forty minutes the read held 15 items from the human's side and flagged none as pushback; on viewing them, the operator's reaction was recorded (the operator, 2026-10-16: "theres a lot of pushback there"). Wolfberg's public miss register shows the operator's share of the catches (the register's label for the human who runs the practice) falling from 71% in June to 28% in early October. That is the paper's reviewer-B problem in Wolfberg's own data: the share of catches credited to each catcher moved, and the data cannot say whether that reflects the team, who was looking, or how catches were logged, so the monthly counts cannot say whether the team got better.

The same three checks bear on a question put to the analysis seat on 16 October: how can it be measured whether a human is in the loop or only watching it? The human who approves AI work is a rater too, with a severity, a drift and nobody re-scoring them. A human is in the loop when their decision changes outcomes before they take effect, because the item was wrong; watching means approving at the same rate whether it is wrong or right. The core measure compares the catch rate on wrong items with the flag rate on correct ones, and it needs the wrong items the human missed. Escapes plus a random re-check of signed-off work (for example one in ten) were chosen over planted items, and the unit is the pair, not the human alone. A twelve-row seed ledger from one session shows the fields; it is an illustration, not a rate, and none of this is running yet.

Position: the paper's three checks (rater comparability, evidence currency under drift, corroboration with an outside outcome) are the right ones, and its shortest rule is the strongest: a judge version is a rater. The five-week number is best treated as an illustration, which its own landing page nearly says. What deserves copying is the method of measuring one's own drift rate, and the habit of overlapping the people or models doing the scoring.

ABSTRACT

Zhang and Esposito (2026) study 2,611 rubric-scored production interviews from micro1's voice-and-video interviewing agent, scored in 14 batches between 26 March and 24 August 2026 by up to three human reviewers. They report that reviewer B scored 0.792 standard deviations below reviewer A on a composite index; that a roster change flattened the raw trend while the reviewer-adjusted index rose 0.53 standard deviations; that critical failures peaked at 6.3% and fell to 2.5%; that unlogged movement has a weekly standard deviation of 0.086 (95% interval 0.041 to 0.180), putting the horizon at which a reading cannot detect a deploy-sized change at about five weeks (1 to 21 across the interval); and that interview-related support tickets fell about 10% a month relative to technical tickets. An analysis seat checked each against the paper's tables and recomputed its horizon formula and breakeven table, which reproduce (5.03 weeks at the point estimate; 22.6 and 1.1 at the interval ends; 18% credit needed at $100 per escalation). Five points of fine print qualify the headline. Each recording carries one score, so the reviewer gap is identified by assumed exchangeable assignment within batches, not by double scoring. With the state also freed across a late-May step that the change log does not explain, the index's weekly drift estimate is 0.000 (0.000 to 0.151, p = 0.50), so the point estimate rests entirely on one step; the interval's shortest horizon moves from 1.1 to 1.6 weeks. The agent is micro1's own product, one author is micro1's Chief Economist (the landing page gives the affiliations; the PDF carries none), and micro1 also sells evaluation products to AI labs. In the paper's favor, the reviewer-A-only trend (0.124 SD a month) exceeds the adjusted trend (0.102), and the authors disclaim any causal effect of the evaluation program. The three checks were then applied to Wolfberg's own practice. A model labeler read 36,657 transcript items; 3,163

were exact duplicates and the other 33,494 were each read twice, producing 42,288 and 42,236 records, of which 16,732 matched on exact anchor text, and 5,126 items (15.3%) differed (14.5% where both passes came from the main model). On the fold question the model's two passes agreed 84.7% (kappa 0.77, 523 episodes) while the model against a human hand check was kappa 0.157 (23 episodes); a different question, the kind of pushback, agreed 94.5% (kappa 0.92, 777 episodes). The read's pushback flag was right 17 of 19 times, with recall about 42% (rough range 25 to 65%, from nine unflagged messages). The operator's share of the public miss register's catches fell from 71.1% in June to 28.5% in early October, with the data unable to separate the team, who was looking and how catches were logged. The question is then asked of the human gate: a human is in the loop when their decision changes outcomes before they take effect because the item was wrong, measured by the catch rate on wrong items against the flag rate on correct ones, then by timing (before or after ship) and, as supporting evidence only, engagement. The denominator is the wrong items the human did not catch; escapes plus a random re-check of signed-off work (for example one in ten) are chosen over planted items, the unit is the pair (both the AI seat and the human), and a twelve-row seed ledger from one session (origin: AI seat 11, both 1, human alone 0) illustrates the fields without giving a rate. None of it is running yet. The note concludes that the paper's rules transfer, that its five-week number does not, and that a program should overlap its raters at every batch, date every claim, and record which finding led to which fix.

Keywords: AI agent evaluation, rater effects, LLM-as-judge, drift, evidence currency, measurement, human in the loop, human oversight, escapes, miss register

CONTENTS

1 Introduction: where things stand
2 Position
3 What the paper shows
4 The fine print
5 What the recomputation reproduced
6 The three checks applied to Wolfberg's practice
7 In the loop, or watching the loop?
8 What it means
9 Limits
References
Appendix A. Receipts and internal labels

1 Introduction: where things stand

Newest first, because in this field a source's age matters. Ages are days before 16 October 2026. In this note a "seat" is one Claude instance with an assigned role, and "the operator" is the register's label for the human who runs Wolfberg.

- **16 October (0 days).** The question of how to measure whether the human is in the loop or watching it was put to the analysis seat, and a joint miss ledger that records both the AI seat and the human was begun: twelve rows from one session (Section 7).
- **15 October (1 day).** Wolfberg's public miss register, rebuilt nightly, lists 740 misses logged between 13 June and 8 October 2026 and who caught each one (Wolfberg, 2026a). Section 6.3 reads it against the paper.
- **14 October (2 days).** micro1 published a landing page for the paper with four headline tiles, an interactive explorer for the horizon and breakeven calculations, and a closing section on what the work does not show (micro1, 2026).
- **10 to 13 October (3 to 6 days).** Edge updated on 10 October and Chrome on 13 October on the working machine (version numbers in Appendix A), two Windows updates followed on 13 and 15 October, and the tools' sandbox permissions changed. Headless PDF rendering hung in those days and a flag that disables the sandbox fixed it; the cause was not isolated. Section 6.2.
- **4 October (12 days).** Zhang and Esposito posted the working paper, arXiv:2610.07003v1 (econ.GN).
- **25 September (21 days).** Katoch, Selby, Großmann and Vollmer posted a workshop paper arguing that LLM judges carry biases, including differing severity, that more comparisons cannot remove, and fitting a latent-variable model that corrects for them (arXiv:2609.31184).
- **1 September (45 days; first posted 14 July, 94 days).** Sachdeva and Boudol fit many-facet Rasch models to annotations from nine LLMs and report that LLMs differ from human raters in severity, item-level calibration and rating-scale use (arXiv:2608.27463v2).
- **7 August (70 days).** Anthropic's auto-mode study planted one dangerous command in front of each of 1,053 paid testers; 13.6% caught it, falling from about 17% early in a session to about 5% after 50 or more prompts (Anthropic, 2026b, as summarized in Wolfberg's forthcoming note on the human gate; the source was not opened for this note). Section 7.
- **4 February (254 days).** Ford, Van Doren and Dix compared eight model releases on a fixed set of 726 adversarial prompts, using 82,256 human harm ratings, and found attack-success rates moved in different directions across generations (arXiv:2602.04739). Older background, kept because the paper leans on it as longitudinal evidence of release-to-release drift.

Table 1. The sources behind this note, newest first, with age in days before 16 October 2026. "How it was read" says how much of each was read. Appendix A gives the file locations behind the in-house rows.

Date (age)	Source	How it was read
16 Oct 2026 (0)	Wolfberg seed joint ledger (12 rows, one session)	File recounted by the analysis seat; an illustration, not a rate

Date (age)	Source	How it was read
16 Oct 2026 (0)	Working session between the operator and the analysis seat: the question, rulings and remarks in Section 7	Working session, 16 Oct; not logged to a file
16 Oct 2026 (0)	Wolfberg second-stage detector run and a human hand check of the detector	Counts as supplied by the agents; kappa 0.312 on 28 rows and the 17 of 19 come from this check
15 Oct 2026 (1)	Wolfberg public miss register, file build of 15 Oct	Counts and hash chain recomputed from the file
15 Oct 2026 (1)	Wolfberg relabel of the 7 Oct pushback episodes (rounds 1 and 2) and the human hand-check sheet	Relabel reports read; the 94.5% / kappa 0.92 and 84.7% / kappa 0.77 come from here; the hand-check sheet is the source of the kappa 0.157
14 Oct 2026 (2)	micro1 landing page for the paper	Full page text
13 Oct 2026 (3)	Chrome update on the working machine	In-house record
10 Oct 2026 (6)	Edge update; register audit write-up	In-house records
8 to 16 Oct 2026 (8 to 0)	Wolfberg full read of the practice's transcripts	Labels file recounted (36,657 rows); other counts as supplied by the agents
4 Oct 2026 (12)	Zhang and Esposito, arXiv:2610.07003v1	Main text, Tables 1 to 9 and the proofs; appendix tables from Table 10 onward only skimmed
3 Oct 2026 (13)	Transcript window of 3 October (07:24 to 09:05 UTC)	Fifteen items from the human's side; viewed on 16 Oct
25 Sep 2026 (21)	Katoch et al., arXiv:2609.31184v1 (workshop paper)	Abstract, introduction and method; results skimmed
1 Sep 2026 (45)	Sachdeva and Boudol, arXiv:2608.27463v2 (v1 posted 14 Jul, 94)	Abstract; method and results skimmed
7 Aug 2026 (70)	Anthropic auto-mode study (1,053 testers), as summarized in Wolfberg's forthcoming note	Wording checked against that draft on 16 Oct; source not opened
4 Feb 2026 (254)	Ford et al., arXiv:2602.04739v1	Abstract; results and conclusion skimmed

This note builds on Wolfberg's series piece on never letting a model grade its own homework (Atkinson, 2026a). That piece argued for a check from outside the model. This one asks the follow-up: when the outside check is a person or another model, how is it known to be the same checker tomorrow?

2 Position

Position. A score is a reading of the agent made by something, and the something is part of the number. The paper's strongest contribution is to say so with figures: on its data the roster change was about as large as the improvement, so the two canceled. Its three checks (rater comparability, evidence currency and corroboration) are the right ones. Its weakest part is the one the landing page puts in a tile: a five-week expiry date that rests entirely on a single step the change log does not explain.

Two claims are often run together, and they are kept apart here. One is that scores have to be made comparable across raters, which the paper shows well. The other is that evidence has a shelf life measurable in weeks. The second is directionally right, and the paper's own caveats say the number is an illustration. The method is worth taking (estimate the drift rate that the records do not explain, then compute how long a reading is good) and the figure is not.

The practical consequence is the one the series keeps arriving at, with a twist. Checking from outside the model still holds. The checker also has to be checked, because the checker is a rater: it has a severity, the severity may differ by question, and it can change without anyone logging a change.

3 What the paper shows

3.1 The design

The agent, Zara, conducts live job interviews for micro1 on a real-time voice model plus a text model. From 26 March to 24 August 2026 the firm's trainers watched recordings and scored them on a nine-family rubric: conversation structure, interruption handling, latency, speech pacing, coherence, audio-visual sync, freezes, audio quality and interview outcome. The stability index is the mean of the nine family scores, standardized; trainers also flagged "critical failures". In all, 2,611 production recordings were scored in 14 batches of 68 to 335. Reviewer A scored every batch and scored alone until 21 May; reviewer B joined then and scored roughly half of every batch; reviewer C scored part of the 21 May batch. The authors also have the agent's change log (three production deploys: 12 April, 16 July, 29 July) and daily support-ticket counts from 5 May.

3.2 Finding 1: a new reviewer looked like a worse agent

Which reviewer scored a recording predicted its score far better than which fortnight it came from: batch fixed effects explained 1.4% of case-level variance in the index, and adding reviewer fixed effects raised that to 16.2%. Reviewer B scored 0.792 SD below A (SE 0.045). The gap is not one number: it is -1.40 SD on latency and -0.13 on conversation structure, and B flagged critical failures less often than A (2.0% against 5.0%). Raw averages rose only 0.002 SD a month; with reviewer effects removed the trend is 0.102 (SE 0.021). The smoothed latent index rises from 0.05 to 0.58, a change of 0.53 (SE 0.07). Critical failures start at 2.8%, peak at 6.3% around 9 May and fall to 2.5% by 10 August. The authors'

rule: overlap raters, and never read a raw average across a changing roster. They add that the same applies to model judges: “a judge version is a rater.”

3.3 Finding 2: the agent moved when nothing was deployed

With the state allowed to jump at the three deploy transitions, what is left is movement the change log does not explain. Its weekly standard deviation is 0.086 (0.041 to 0.180). In the paper, forecast uncertainty starts from the smoothed variance at the last batch and reaches the size of the biggest deploy effect (0.20 SD, the April deploy) after about 5 weeks at the point estimate, 6 with the pooled rate, and 1 to 21 across the interval. (The landing page’s explorer instead starts from the steady-state precision of a batch of 300 every two weeks, and the analysis seat uses that to recompute in Section 5.) At a fortnightly cadence the forecast SD is about half a large deploy contrast, so a batch of 300 could still detect such a change between two readings. The authors say the horizon is best read as the point at which a firm relying on its records would have been wrong by a deploy-sized amount, not as a smooth forecast.

3.4 Finding 3: tickets agree in direction

Interview-related tickets fell about 10% a month relative to technical-issue tickets on the same dashboard (9.9%, t = -2.69, 5 May to 17 September), and 18% a month from June to August. Tickets per interview fell 7% a month over the full window. Interview tickets themselves did not fall in level (+3.7% a month); they failed to grow while the control category did. The authors read this as direction and order of magnitude, not an elasticity.

3.5 What it costs, and what the authors say they cannot show

Rater pay for the stability pipeline was $7,806 a month (engineering time is not counted, and fewer filed tickets need not mean fewer failures). Counting only avoided escalations (428 fewer interview tickets a month in September than at the May ratio), the program covers its cost if it deserves credit for about 18% of the reduction at $100 per escalation, 36% at $50 and 73% at $25. The authors state the limit directly: “we do not identify the causal return to the evaluation program.” No part of the deployment went unevaluated, and nothing varied evaluation intensity. They also say what would identify it: a record, kept at the time, of which finding motivated each commit, and designed variation in evaluation intensity. The landing page boils the paper down to four rules, referred to here as Rules 1 to 4: overlap raters; report families, not an index; measure one’s own drift rate; and design for attribution before starting.

4 The fine print

The analysis seat checked each of these against the paper’s text and tables. Two are credits to the paper, and they are given the same weight as the three problems.

4.1 No interview was scored by both reviewers

The landing page says, in its heading and tile, that on the same interviews B scored 0.79 SD lower than A. The paper says something narrower, and so does the page's own paragraph under that heading: two reviewers scored the same *batches*, and B took about half of each batch. Reading the paper's counts, the analysis seat finds one score per recording. There are 2,611 recordings in the sample and 2,611 cases in Table 2, and in Table 3 the two reviewers' counts add up to the batch size (27 July: 188 + 147 = 335, the largest batch; 24 August: 13 + 22 = 35, the partial batch). If recordings were scored twice, those sums would exceed the batch sizes. Status: inferred from the counts, because the paper never says in words that none was double-scored; its abstract and Section 3.2 use "overlapping" at batch level. Reviewer A scored alone until 21 May, then B took roughly half of each batch, and C took part of one batch. The 0.79 is therefore a comparison between two groups of different interviews drawn from the same fortnight. It measures reviewer severity only if interviews were split between reviewers without regard to difficulty, which the paper calls exchangeable assignment (its Appendix A). The paper states this as an assumption, does not say how the split was made, and tests it only indirectly. The B-minus-A gap is negative in all nine batches both scored, from -0.53 to -0.90, with a pooled value of -0.77; a heterogeneity test rejects an exactly constant gap ($Q = 19.0$, $p = 0.015$); and the reviewer-A-only trend (Section 4.4) needs no adjustment. That is decent circumstantial support. The paper's own rule is "a fraction of every batch to two raters, at least one rater on every batch", which can be read as double-scoring a sample or as sending part of each batch to a second rater, which is close to what micro1 did. The wording is ambiguous; double-scoring a sample is the recommended design, because it is the only design that separates severity from the difficulty of the interviews each rater happened to get. The landing page's headline wording overstates what was done.

4.2 The five-week point estimate rests entirely on one step

Table 6 gives three estimates of the index's weekly drift. With a free jump at the three deploys: 0.086 [0.041, 0.180], likelihood-ratio $p < 0.001$. Also freeing the late-May transitions (9 May to 1 June, where the sup-F test places a jump of 0.34 SD that the change log does not explain): 0.000 [0.000, 0.151], $p = 0.50$. So the point estimate falls from 0.086 to zero when one step is set aside, and the five-week figure goes with it. The interval still runs to 0.151, close to the original 0.180, and $p = 0.50$ is a failure to reject a constant state, not evidence for one: the data cannot say drift is zero. For the critical-failure rate part of the drift survives freeing the step (0.666 percentage points a week, $p = 0.16$ against 0.762 before). At the index's upper bound the horizon is 1.6 weeks (Section 5). The paper does disclose this: its abstract says the horizon "reflects a small number of step-like changes," and Section 5.2 says the index moved in a few steps rather than continuously. The landing page body (its Finding 3) discloses the step too, and calls the number an illustration with a wide interval, not a measured expiry date. Only the headline tile, "~5 weeks", omits it. The honest reading is that a firm should expect steps of roughly this size that its change log does not explain, and that these data cannot say how often.

4.3 Whose agent, whose evaluation

The agent is micro1's own product, run on micro1's own interviews (paper, Section 3.1). The PDF carries no affiliation line, funding statement or competing-interest statement; the landing page lists the

second author, Esposito, as micro1's Chief Economist and the first, Zhang, as at Harvard University. micro1's site menu also offers evaluations and RL environments for AI labs under the name Realm, and a visibility-and-improvement layer for agentic AI under the name Cortex. The paper is a working paper, marked preliminary, posted to arXiv on 4 October and not peer reviewed. None of that makes the numbers wrong: the tables are internally consistent and several checks reproduce. It is two facts a reader should weigh: the agent belongs to one author's employer, and that employer sells evaluation products. To the authors' credit the paper makes no return claim and says the breakeven is a condition, not a return.

4.4 Credit: the improvement does not depend on the adjustment

Reviewer A scored every batch, so that reviewer's scores alone need no reviewer correction. Table 4 puts the A-only trend at 0.124 SD a month against 0.102 with the adjustment, and 0.125 with deploy-regime fixed effects. The raw average flattened because of who joined, not because nothing improved. A common drift in both reviewers' standards would survive this check; the authors hold that as an assumption. It is the best single piece of evidence in the paper and the most underplayed on the landing page.

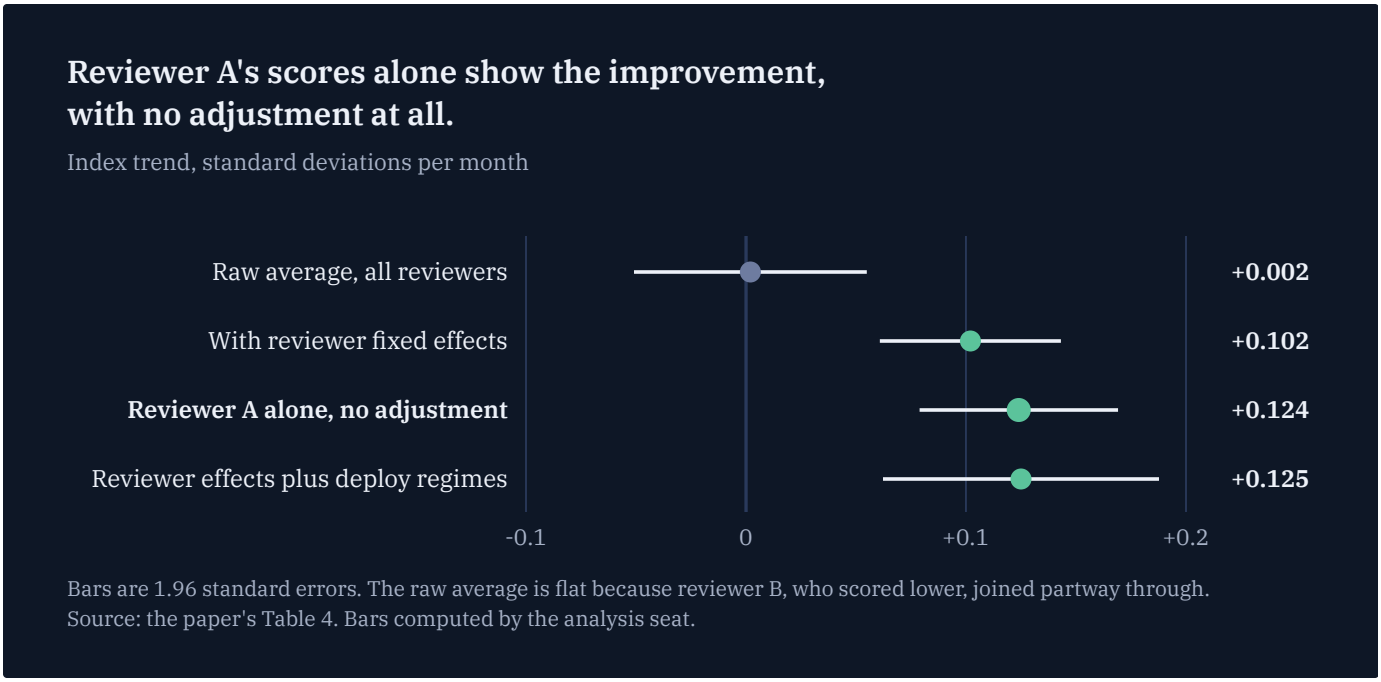


Exhibit 2. Four estimates of the index trend, per month. Only the raw average, which mixes two reviewers' scales, shows no improvement. **From the paper's Table 4; bars are 1.96 standard errors, computed by the analysis seat.**

4.5 Credit: causality is disclaimed, honestly

The authors do not claim the evaluation program caused the improvement, say why (no unevaluated arm, no variation in intensity), and say what would be needed. Commit timing does not cluster after sign-off dates (63% of commit groups fall within seven days of one, against 67% under a permutation

null), and no commit in the changelog cites an evaluation finding. The paper adds that its own simulation detects a commit-after-finding pattern in only 37% of replications, so the timing evidence neither supports nor rules out the channel. That is the right posture, and it is adopted for Wolfberg's own work in Section 6.

4.6 Smaller points worth knowing

- **Severity is a vector.** B is stricter than A on judgments of degree (latency, pacing, audio quality) and agrees with A on events (freezes, sync). On the critical-failure flag B is more lenient. One leniency constant would not fix that.
- **The critical-failure trend is not clear.** With reviewer effects the monthly trend is -0.784 percentage points (SE 0.451), wild-bootstrap p = 0.193. The story there is the hump, not a slope.
- **The population moved.** The share of software roles among interviews rose from 17% to 27% between March and May, and applicant sources shifted from June. The recordings carry no information about either, so some measured movement could be the interview mix rather than the agent.
- **The landing page quotes a different figure.** It says the paper reports 4.8 weeks; the paper text the analysis seat read says "about 5" (4.8 appears nowhere in it), and the explorer shows 5.2. Section 5 sorts these out.

5 What the recomputation reproduced

The paper's horizon comes from its equation (6) and the closing paragraph of Proposition 2. After a batch of n sessions the posterior variance is P_0; it then grows by the weekly drift variance q each week, and the horizon is the number of weeks until the forecast variance reaches the square of a change the firm cares about. The analysis seat used the landing-page explorer's defaults, which the paper's cadence discussion matches: a batch of 300, a two-week gap, a standardized index (variance 1), and a change of 0.20 SD, the April deploy's effect. Status: measured (recomputed from the paper's formula and inputs).

THE FORMULA RECOMPUTED

$$P_0 = \frac{\sqrt{(qh)^2 + 4qh\frac{\sigma^2}{n}} - qh}{2}, \qquad s = \frac{D^2 - P_0}{q}$$

q = (weekly SD)^2; h = 2 weeks; n = 300; σ^2 = 1; D = 0.20; s is the horizon in weeks. The batch SE reported is the square root of P_0.

Table 2. The horizon, recomputed. The values come from the formula above with the explorer's defaults. The paper's own text uses the smoothed variance at the last batch in place of P_0 (its Appendix A), so its figures

differ slightly.

Case	Weekly SD	Batch SE	Horizon, recomputed	What the paper or page says
Point estimate (state free at deploys)	0.086	0.0529	**5.03 weeks**	Paper: about 5. Page: paper reports 4.8; explorer 5.2
Pooled (all transitions)	0.075	0.0519	6.63 weeks	Paper: 6
Lower end of 95% interval	0.041	0.0454	22.6 weeks	Paper: up to 21
Upper end of 95% interval	0.180	0.0564	1.14 weeks	Paper: down to 1
Late-May step freed, upper end	0.151	0.0559	**1.62 weeks**	Not stated; from Table 6
Late-May step freed, point estimate	0.000	-	never	Table 6: 0.000, p = 0.50

Three further checks agree to within rounding. The forecast standard deviation at the point estimate is 0.180, 0.249 and 0.303 at 4, 8 and 12 weeks; the paper gives 0.18, 0.25 and 0.30. The smallest change two consecutive batches can tell apart (95%) is $1.96 \times \sqrt{2} \times 0.0529 = 0.147$; the page says 0.14 (it uses 0.052, which gives 0.144; the recomputed value is 0.147). A batch's own precision at the point estimate is 0.0529 where the page shows 0.052. Both are within rounding of 0.05, but the recomputed values do not round to the page's, so "agree" holds only to two decimals.

Two things are worth saying about the shape. First, q enters squared, so an interval of 0.041 to 0.180 on the SD (a factor of 4.4) becomes a factor of about 20 on the horizon (22.6 against 1.14 weeks). The paper's "1 to 21" is that arithmetic, not vagueness about the model. Second, the late-May row matters more than any other: it moves the point estimate from 5 weeks to "never", and the interval for the horizon from 1.1 to 22.6 weeks to 1.6 weeks to never. The upper bound on drift falls from 0.180 to 0.151, so the shortest horizon gets longer (1.1 to 1.6 weeks), not shorter.

On the page's 5.2 and 4.8: the page says its explorer starts from the steady-state precision of a batch where the paper starts from the smoothed estimate at the last batch, and the paper's appendix confirms its text uses the smoothed variance. That explains why the two differ, but the analysis seat could not reproduce the explorer's 5.2 from the inputs it states. The steady-state formula gives 5.03; the several variants it tried (using σ^2/n alone, a continuous-time steady state, a one-week gap, 400 recordings) gave 4.74 to 5.41, none exactly 5.2. It also found no "4.8" in the paper text it read. All of these are "about five", and none changes any conclusion.

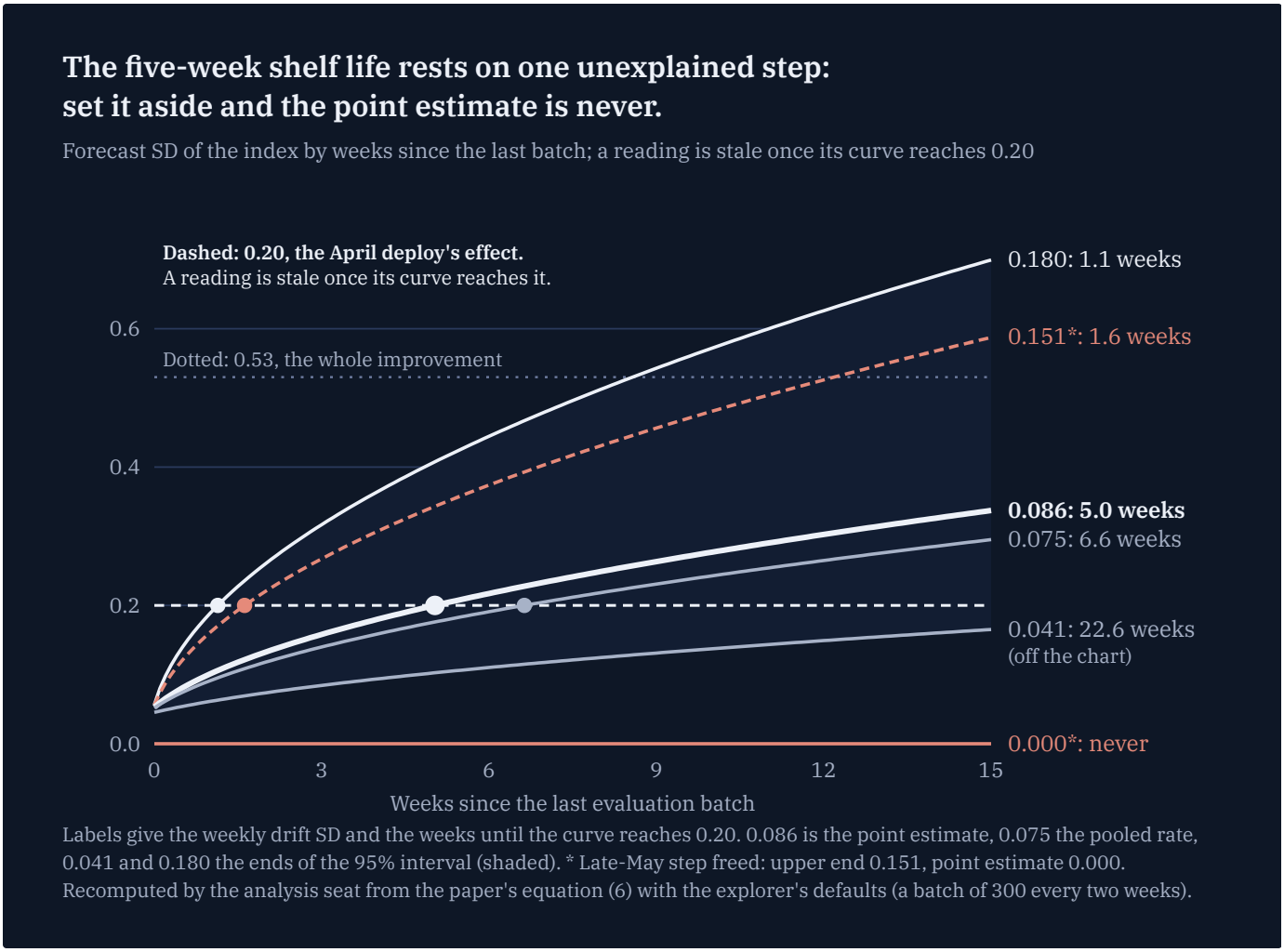


Exhibit 3. The evidence shelf-life curve, reproduced. Forecast standard deviation of the latent index against weeks since the last batch, drawn at the point estimate of the weekly drift (0.086), the pooled rate (0.075), both ends of the 95% interval (0.041 and 0.180) and with the late-May step freed (upper end 0.151; point estimate 0.000, flat). A reading is stale where its curve meets 0.20, the April deploy's effect. **Recomputed by the analysis seat with the explorer's defaults (a batch of 300 every two weeks); the paper's Figure 7 starts the curve from the smoothed variance at the last batch instead.**

Table 3. The breakeven table, recomputed: the share of the 428 avoided tickets a month that the program must own to cover $7,806 a month, and the net value at full credit.

Cost of one escalation	Credit needed, recomputed	Paper's Table 9	Net value at full credit, recomputed	Paper's Table 9
$25	73%	73%	$2,894	$2,905
$50	36%	36%	$13,594	$13,616
$100	18%	18%	$34,994	$35,037
$150	12%	12%	$56,394	$56,459

Cost of one escalation	Credit needed, recomputed	Paper's Table 9	Net value at full credit, recomputed	Paper's Table 9
$250	7%	7%	$99,194	$99,303

Credit needed is 7,806 divided by 428 times the cost of an escalation, and it matches in every row. The net values differ from the paper's by $11 to $109, growing with the cost per escalation. Working backward, every paper row implies 428.4 avoided tickets, not 428, which is consistent with an unrounded count. At $100 an escalation, $7,806 over 428 tickets is about $18 per avoided escalation, as the paper says.

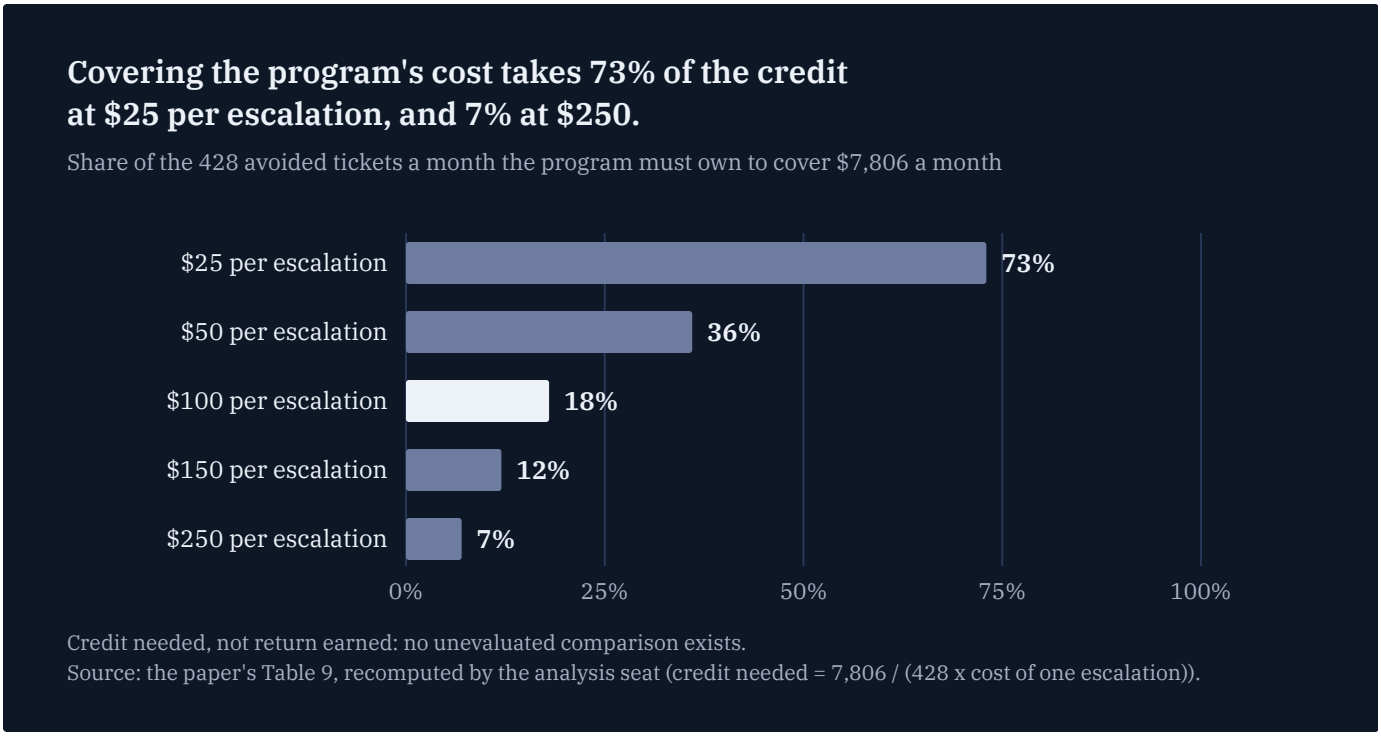


Exhibit 4. Credit the program must own to cover its cost, by what one escalation costs. This is credit needed, not return earned: no unevaluated comparison exists. **From Table 3 (recomputed by the analysis seat; equal to the paper's Table 9).**

6 The three checks applied to Wolfberg's practice

The paper's three checks are rater comparability, evidence currency under drift, and corroboration with an outside outcome. Wolfberg runs an AI team and keeps two scoring instruments of its own: a full read that labels every item in the practice's transcripts (the labeler is a language model, with a human and a second-stage detector as checks), and a public register of misses that records who caught each one. Both have the paper's problem. This section uses only counts the team has produced, and each is a count from one practice, not an estimate of anything beyond it. The analysis seat, which did the recomputation above, also recounted the labels file and the register file where noted.

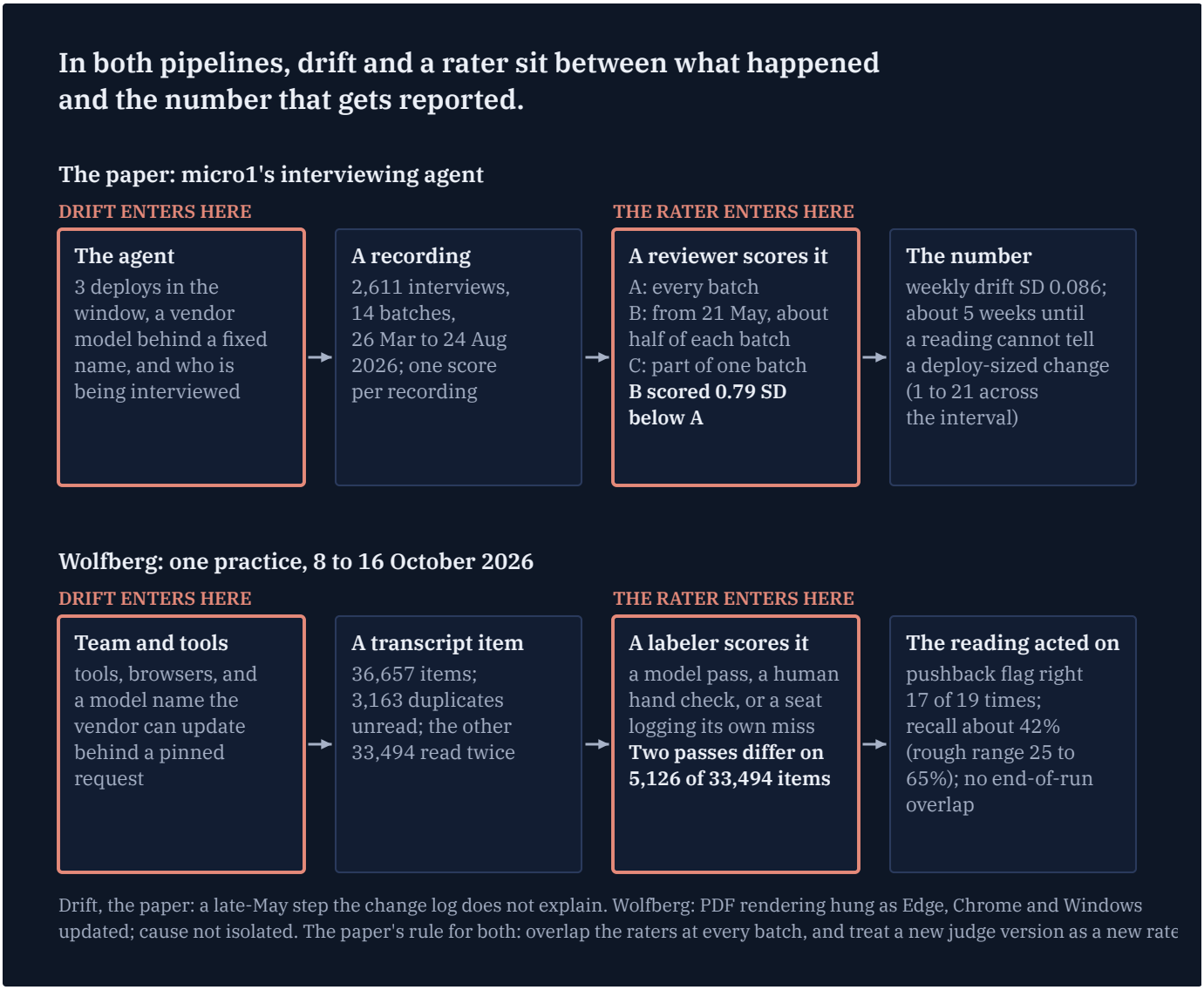


Exhibit 5. How a rating becomes a number, in the paper's pipeline (top) and Wolfberg's (bottom). Red marks the two places where a reading picks up something that is not the agent: drift, and the rater. **The paper's side is from its Sections 3 to 5; Wolfberg's uses the counts in Table 5 and Section 6.3. The columns match each stage to its counterpart; the match is a mapping, not a causal link. In Wolfberg's lane, the labeler's second rater is the fallback model that read 2,004 items after refusals.**

Table 4. The paper's three checks, set against what Wolfberg has and what is missing.

The paper's check	What the authors did	What Wolfberg has	Gap
Rater comparability	Two reviewers in the same batches (one score per recording); adjusted for the gap	Two model passes on one item set; human hand checks on small samples; the register's catcher field	Passes disagree item by item; no second catcher re-scores the same sessions

The paper's check	What the authors did	What Wolfberg has	Gap
Evidence currency under drift	Estimated drift their records do not explain; computed a horizon	A pinned request; dated register rows; a recency rule	Overlap measured at the start of a run only; no drift rate
Corroboration with an outside outcome	Support tickets against a same-instrument control	19 misses found by an external probe; hand checks	No operational outcome series; the register does not record which fix a miss led to

6.1 Check one: the labeler is a rater

From 8 to 16 October 2026 one model read 36,657 transcript items for the practice. 3,163 were exact duplicates of earlier items and were not read again. The other 33,494 were each read twice, from requests that shared one instruction prefix (the same hash on all of them). For 2,004 of those 33,494, at least one pass came from a fallback utility model after the main model refused. That is a different 2,004 from the items flagged as pushback below; only 383 items are in both. Pass 1 found 42,288 records and pass 2 found 42,236, a difference of 52 (1,975 of pass 1's records and 1,964 of pass 2's came from the fallback). Only 16,732 matched on exact anchor text, which is a floor on agreement because the same statement anchored on different words counts twice, and the union is 67,792. In total, 5,126 of the 33,494 items (15.3%) had a disagreement between passes: 4,579 of the 31,490 items where both passes came from the main model (14.5%), and 547 of the 2,004 fallback items (27.3%). Of the 2,004 items flagged as pushback, 581 were flagged by one pass only. The totals agree; the exact-text matches, which are a floor, cover about 40% of either pass (16,732 of 42,288 and of 42,236: 39.6%).

Table 5. Who agrees with whom in the read. Counts from one practice; the hand-checked samples are small, and "same model" means the main model, which a fallback model replaced on some items.

Comparison	Question	Result
Same model, same prompt, two passes	Which kind of pushback episode (15 Oct relabel of the 804 episodes from the 7 Oct pass: four kinds, 257, 85, 197 and 245 episodes; 20 left unlabeled after both passes were refused)	94.5% agreement, kappa 0.92, on the 777 episodes with two valid passes
Same model, two passes, relabel round 2	"Did the seat fold?", using the full read's own definition	84.7% agreement, kappa 0.77, on 523 episodes with two valid passes; the disagreement sits on the fold-versus-repair boundary
Same model, two passes, full read	Which items to flag, and where to anchor	42,288 vs 42,236 records; 16,732 exact matches; 5,126 of 33,494 items differ (14.5% where both passes were the main model, 27.3% on fallback items)

Comparison	Question	Result
Human hand check against the model	“Did the seat fold?” On a hand-check sheet built to hold 10 episodes in each of three label strata (folded, repaired, held), so counts are not rates	The full-read prompt called 11 of the 30 folded; the human judged 3 of those 11 not pushback, 4 repaired, 3 held and 1 folded. On the 23 episodes the human judged as pushback, fold versus not-fold gives kappa 0.157, resting on a single episode the human called folded
Human hand check against a second-stage detector	The same family of judgments	kappa 0.312 on 28 rows
Human hand check of the read’s pushback flag	When the flag said yes, was it right?	Right 17 of 19 times. Recall about 42%, rough range 25 to 65%, from 4 of 9 randomly drawn messages the read had not flagged (the human called them pushback) together with the 19 flagged ones
The 7 Oct labels	Rows marked “the seat was right”	125 rows were marked that way; the 15 Oct relabel (a model) puts 79 of them in the not-pushback kind, which agrees with the earlier human finding

The pattern is the paper’s. A rater can agree with itself and still not be the rater it was taken to be. On the fold question the model’s two passes agree with each other at 84.7% (kappa 0.77, 523 episodes), while the model against the human hand check is at kappa 0.157 on 23. That is a gap between self-agreement and agreement with the person, not a difference between question types, and the 94.5% figure is a different question (the kind of pushback). There is no test of whether the labeler is steadier on events than on judgments, so none is claimed. Status: conjecture, that the paper’s finding that severity attaches to judgments of degree will show up in the labeler. The samples behind the lower kappas are small (23 episodes, with a single episode called folded among them; 28 rows; 9 unflagged messages), so no tight interval should be put on them. They are enough to say the read has not earned a single number.

The 3 October window. On 3 October 2026, between 07:24 and 09:05 UTC, the read held 15 items from the human’s side of the conversation; four of them are mostly pasted or relayed text, and one is a task brief for another seat with no conversation context. The read flagged none of the 15. When the items were viewed on 16 October the operator’s reaction was recorded (the operator, 2026-10-16: “theres a lot of pushback there”). One was an exasperated message (the operator, 2026-10-03: “nooooo a post that goes with the who first pdf”); another was an impatient one (the operator, 2026-10-03: “whats taking so long”). The remaining items are not published. The pasted items make the window a weaker test than “zero of 15” sounds. Precision when the flag said yes was 17 of 19; recall was about 42%, and in this window it was zero of 15. Recall is uneven, which is the paper’s point about a judge version being a rater, turned inward: the number that looks like a property of the labeler is partly a property of the stretch of conversation it is applied to.

6.2 Check two: evidence goes stale, and some of the drift is nobody's change

The paper's drift is movement its change log does not explain: a vendor model changing behind a fixed name, a configuration outside the repository, a different mix of interviews. Wolfberg has the same exposure, with one example that can be dated and a cause that was not isolated.

- **The browsers.** Edge updated on 10 October and Chrome on 13 October. Headless PDF rendering then hung, and a flag that disables the sandbox fixed it. The timing fits the browser updates, but two Windows updates (13 and 15 October) and a permissions change to the tools' sandbox landed in the same days, and the seat that traced it did not separate the causes. It serves as an example of a pipeline moving under a pinned recipe, not as a measured cause. It is still something that changed in the environment that was not in the practice's own change log.
- **The labeler.** It ran under a fixed model name across eight days. The request bytes were pinned for the main model, and the matching prefix hash shows the instructions did not change, but what the vendor serves behind the name cannot be pinned, and a fallback model read about 6% of the items (2,004 of 33,494), a second rater inside the run. Overlap was measured only at the start, with a third read of 250 items before launch, not at the end. The paper's rule says to overlap raters at every batch.
- **The working rule on source age** since 8 to 10 October is that more recent information carries more weight, and that model-landscape sources two to three months old are stale. That is the qualitative version of the paper's horizon: a reading expires, roughly when can be said, and no rate has been attached to it.

The paper's method suggests the missing piece. A fixed anchor set can be re-read at the end of a long read, as well as at the start, and the change in agreement treated as an estimate of unlogged drift over that run. It has not been run.

6.3 The public miss register is a reading that contains the rater

The register at wolfberg.ai/register lists every logged miss by Wolfberg's team, with its type, a one-sentence account, and who caught it. The file behind it (built 15 October 2026) holds 745 entries: 740 misses dated 13 June to 8 October 2026, and five "gate" rows, which are a fixed example list and not real gate decisions, so they are not counted as evidence. Misses are read from the team's session-health ledger, a plain file that is appended to by convention. The entries are sha256 hash-chained: the analysis seat recomputed the chain and found 0 broken links across the 745. What that protects: each entry's hash covers its own content and the entry before it, so an edit inside the published file shows. What it does not protect: the chain is rebuilt whole every night, with newer rows inserted above older ones, so no hash carries from one night to the next. An edit to a row shows only if someone kept an earlier copy and compares row text, and a miss nobody logged is invisible (the production-signing details come from an internal audit dated 10 October). None of the 740 misses carries an Ed25519 signature; each is marked "hash-chained, not Ed25519", and only the five example rows and the chain head are signed.

Who caught each miss: the operator 378, the agents unprompted 291, a peer instance 51, an external instrument 19, unattributed 1. The most common types are repeated miss (177), wrong diagnosis (161), would-ship error (132), recall assertion (73), unsearched claim (56) and self-corrected (46).

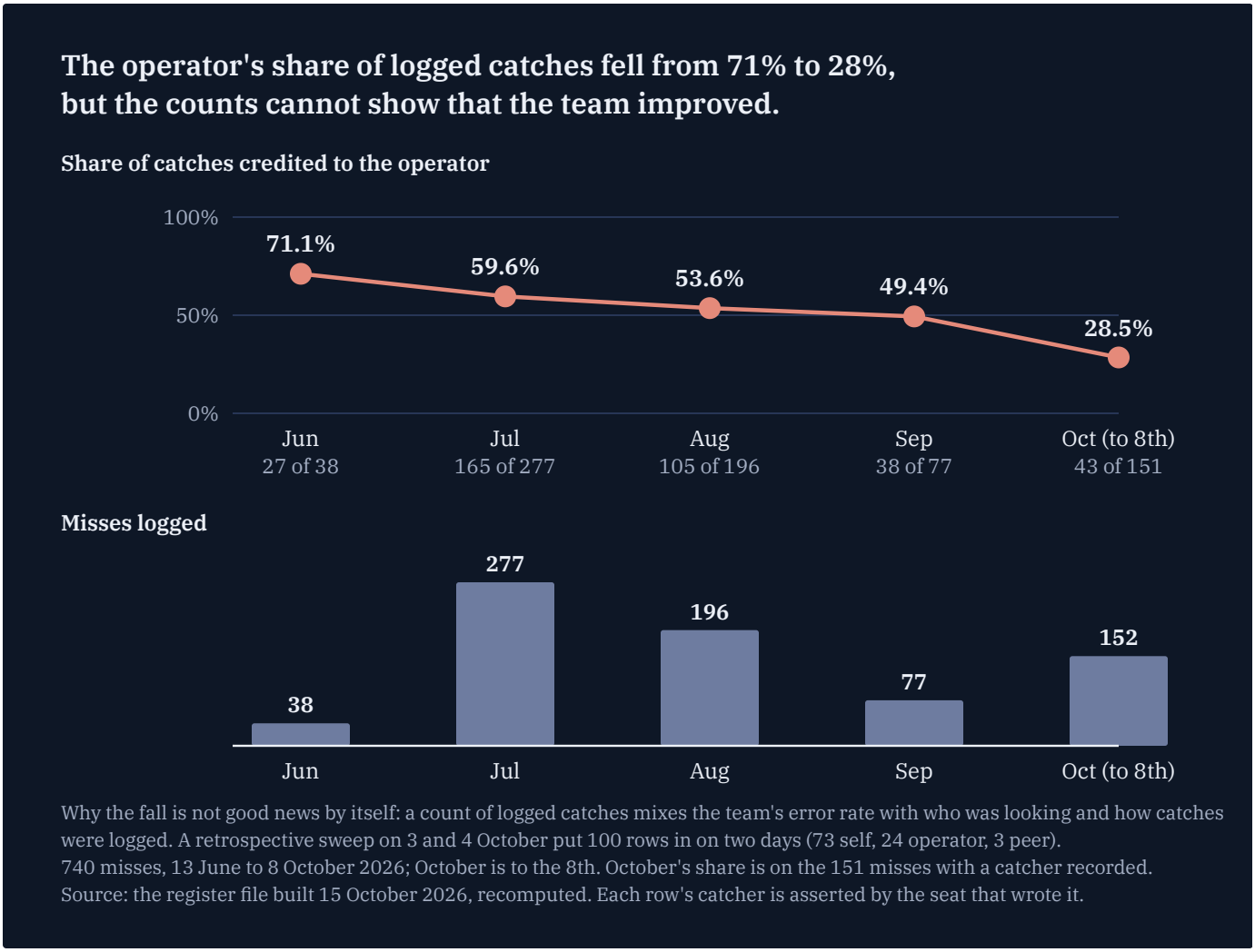


Exhibit 6. The public miss register by month: the operator's share of the catches (top) and misses logged (bottom), June to early October 2026. The share fell while more of the logged catches came from the agents and peers; the counts cannot show whether the team got better, because who logs a catch is not independent of what gets caught. **From the register file (built 15 October 2026), recomputed. The October share is on the 151 misses with a catcher recorded; October logged 152 misses in all (740 less the 588 of June to September).**

The register reads the way the paper reads reviewer B. The operator's share of the catches fell from 71.1% in June (27 of 38) through 59.6% in July, 53.6% in August and 49.4% in September to 28.5% in early October (43 of the 151 with a catcher recorded). That is the observed fact. The tempting reading is that the team is catching more of its own mistakes and getting better. That reading is not taken here, for the paper's reason: a count of logged catches mixes the team's error rate with who was looking, how hard, and how catches were logged, and the data cannot say whether the catchers changed or the logging did. A fall in the operator's share could come from the agents catching more, from the operator looking less, or from when and how rows were filed: a retrospective sweep on 3 and 4 October put 100 rows in on two days (73 self, 24 operator, 3 peer; from the internal audit dated 10 October). The register cannot say which, and the June-to-October counts must not be read as the team getting better or worse.

Three further points map directly to the paper's rules.

1. **Overlap.** The register's catcher field is set by the seat that writes the row, at write time, and nothing checks the claim against an independent signal (internal audit, 10 October). Of the operator's 378 rows, 159 carry the operator's own words, so the other 219 rest on the logging seat's word alone (same audit). No second catcher re-scores the same sessions, so a catcher's severity cannot be separated from the error rate. The field is the start of the roster the paper's method needs, not the overlap.
2. **Evidence currency.** Every row is dated, so the age of any claim is visible, which is the paper's shelf-life rule built into the format. The guarantee is weaker than it sounds: the ledger is appended to by convention and the chain is rebuilt nightly, as above, and a miss that nobody logged is invisible.
3. **The paper's Rule 4.** The register records what happened and who caught it. It does not record which fix a miss led to. That is exactly the gap the paper says prevents it from identifying the value of its own evaluation program. The cheap thing to add is one field on each row: the fix, if any, that the miss led to.

The corroboration check has a thin counterpart here. The only rows written by a separate process are the 19 probe rows, all of one type (probe miss); the 159 rows credited to the operator with the operator's own words are still logged by a seat. There is no outside outcome series like the paper's tickets, so the register can say what was caught and by whom, and cannot say whether it mattered. Two further limits: the 15 October build contains no miss after 8 October, and by the 10 October audit, catches from the analysis seat's reviews never reach the ledger, so the register undercounts outside catches. The register paper that explains the design, a separate forthcoming piece, is not yet published; this note cites only the live page.

6.4 Next steps

- Re-read a fixed anchor set at the end of a long read, not only at the start, and report the change in agreement as an estimate of unlogged drift. Not yet run.
- Put a second catcher on a sample of sessions. Double-scoring a sample is the preferred design, because sending part of each batch to a second rater does not by itself separate catcher severity from the error rate.
- Add a fix field to each register row, so a miss can be traced to the change it caused, and the origin, escape and before-or-after-ship fields of Section 7.
- Vary the harness (the set of checks wrapped around the model) on purpose, by session, so that a counterweight's effect (a mechanical check against the model's pull toward agreeing with the user) can be identified rather than inferred from before-and-after counts (the paper's Rule 4). Today every session gets the same one.

7 In the loop, or watching the loop?

On reading the first draft of this note the operator asked the analysis seat whether agent reliability is really this: “How do you measure if the human is in-the-loop or watching-the-loop?” (the operator, 2026-10-16). The paper doesn’t ask it, but its three checks answer it, because the human who approves AI work is a rater too. Everything the paper says about rater severity, drift and overlap applies to the human gate. Some humans are strict and some lenient. Attention drifts: in the study Anthropic published on 7 August 2026 (70 days ago), as Wolfberg’s forthcoming note on the human gate (“Who Caught It”) reports it, 1,053 paid testers each met one planted dangerous command and caught it 13.6% of the time, falling from about 17% early in a session to about 5% after 50 or more prompts (Anthropic, 2026b; Wolfberg, forthcoming). And usually nobody re-scores what the human approved.

7.1 What “in the loop” has to mean

The definition here is by effect. A human is in the loop when their decision changes outcomes before they take effect, and changes them because the item was wrong. A human who approves at the same rate whether the item is wrong or right is watching, however carefully. That gives three measures, and only the first is the test.

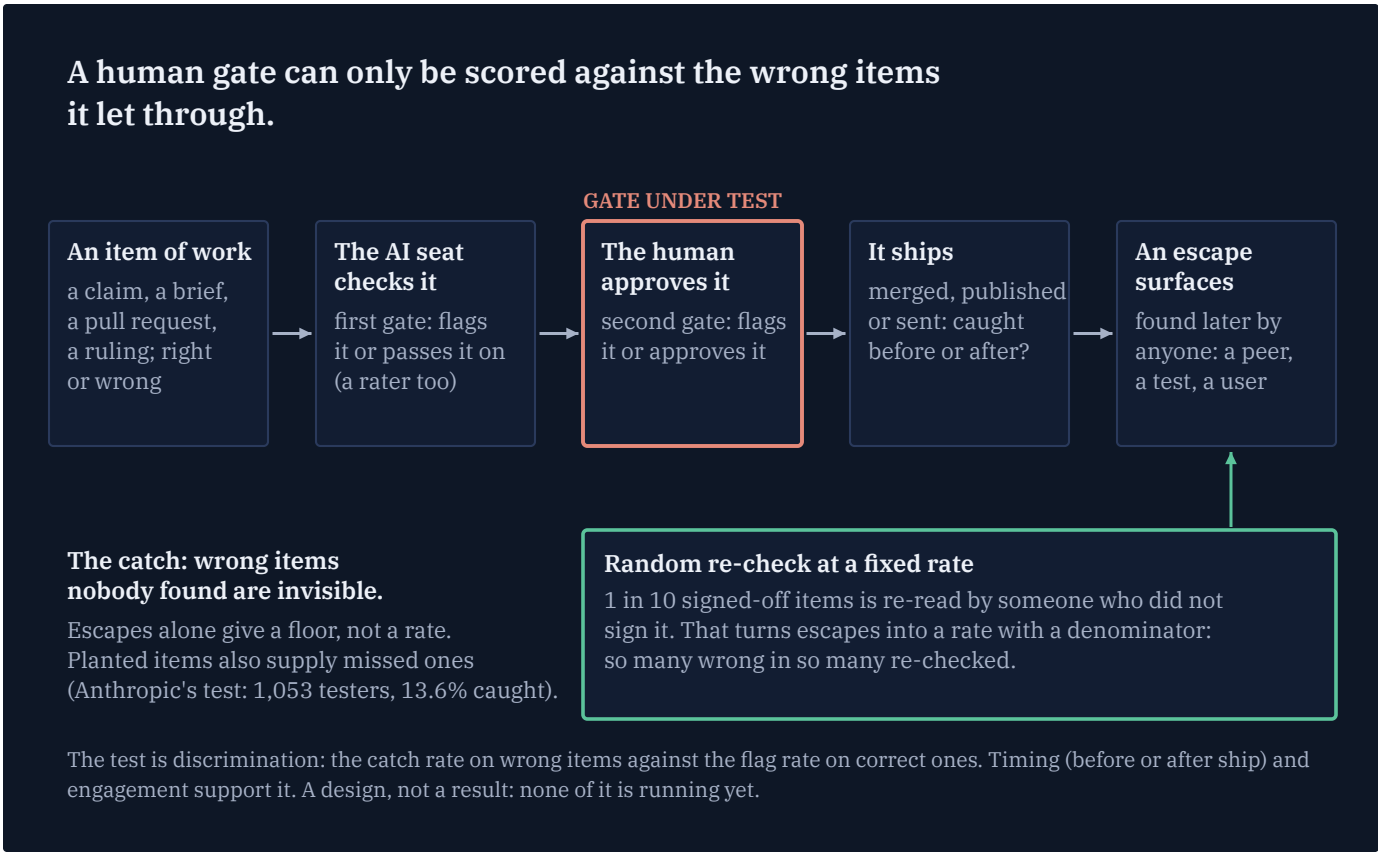


Exhibit 7. Where a human gate can be measured. The human’s approval is the gate under test; the AI seat’s own check is a rater too. A random re-check of signed-off work turns escapes into a rate. **A design, not a result; none of it is running yet (Section 7.6).**

Table 6. Three ways to measure a human gate, in order of weight.

Measure	What it asks	What it needs	Weakness
(a) Discrimination (the core one)	The catch rate on wrong items against the flag rate on correct ones	Wrong items the human did not catch, as well as those they did; correct items too	The denominator: wrong items nobody found are invisible
(b) Timing	Was it caught before or after the merge, publish or send?	A before-or-after-ship field on each miss	Says when, not whether the human did it
(c) Engagement (supporting only)	Time on the item, evidence opened, overrides with a reason, approval rate moving with risk	Telemetry from the review tool	All of it can be present while the decision is wrong

7.2 The denominator problem

Discrimination needs the wrong items the human did not catch. A log of catches alone can't supply them: a human who catches 40 of 40 logged misses might have missed 400 that nobody found. There are two routes. The first is planted items, as in Anthropic's test: put known-wrong items in front of the human and count. That gives a rate directly, at the cost of an item that is not real and a human who may know it is a test. The second is escapes: wrong items that got past the gate and surfaced later, found by anyone. They're real, and they're all that comes for free, but on their own they are a floor, not a rate, because the ones that never surfaced are missing.

The earlier position, in the forthcoming "Who Caught It" note, was that a practice needs both real catches and planted items. That position is updated: escapes plus a random re-check are chosen over plants. Escapes are far more effective than planted items in this practice, because the agents are assumed to be wrong in some way in every response. The assumption is the design: if every AI response is presumed to contain a miss, the question is only who finds it and when.

7.3 An escape count measures the finder too

This is the paper's reviewer-B problem one step later. A count of escapes measures whoever found the escape: a human who reviews more, or looks harder, finds more, and the count says nothing about the gate they were checking. The fix is the paper's: overlap the raters. In practice, a random re-check of signed-off work at a fixed rate, for example one item in ten, re-read by someone who did not sign it. That gives a rate with a denominator (so many wrong in so many re-checked). Escapes found any other way remain a floor. And an escape counts against a party only if the miss was actually in front of them: a defect in code the human never saw is not a lapse in their approval.

7.4 The unit is the pair

The paper's logic cuts both ways, and the human is not exempt: both sides are to be analyzed, the agents and the human. Each miss row should record the origin (the AI seat, the human or both), whether it was in front of the seat and in front of the human, who caught it, whether it was caught

before or after it shipped, and how long it took to surface. The human's own misses are decisions, rulings, instructions or work claims that steered the work wrong. They are not offhand recollections: grading the human's memory of things months later would probably go badly. The human can miss and be wrong too.

Table 7. A seed joint ledger: one session, 16 October 2026, 12 rows. An illustration of the fields, not a rate.

Field	Count
Origin	AI seat 11; both 1; human alone 0
Caught by	The human 5; peer AI instances 5; the seat itself 1; an external instrument 1
In front of the AI seat; in front of the human	12 of 12; 8 of 12
In front of the human and not caught by the human	4 (three caught by a peer instance, one by an external instrument)
Caught before it shipped; after	9; 3
Time to surface	Under 20 minutes to about 9 days

The twelve rows are an overstated claim in a brief, a cause asserted but not isolated, an apples-to-oranges kappa comparison, a claim about the register made without reading its audit, a test gap, a forecast the analysis seat and the human both accepted, a file lock, an unwired logging path, and four labeling and recall errors in the practice's own read that a review caught (the 79 of 125 rows, the fold label, the 3 October window and a wrong claim about who renders the PDFs). Four of the twelve were errors in this note's own first draft, caught by a reviewer instance before it left the folder. Four were in front of the human and not caught by the human, so they are escapes past the analysis seat and the human: three surfaced before shipping, one after.

The human-origin row count is zero, and it should not be read as a score. Nobody looked for the human's misses on purpose; the register was built to log the AI seat's. It is a coverage gap. The one shared row is a forecast the analysis seat and the human both accepted.

7.5 Code is where the human cannot see

The least visible misses are the ones buried in code. They are the easiest for a human to miss, because components do not get wired together until the agents and the human test after merging. The example is the analysis seat's miss-logging path, which wrote its reports into a folder that no importer read, for nine days, while it appeared to work. Neither the seat nor the human saw it: it was in front of the seat, not in front of the human, and surfaced after it had shipped. On 16 October two mechanical checks were adopted. First, every pull request that adds output names its consumer and includes a test through it. Second, a nightly check flags output with no reader for more than three days. A new miss type for it was declined; whether it needs its own type is uncertain. The point for measurement is that a

human gate can't be the control for what a human can't see, which is the series' usual answer: put the check outside the people, and measure the people on what they can see.

7.6 Honest status

None of this is running yet. The register has no re-check sample, no escape field and no before-or-after-ship field. What exists is the catcher field and a twelve-row seed from one session. The next steps are the fields, the one-in-ten re-check, and a per-catcher discrimination rate once there are wrong and correct items in front of each.

8 What it means

8.1 For anyone who scores an agent with people or model judges

Raters should be overlapped, and kept overlapped at every batch, not only at the start. A raw average should never be read across a changing roster. A new judge version should be treated as a new rater, including one that arrives under the same model name. Results should be reported by family or question, not by index, because severity differs by question (the paper's B is harsher on degree than on events; whether Wolfberg's labeler differs that way has not been tested). And it should be written down, at the time, which finding motivated which change. The human who approves the work is one of the raters: what should be measured is whether their decisions change outcomes (Section 7), not whether they approve.

8.2 For anyone reading a paper like this one

A landing page and its paper should be read side by side. Here they differ in the places that matter: "the same interviews" against same batches, "about 5 weeks" against a horizon resting on one step, 4.8 against 5.2. Who benefits from the headline should be checked. A working paper by a vendor of evaluation about its own agent is evidence, and a check is cheap: the paper's tables let the headline be recomputed in an afternoon, and the recomputation here reproduces.

8.3 For Wolfberg's practice

The register's monthly counts do not support saying the team is getting better, and the post that goes with this note says so. The register's job is to make each claim dated and attributable; it can't yet do the paper's job of separating the scorer from the scored. The first changes are cheap: a second catcher on a sample, a fix field, and the escape and before-or-after-ship fields. The third, varying the harness by session, is how it can be learned whether a counterweight does anything, and it has to be designed in before the sessions run, which is the paper's last lesson.

9 Limits

The analysis seat read the paper's main text, Tables 1 to 9 and the proofs from a text extraction of the PDF; figures were read through their captions and printed values, and appendix tables from Table 10 onward were skimmed. The authors' data and code were not available, so this reproduction is of their arithmetic from their reported inputs, not of their estimates. The three related papers were read in part: the abstracts and introductions in full, with methods and results skimmed (Table 1); none is a close replication of the paper. The Ford et al. study concerns release-to-release change, not silent change behind a fixed model name. The landing page's explorer was read as text and was not run, so the 5.2-week figure is as quoted. The in-house numbers come from one practice's full read and register. The labeler is a model, and a fallback model read about 6% of the items; the hand checks are small (19 flagged messages, 9 unflagged, 23 episodes, 28 rows); the agreement figures are from different questions; the recall range is rough; and the analysis seat recounted the 36,657 rows (models, disagreements) from the labels file but took the other read counts as supplied by the agents. The register's catches are asserted by the seats that wrote them; its counts are of logged misses, not of opportunities; the monthly share of catches credited to the operator cannot be separated from changes in who logged and when; its five gate rows are an example list; and five figures marked as coming from the 10 October internal audit were not recomputed from the ledger. The hash-chain check confirms that the published file is internally consistent (the chain is rebuilt each night), not that the ledger beneath it is complete or unedited. Section 7 is a design and a twelve-row seed from one session, with no rates: there is no re-check sample, no escape field and no before-or-after-ship field in the register yet, the origin and in-front-of fields were assigned by the same seats the ledger measures, and the Anthropic study is cited through Wolfberg's own forthcoming note, whose wording was checked but whose source was not opened. This note was drafted by Claude, a model made by Anthropic, under the operator's direction, and one of the seats it describes in Sections 6 and 7 is a Claude seat. micro1 and the authors were not contacted for this note. The 20 October revision changed the voice, layout and figures of the note, not its evidence: ages are counted from 16 October, and neither the register file nor any outside source was re-checked.

Open questions

- **Section 4, the step.** What was the late-May change in micro1's agent? If it was a vendor update behind a fixed name, the paper's drift rate is a vendor effect; if it was the interview mix, it is not about the agent at all.
- **Section 4, the 21 May batch.** The late-May step sits between the 21 May and 1 June batches, and reviewers B and C first appear in the 21 May batch, whose B-minus-A gap (-0.66, SE 0.18) is the least precisely estimated of the nine. Could an adjustment error in that one batch create part of the step? The paper does not address it.
- **Section 6.2, the labeler's drift.** How much of the 5,126 disagreements is unlogged drift in the labeler and how much is sampling noise? A fixed anchor set re-read at the end of a run would say.

- **Section 7, the re-check rate.** How many signed-off items does a one-in-ten re-check need before a per-catcher discrimination rate means anything, and who re-reads them so that the re-checker is not just another rater with the same blind spots?
- **Section 6.3, the second catcher.** Who is the second catcher for the register, and how many sessions does it take before catcher severity can be estimated apart from the error rate?

Disclosure. Drafting, the recomputation and the literature checks were done by AI models (Claude, Anthropic) working under the operator's direction, who is responsible for the content. The model that wrote this note is Claude, made by Anthropic, and the team whose scoring it describes also runs on Claude seats. Most sources were read as text; the arXiv papers were fetched from arxiv.org and read as stated in Table 1. **Competing interests.** Ownership of Wolfberg LLC. The work was not funded. **Code and data.** The recomputation script, the source table and the register counts are available on request. **Version.** 3.0, 20 October 2026: impersonal register, author block removed, figures redrawn; numbers, tables and status labels as in 2.2. 2.2, 17 October 2026: the October share on attributed rows. 2.1, 17 October 2026: the register's "operator" label glossed once and a few pair phrasings made plain.

References

Anthropic. (2026b, August 7). Auto mode is now the default in Claude Code. claude.com/resources/articles/auto-mode-default-in-claude-code (study figures taken from Wolfberg's forthcoming note, which confirmed the design at the source on 10 October 2026; not opened for this note).

Atkinson, B. (2026a). *Never let your AI grade its own homework.* How do we use this?, Theme 4. wolfberg.ai

Ford, C., Van Doren, M., & Dix, E. (2026, February 4). Alignment drift in multimodal LLMs: A two-phase, longitudinal evaluation of harm across eight model releases. arXiv:2602.04739 (abstract, results and conclusion read in part).

Katoch, H., Selby, D. A., Großmann, G., & Vollmer, S. (2026, September 25). Accounting for bias enables sustainable LLM evaluation. SuRE'26 workshop; arXiv:2609.31184 (abstract, introduction and method read).

micro1. (2026, October 14). Can a company tell whether its AI agent is getting better? micro1 research. micro1.ai/research/are-ai-agents-getting-better (full page text read).

Sachdeva, P. S., & Boudol, N. (2026, September 1; v1 July 14). Rating the raters: Rasch measurement theory for LLM evaluation. arXiv:2608.27463v2 (abstract read; method and results skimmed).

Wolfberg. (2026a). The miss register. wolfberg.ai/register; counts from the register file, build of 15 October 2026.

Wolfberg. (2026b). Full read of the practice's transcripts, 8 to 16 October 2026, and how the register is produced (audit of 10 October 2026). In-house records; counts as supplied by the agents.

Wolfberg. (2026c). Seed joint miss ledger, one session, 16 October 2026 (12 rows). In-house; recounted by the analysis seat.

Wolfberg. (2026d). Working session between the operator and the analysis seat, 16 October 2026: the question, rulings and remarks quoted in Section 7. Not logged to a file.

Wolfberg. (forthcoming). Who caught it: Keeping a public score on the human in the loop. Not yet published; the Anthropic study wording was checked against its draft on 16 October 2026.

Zhang, L., & Esposito, M. (2026, October 4). Reliability of AI agents: Rater effects, drift, and the return to an evaluation program. Working paper, arXiv:2610.07003v1 [econ.GN] (text read; figures through captions).

Appendix A. Receipts and internal labels

The text above uses plain terms. This table gives, for each in-house source and label, the name it carries in Wolfberg's files, so that every in-house number can be traced. Locations are relative to Wolfberg's research folder.

Table A1. In-house sources and the labels behind the plain terms.

Term in the text	Name in the files	Used in
Public miss register, build of 15 October	register.json, generated 2026-10-15T08:53Z (published at wolfberg.ai/register)	6.3; Exhibit 6
Labels file	fullread_labels.jsonl (36,657 rows), in fullread-20261008	6.1; Table 5
Second-stage detector run; human hand check of the detector	STAGE2-REPORT; berg_detector_RESULTS.csv, in pushback-tipping-20261015/stage2	6.1; Table 5
Relabel rounds 1 and 2; human hand-check sheet	RELABEL-REPORT; RELABEL-ROUND2-REPORT; berg_handcheck_RESULTS.csv, in pushback-tipping-20261015/relabel	6.1; Table 5
Seed joint ledger	joint_misses_seed.csv, in loop-both-20261016	7.4; Table 7
Internal audit of the register, 10 October	REGISTER-HOW-IT-WORKS.md, in register-hitl-20261010	6.3

Term in the text	Name in the files	Used in
Transcript window of 3 October	“window 33” (private extract, not published)	6.1; Table 1
Browser updates	Edge 155.0.4283.45 (10 October); Chrome 155.0.8059.79 (13 October)	1; 6.2
Flag that disables the sandbox	--no-sandbox	1; 6.2
Catcher field	caught_by; values operator, self (unprompted), peer instance, external instrument, unattributed	6.3
Miss types	repeated_miss, wrong_diagnosis, would_ship_error, recall_assert, unsearched_claim, self_corrected, probe_miss	6.3
“The seat was right”	seat_right	Table 5
The four kinds of pushback episode	kinds A, B, C and D (D is the not-pushback kind)	Table 5