THE ADD-BACK · EPISODE 03

The Verification Tax

Arc 1 · Month 18–30, value creation plan review

In the room: Operating partner, portfolio company CEO and CFO, board

The question: Why hasn't the workstream converted to margin?

THE POSITION

SIGHTING

Two years in. The automation works, the output is real, and the margin never arrived. The board pack says adoption, change management, another two quarters.

THE CONSENSUS READ

An implementation problem with an implementation fix.

THE MECHANISM

In the only controlled market where machines paid other machines for knowledge work, 39% of transactions ended in a payment dispute and the correlation between quality and payment collapsed to 0.16 for mid-grade work. Evaluators could separate excellent from failed and nothing in between. That is a checking cost, it lands in salaried time your P&L cannot see, and the study found it does not self-correct.

THE EXPOSURE

The workstream's binding constraint isn't production. It's verification, and verification isn't in the business case.

THE TEST

Ask what percentage of automated output is checked or reworked by a human, and who. Under ten percent, this doesn't apply to you.

THE BRIDGE — CLAIMED TO REALISED

The portfolio-company sizing is illustrative and applied to a composite workstream. The Diagon figures are real, published and cited with their sample; the two are kept separate throughout.

LINE	MOVES	RUNNING
Claimed — $900,000 run-rate margin from an automated knowledge-work process, 24 months in, tracking at ~40% of plan		**$900,000**
1. The production case was right [changes the story]	-$0	$900,000 of production value, real
Give the workstream its due first, because the failure is not the one people assume.		
2. The dispute rate	-$220,000 to -$320,000	$580,000 to $680,000
Same study, the number nobody quotes: approximately 39% of transactions ended in dispute by round 24, ±2 points across three seeds, still rising at the 48-round…		
3. Where the correlation collapses, and why it worsens as you scale	-$120,000 to -$180,000	$400,000 to $560,000
Within-bin correlation between quality and payment drops to r = 0.16 for tasks scoring below 0.5.		
4. Where I was wrong — I recommended a better evaluator [where I was wrong]	-$0	$400,000 to $560,000
Reading a 39% dispute rate, my recommendation was to improve evaluation: tighter acceptance criteria, structured rubrics, a second-pass automated checker. Cost: The correction is not better evaluation but scope selection: the workstream converts to margin where output quality is cheaply…		
5. What actually converts [residual]	+$0	$400,000 to $560,000
The shortfall is concentrated in the extension phase rather than the pilot, which is why it took eighteen months to become visible and why the pilot metrics were honest.		
Realised		**$400,000 to $560,000 against $900,000 planned**

THE DOWNSIDE

Multiple: 9x, illustrative

The gap between a $900,000 planned workstream and $400,000–$560,000 realised, at an illustrative 9x. Substitute your own comps.

$3.1M to $4.5M of enterprise value — in the plan and not in the business

When it surfaces. Verification cost is invisible in a pilot — small volume, motivated team, senior people checking. It becomes material at scale, which is month 18–30, which is exactly when the VCP review happens and roughly two years before you need the margin in the trailing twelve at exit. That timing is the good news: discovered at month 24 a re-scope has 18–30 months to convert. Discovered in sell-side prep it is a number in the model that isn't in the P&L, and the buyer's diligence will find the reviewers.

The Verification Tax

The question: Why hasn't the workstream converted to margin?

Paste into a request list or a management agenda. Each question resolves to an artifact, not to a characterisation.

THE DILIGENCE PACK

1. What percentage of automated output is reviewed, corrected or reworked by a person before it is used? Provide the sample and the method by which that percentage was established.

Artifact: A measured sample, not an estimate from the process owner.

2. Name every individual who performs that review, their department, their cost centre and the hours per week.

Artifact: Named list with cost centres.

If the reviewers sit outside the automated process's cost centre, the savings and the cost are in different places in your P&L.

3. Provide the rework rate broken out by transaction complexity band.

Artifact: Transaction-level data.

You are testing whether rework rises with complexity. The study says it will; if it doesn't in your case, this mechanism doesn't apply and the shortfall is something else.

4. For the workstream's next planned extension, state how output correctness will be established, and by whom, at volume.

Artifact: A written answer before funding is released.

'Spot checks by the team' is not an answer; it is an unbudgeted headcount.

5. What proportion of automated output is checkable against a system of record versus requiring human judgment?

Artifact: A classification of the work.

This is the scope-selection number from Line 4 and it determines the defensible size of the workstream.

DISQUALIFIER

If nobody can produce the number in question 1, the workstream has no measured verification cost — which means the business case has never been tested against its binding constraint.

SOURCES — EVERY MEASUREMENT WITH ITS SAMPLE

CLAIM	SOURCE	SAMPLE / CLASS
Market-traded agents earned ~1.6x profit per task ($2.62 vs $1.66) and >3x median per-task return on compute ($1.72 vs $0.54)	Shang, Liu & Jin, When Agent Markets Arrive (Diagon), arXiv 2604.06688v2	1,957 transactions, 3 seeds, 5 model families, simulated market (measured)
~39% of transactions ended in dispute by round 24, ±2 points across seeds, still rising at the 48-round extension; does not self-correct within the studied horizon	Shang, Liu & Jin (Diagon), arXiv 2604.06688v2	1,957 transactions, 3 seeds; 24-round base, 48-round extension (measured)
Within-bin quality-to-payment correlation r = 0.16 for tasks scoring below 0.5; evaluation bottleneck tightens with task complexity	Shang, Liu & Jin (Diagon), arXiv 2604.06688v2	1,957 transactions, 3 seeds, effect sizes with CIs (measured)
25–35% of a $900K workstream consumed by verification load	Applied at a conservative fraction of the study's rate — our sizing, not the study's finding	n/a — illustrative application (illustrative)
9x entry multiple	Illustrative mid-market comp	n/a — illustrative, arithmetic exposed (illustrative)

One standing caveat, and one specific to this episode. Every number in this show is somebody else's measurement, and I'll tell you whose, with the sample. None of it is diligence on your deal. Do that yourself. Specific to this episode: the 39% and the 0.16 come from a simulated market of five model families executing synthetic tasks over 24 rounds. Human reviewers evaluating real work, under a real contract, with a relationship at stake, plausibly evaluate better — I would expect them to. What I would defend is the direction and the shape: verification of mid-quality knowledge work is hard, gets harder with complexity, and does not improve on its own. What I would not defend is importing the number into your model.