

Peer Review

Review of: "Can You Trust the Graph Beneath the Agent? A Reliability Metrology for Enterprise Skills Graphs: The FACTS Dimensions and the Skills Graph Reliability Index"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper argues that enterprise skills graphs, the data layer under agentic HR tools, are judged almost only by coverage, and that coverage says nothing about whether an agent may safely act on the graph. In its place, the author proposes a reliability construct with five dimensions (Freshness, Accuracy of stated confidence, Contestability, Traceability, Stability: FACTS), derived in §6.1 from five ways the defence of an automated decision can fail. The dimensions are combined in §9 into a Skills Graph Reliability Index (SGRI) by a non-compensatory, deny-overrides rule, and the result is mapped to three bands: act, human-gate, do not automate. Fairness is handled as stratified reporting across all five dimensions (§6.7), not as a sixth one.

The second claim is mechanistic. The author names a failure mode, *evidential circularity*: agent outputs written back into the graph without provenance labels end up in the validation sample, so measured calibration improves while true calibration gets worse. §5 tests this in a simulation of 20,000 edges, and Table 2 reports that at a write-back rate of 0.50, the measured ECE (0.067) is lower than the control's (0.078) while the true ECE (0.178) is about 4.8 times the control's (0.037). The paper is explicit that the link between SGRI and decision quality is untested (P1, §12.1).

2. Whether the conclusion holds on the data

I enjoyed this paper, and I want to say concretely why. The derivation in §6.1 does real work: Table 3 shows what happens to seven commonly proposed extra dimensions, and the argument that completeness belongs to coverage rather than reliability is clean. The boundary section (§7) concedes the strongest objection instead of stretching the construct, and the formulation in §7 that reliability "sets a ceiling on achievable decision quality" is falsifiable in a way most framework papers never manage. The brittleness objection to a raw minimum is raised by the author himself in §9.1 and answered with per-dimension floors and an amber margin. The limitations in §12.3 are honest. That is a lot of good practice for a conceptual paper.

Most of the paper is analytic, so "holding on the data" applies mainly to two places: the simulation in §5 and the worked illustration in §10.2. I checked the arithmetic I could.

Table 2. The headline factor is right: $0.178 / 0.037 = 4.81$ (Table 2, column True ECE, rows 0.50 and 0.00 control). The ratio column mostly reproduces from the rounded inputs: $0.099 / 0.074 = 1.34$ and $0.141 / 0.070 = 2.01$. Two cells differ slightly: for the control, $0.037 / 0.078 = 0.474$, which rounds to 0.47, not the printed 0.48; for $w = 0.50$, $0.178 / 0.067 = 2.657$, which rounds to 2.66, not 2.65. These are almost certainly computed from unrounded values, which is fine, but a footnote saying so would prevent the question.

What the simulation isolates. This is my main substantive point. In §5.1, the written-back point estimates enter the evidence pool, and that pool feeds two things at once: the model's confidence (the Laplace-smoothed mean of the pool) and the organisation's validation labels. The control, $w = 0.00$, is described as "complete provenance labelling, under which no synthetic observation enters the evidence pool". So the control removes write-back entirely; it does not model labelled write-back. As a result, the rise in true ECE in Table 2 may come mostly from self-reinforcement of the confidence (the model averaging its own outputs), while the fall in measured ECE comes from contamination of the validation labels. Only the second is evidential circularity in the sense of §4. A 2×2 design (write-back into the confidence computation yes/no $\times$ synthetic entries admitted into the validation sample yes/no) would separate the two and make the demonstration specific to the mechanism the paper names.

The control's measured ECE. In the control, measured ECE (0.078) is roughly twice the true ECE (0.037), Table 2, row 0.00. I read this as the effect of observation error 0.12 and stale observations in the pool on the validation labels. That matters for the claim in §5.2 that contamination "inverts the ranking the instrument produces": part of why the contaminated graph at $w = 0.50$ looks better than the control is

that the control's own measurement is inflated by label noise. I would say so explicitly, and I would add a sensitivity run with lower observation error to show that the inversion survives.

Uncertainty and scope. Table 2 gives means over five seeds with no spread. With 20,000 edges, the seed-to-seed variation is probably small, but the differences between measured ECEs across conditions are also small (0.078, 0.074, 0.070, 0.067). A standard deviation or range per cell would show whether these steps are larger than the noise. The parameters are, as §5.3 says, plausible rather than estimated; a one-at-a-time sweep over drift probability (0.03), evidence arrival (18 percent), and pool cap (twelve) would show how much of the effect depends on those choices. As rendered on Qeios, I could not read Figure 4, so I could not check the "cross at cycle eight" statement in its caption against numbers.

Worked illustration, Table 7. I verified that Graph B's equal-weight mean is $(46 + 22 + 38 + 81 + 44) / 5 = 231 / 5 = 46.2$, which matches the "weighted mean of 46" in §9.1 and §10.2 if equal weights are meant (the text never says what the weights are). Graph A's SGRI of 79 equals its minimum (Contestability, 79), and Graph B's 22 equals its minimum (Accuracy of confidence). One small point: the Figure 5 caption says the mean "sits above three of the five dimension scores". With the unrounded mean of 46.2, Freshness at 46 is also below it, so it is four of five; with the rounded 46, it is three. The illustration uses a single floor of 60 for all dimensions, while §9.1 says "each dimension carries its own floor". And the margin that separates amber from red is never given a number, so I could not confirm the RED assignment for Graph B's Freshness (46) or Stability (44) on their own. Calibration at 22 is clearly red under any sensible margin.

Introductory evidence. §1 says that among the high-coverage group "roughly 40 percent were not confident in the accuracy" of their data, while the caption of Figure 1 says that among those enterprises "most are not confident in its accuracy". 40 percent is not most. The same sentence places the 92.6 percent planning major changes inside the high-coverage group; since both numbers carry the argument, giving the group size behind them would help. As rendered on Qeios, I could not read Figure 1 itself.

Overall, the conceptual claims are stated at the strength the arguments support, and the paper is careful not to claim predictive validity. The simulation supports the claim that "the failure mode is real in principle", which is exactly what §5.3 says; it would support the narrower claim about validation contamination more convincingly with the 2×2 design above.

3. Reproducibility

The paper has a Reproducibility note, which I appreciate: it lists 20,000 edges, drift 0.03, evidence arrival 18 percent, observation error 0.12, pool cap twelve, eight burn-in and twelve measured cycles, five seeds, and ECE over ten equal-width bins. That is more than many simulation sections give. No language model is used, so model versions and decoding parameters do not apply here.

What I could not find, and what I would need to reimplement Table 2 without guessing:

- Code or a notebook. None is linked. The simulation is small, and publishing it would cost little and settle every question below.
- The seed values themselves and the software used.
- How the validation sample is drawn: its size, whether it is one observation per edge or several, whether the label is the most recent observation or a random one, and whether edges in the validation sample are excluded from the confidence computation. If the same pool yields both confidence and label, confidence and label are not independent even in the control.
- What the agent's "point estimate" is: the thresholded confidence (0/1) or the confidence itself. The strength of the self-reinforcement depends on this.
- The initial prevalence of the latent true state and how the evidence pool is initialised before burn-in.
- How "synthetic share of evidence" (Table 2, last column) is computed: over the whole pool or over the validation sample.

For the index itself, the normalisation to 0–100 in §9.1 is not specified for any dimension. ECE and Brier are lower-is-better scores between 0 and 1, so some inverting and rescaling is needed; how an ECE becomes an Accuracy score of 88 or 22 (Table 7) is not stated. The same is true for the contestability composite (existence, resolution rate, latency) in §6.4. Without these mappings, two organisations cannot compute comparable SGRI, which is the stated purpose in §11.2.

4. What specifically to fix

1. **§5.1 and Table 2.** Add a labelled write-back condition: synthetic observations enter the confidence computation but are excluded from validation. Better still, run the full 2×2 described in section 2 of this review, so that true-ECE degradation from self-reinforcement is separated from measured-ECE improvement from validation contamination.

2. **§5.1.** Specify the validation sample (size, which observation is the label, overlap with the confidence computation) and define the agent's point estimate.
3. **Table 2.** Add a standard deviation or min–max across the five seeds to each cell. A footnote that the ratio column is computed from unrounded values would explain 0.48 against 0.474 and 2.65 against 2.657.
4. A short sensitivity analysis over observation error, drift, and evidence arrival (§5.3 already anticipates this). Even one appendix table would do.
5. **Reproducibility note.** Link the simulation code with seeds; the note currently says the procedure can be "reimplemented independently", and code would make that true in practice.
6. **§9.1.** Give an explicit normalisation function for each FACTS dimension, at least an illustrative one. For Freshness, a formula such as a mean of decay weights $2^{-\text{age}/\text{half-life}}$ over decision-relevant edges would make "Percentage of edges younger than their skill half-life, decay-weighted" (Table 4, §6.2) unambiguous; as written, it mixes a threshold and a weighting.
7. **§9.1 and Table 6.** Put a number on the amber margin in the worked illustration, and either use per-dimension floors in Table 7 or state that the common floor of 60 is a simplification of the rule in §9.1.
8. **Figure 5 caption.** "Three of the five" is correct only against the rounded mean; with 46.2, it is four. State which, and say that the mean is equal-weight.
9. **Figure 1 caption vs §1.** Align "most are not confident" with the "roughly 40 percent" in the text, and give the size of the high-coverage group behind the 40 and 92.6 percent.
10. **§1 and §2.3: missing references.** The Josh Bersin Company vision (130 agents, 95 capabilities), the Steve Goldberg quotation, the Digital HR Leaders podcast episode, and the HR.com 2025–2026 figures (42 and 32 percent) have no entries in the reference list. §2.1 describes a "substantial practitioner and vendor literature" without citing any of it. The JDXpert report, the main empirical anchor of the paper, is listed without a URL.
11. **§6.7.** Reporting the worst subgroup score is sound in principle, but for small groups, a minimum over subgroups is biased downward, and ECE on small samples is biased upward. I would add a minimum subgroup size or a confidence-interval rule before a subgroup failure closes the gate; otherwise, P5 could be confirmed by sampling noise alone.
12. **§6.4.** A low contest rate is ambiguous: few errors, or people who cannot or dare not contest. It would help to say how the contestability metric distinguishes these, since §6.4 relies on contest events as a

ground-truth stream, and Table 5 already flags their self-selection.

13. In P1 (§12.1), recommendation quality “judged higher in quality by independent human review” is itself an oracle in the sense of §8. One sentence applying the oracle-provenance disclosure to the outcome measure would keep the paper consistent with its own rule.

5. Recommendation

Minor revisions. This is a carefully argued design-science paper with a construct that is derived rather than listed, a clearly stated boundary, and falsifiable propositions. The simulation is a useful demonstration, and I would like it to target the named mechanism more precisely (item 1) and to be published with code. The remaining fixes are specifications (normalisation, margin, validation sampling) and citation gaps. None of them changes the paper’s argument; all of them would make SGRI something two organisations could actually compute the same way.

Declarations

Potential competing interests: No potential competing interests to declare.